Stable and Fast Learning with Momentum and Adaptive Rates

*Joseph E. Gaudio, †Travis E. Gibson, *Anuradha M. Annaswamy
*Massachusetts Institute of Technology, †Brigham and Women’s Hospital, and Harvard Medical School

Introduction

Momentum and adaptive rate methods have become the state-of-the-art for training machine learning models. This poster answers the question: How can similar algorithms, incorporating momentum techniques and adaptive rates, be used for provably correct online learning and adaptive control in dynamical systems?

Momentum-Based Algorithm Derivation [2]

\[ L(\theta, \dot{\theta}, t) = e^{\theta + \alpha t} \left[ D_\alpha(\theta) e^{-\gamma t} L(\theta) \right] + \frac{1}{2} \dot{\theta}^2 + \frac{1}{2} e^{-\gamma t} \theta^2 \]

- \( h(\theta) = \frac{1}{2} \theta^2 \)
- \( L = \frac{1}{2} \left[ \| \Delta \|^2 + e^{-\gamma t} \theta^2 \right] \)
- \( \dot{\alpha} = \ln(\Delta N_a(t)) \theta \)
- Design: \( \gamma, \beta > 0 \)
- *Ideal scaling condition* \( \beta \) is not needed

Normalizing Signal

\[ N(t) = (1 + \mu e^\phi(t)) \]

Algorithm 1 [2]

\[ L(\theta, \dot{\theta}, t) = e^{\theta + \alpha t} \left[ \frac{1}{2} \dot{\theta}^2 + \frac{1}{2} e^{-\gamma t} \theta^2 \right] \]

- Minimize functional \( J(\theta) = \int L(\theta, \dot{\theta}, t) dt \)
- Euler-Lagrange Equation: \( \left[ \theta = \int \dot{\theta}(t) dt \right] \]

Second Order ODE

\[ \ddot{\theta} + \gamma \dot{\theta} - N(t) \theta = 0 \]

Parameter Update with Momentum

\[ \dot{\theta} = -\gamma \dot{\theta} + N(t) \theta \]

- Taking \( \beta \rightarrow \infty \) (strong friction limit) results in the standard first order MRAC update: \( \dot{\theta} = -\gamma \dot{\theta} + N(t) \theta \)
- *Ideal scaling condition* \( \gamma \) enforces symmetric mixing step

Algorithm Goals

- Goal (primary): Stable and fast learning of \( \theta \) through \( \dot{\theta} \)
- Goal (secondary): Adjust \( \dot{\theta} \) (model weights) so that \( e(t) \rightarrow 0 \)
- Two algorithms are proposed to accomplish primary and secondary goals:
  - Algorithm 1: Momentum-like approaches based on high-order tuning
  - Algorithm 2: Approaches with time-varying adaptive rates

Momentum Comparison of Approaches

Algorithm Comparison

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \vartheta - \theta )</td>
<td>( \vartheta - \theta - \alpha t )</td>
</tr>
<tr>
<td>( \dot{\vartheta} - \dot{\theta} )</td>
<td>( \dot{\vartheta} - \dot{\theta} - \gamma \dot{\theta} )</td>
</tr>
</tbody>
</table>

- Natural parametrization of the algorithm as a function of the feature as compared to time
- Algorithm does not change from an overdamped to underdamped system as time progresses, and is thus capable of running continuously as features are processed, with no restart heuristic
- Online processing of the data, without a priori knowledge of its future nature
- Primary goal occurs with persistent excitation of system regressor
- Secondary goal achieved without persistent excitation
- Proven stable regardless of the initial condition, thus an optimization problem-specific schedule on the parameters of the problem is not required to set for each initial condition

Online Learning and Adaptive Control

- **Model Estimate:** \( \hat{\theta} \)
- **Plant Unknown:** \( \theta \)
- **Error:** \( e(t) = \theta(t) - \hat{\theta}(t) \)

Algorithm 2

Parameter Update with Adaptive Rate

\[ \theta(t) = -\gamma(\theta(t)) e^\varphi(t) \]

- For static parameters, primary and secondary goals achieved: exponent parameter convergence \( \theta \rightarrow \theta^* \) and model tracking error convergence \( e(t) \rightarrow 0 \)
- Time-varying parameters \( \theta(t) \) with bound proportional to: \( \| \theta - \theta^* \| \leq \bar{e} \)
- Less restrictive finite extinction properties as compared to persistent excitation. Holds on to excitation with exponential forgetting
- Adaptive rate adjustment based on regressor excitation history instead of gradient history
- Adagrad-like Algorithm for Online Learning and Adaptive Control

Concluding Remarks

- Tools rigorously developed in the field of adaptive control can be employed to provide for provably correct online learning for momentum and adaptive rate based methods
- There are numerous other similarities in problem statements, tools, concepts, and algorithms between the fields of adaptive control and optimization in machine learning
- See [8] for many other areas of opportunity for combining insights from both fields to solve new problems


This work was supported by the Air Force Research Laboratory, Collaborative Research and Development for Innovative Spacebased Operational Concepts (CREDISC), grant FA-8650-16-C-5002 and the Boeing Strategic University initiative.