Stable and Fast Learning with Momentum and Adaptive Rates

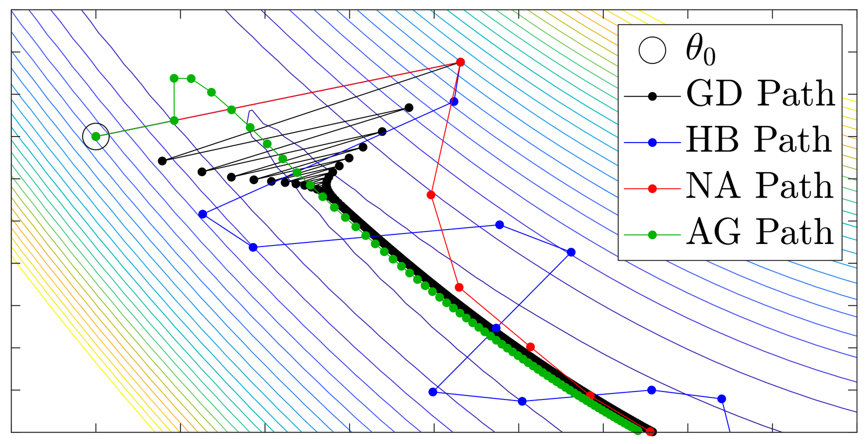
*Joseph E. Gaudio, †Travis E. Gibson, *Anuradha M. Annaswamy

*Massachusetts Institute of Technology, †Brigham and Women's Hospital, and Harvard Medical School

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?



Gradient Descent : $\theta_{k+1} = \theta_k - \gamma\nabla_\theta L(\theta_k)$

Heavy Ball : $\theta_{k+1} = \theta_k - \gamma\nabla_\theta L(\theta_k) + \beta(\theta_k - \theta_{k-1})$

Nesterov Accel : $\theta_{k+1} = \theta_k - \gamma\nabla_\theta L(\theta_k + \beta(\theta_k - \theta_{k-1})) + \beta(\theta_k - \theta_{k-1})$

AdaGrad : $\theta_{k+1} = \theta_k - \gamma\Gamma_k\nabla_\theta L(\theta_k)$

Continuous Time Formulations

[1] Continuous Nesterov Accel : $\ddot{\theta} + \frac{3}{t}\dot{\theta} = -\gamma\nabla_\theta L(\theta)$

Continuous Adaptive Rates : $\dot{\theta} = -\gamma\Gamma(t)\nabla_\theta L(\theta)$

Online Learning and Adaptive Control

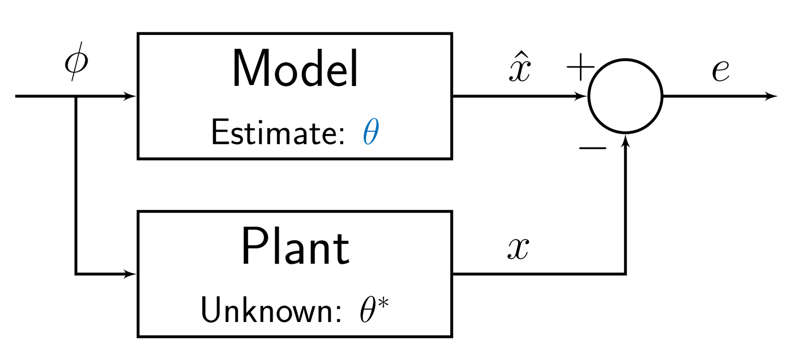


Figure: Dynamical Error Model Formulation

$$\dot{e}(t) = Ae(t) + b\tilde{\theta}^T(t)\phi(t) \quad (1)$$
$$\tilde{\theta}(t) = \theta(t) - \theta^* \quad (2)$$

Assumption: Model is an exact approximation

- $\phi(t)$ is a time-varying regressor

Algorithm Goals

- Goal (primary): Stable and fast learning of θ^* through $\theta(t)$
- Goal (secondary): Adjust $\theta(t)$ (model weights) so that $e(t) \to 0$
- 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

[1] W. Su, S. Boyd, and E. J. Candès, "A differential equation for modeling nesterov's accelerated gradient method: Theory and insights," *Journal of Machine Learning Research*, vol. 17, no. 153, pp. 1–43, 2016.

[2] J. E. Gaudio, T. E. Gibson, A. M. Annaswamy, and M. A. Bolender, "Provably correct learning algorithms in the presence of time-varying features using a variational perspective," *arXiv preprint arXiv:1903.04666*, 2019.

[3] A. Wibisono, A. C. Wilson, and M. I. Jordan, "A variational perspective on accelerated methods in optimization," *Proceedings of the National Academy of Sciences*, vol. 113, pp. E7351–E7358, nov 2016.

[4] A. S. Morse, "High-order parameter tuners for the adaptive control of linear and nonlinear systems," in *Systems, Models and Feedback: Theory and Applications*, pp. 339–364, BirkhÃ¤user Boston, 1992.

[5] J. E. Gaudio, A. M. Annaswamy, M. A. Bolender, and E. Lavretsky, "Adaptive rates for stable and fast learning in dynamical systems." Invention Disclosure, 2019.

[6] J. E. Gaudio, T. E. Gibson, A. M. Annaswamy, M. A. Bolender, and E. Lavretsky, "Connections between adaptive control and optimization in machine learning," *arXiv preprint arXiv:1904.05856*, 2019.

This work was supported by the Air Force Research Laboratory, Collaborative Research and Development for Innovative Aerospace Leadership (CRDInAL), Thrust 3 - Control Automation and Mechanization grant FA 8650-16-C-2642 and the Boeing Strategic University Initiative.

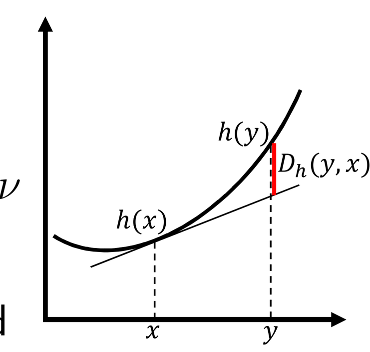
Momentum-Based Algorithm Derivation [2]

Bregman Lagrangian [3]

$$\mathcal{L}(\theta,\dot{\theta},t) = \mathrm{e}^{\bar{\alpha}_t+\bar{\gamma}_t}\left(D_h(\theta + \mathrm{e}^{-\bar{\alpha}_t}\dot{\theta},\theta) - \mathrm{e}^{\bar{\beta}_t}L(\theta)\right)$$

damping — kinetic energy — potential energy

- $h(\cdot) = \frac{1}{2}\|\cdot\|^2$
- $L = \frac{d}{dt}\left\{\frac{e^TPe}{2}\right\} + \frac{e^TQe}{2}$
- $\bar{\alpha}_t = \ln(\beta\mathcal{N}_t)$, $\bar{\beta}_t = \ln\left(\frac{\gamma}{\beta\mathcal{N}_t}\right)$, $\bar{\gamma}_t = \int_{t_0}^t \beta\mathcal{N}_\nu d\nu$
- Design: $\gamma, \beta, \mu > 0$
- "Ideal scaling condition" $\dot{\bar{\beta}}_t \leq \mathrm{e}^{\bar{\alpha}_t}$ not needed



Normalizing Signal

$$\mathcal{N}_t \triangleq (1 + \mu\phi^T\phi) \quad (3)$$

Algorithm 1 [2]

$$\mathcal{L}(\theta,\dot{\theta},t) = \mathrm{e}^{\int_{t_0}^t \beta\mathcal{N}_x dx}\frac{1}{\beta\mathcal{N}_t}\left(\frac{1}{2}\dot{\theta}^T\dot{\theta} - \gamma\beta\mathcal{N}_t\left[\frac{d}{dt}\left\{\frac{e^TPe}{2}\right\} + \frac{e^TQe}{2}\right]\right) \quad (4)$$

damping — kinetic energy — potential energy

- Minimize functional $J(\theta) = \int_{\mathbb{T}} \mathcal{L}(\theta,\dot{\theta},t)dt$
- Euler-Lagrange Equation: $\frac{d}{dt}\left(\frac{\partial\mathcal{L}}{\partial\dot{\theta}}(\theta,\dot{\theta},t)\right) = \frac{\partial\mathcal{L}}{\partial\theta}(\theta,\dot{\theta},t)$

Second Order ODE

$$\ddot{\theta} + \left[\beta\mathcal{N}_t - \frac{\dot{\mathcal{N}}_t}{\mathcal{N}_t}\right]\dot{\theta} = -\gamma\beta\mathcal{N}_t\phi e^TPb \quad (5)$$

Parameter Update with Momentum

$$\text{Gradient-Like Step} \quad \dot{\vartheta} = -\gamma\phi e^TPb$$
$$\text{Mixing Step} \quad \dot{\theta} = -\beta(\theta - \vartheta)\mathcal{N}_t \quad (6)$$

- Taking $\beta \to \infty$ (strong friction limit) results in the standard first order MRAC update: $\dot{\theta}(t) = -\gamma\phi(t)e^T(t)Pb$
- "Ideal scaling condition" $\dot{\bar{\beta}}_t = \mathrm{e}^{\bar{\alpha}_t}$ enforces symmetric mixing step

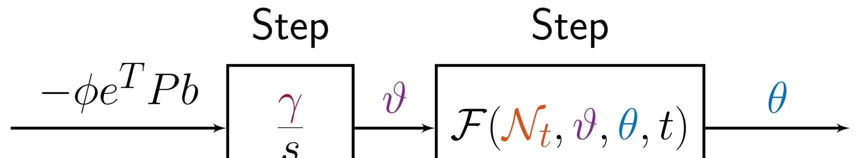


Figure: Block diagram of the algorithm in (6).

Momentum Comparison of Approaches

Algorithm Comparison

Parameterization from [3]	Our Approach
$\mathcal{L} = \frac{t^{p+1}}{p}\left(\frac{1}{2}\dot{\theta}^T\dot{\theta} - Cp^2t^{p-2}\frac{1}{2}e_y^2\right)$	$\mathcal{L} = \mathrm{e}^{\int_{t_0}^t \beta\mathcal{N}_x dx}\frac{1}{\beta\mathcal{N}_t}\left(\frac{1}{2}\dot{\theta}^T\dot{\theta} - \gamma\beta\mathcal{N}_t\left[\frac{d}{dt}\left\{\frac{e^TPe}{2}\right\} + \frac{e^TQe}{2}\right]\right)$
$\ddot{\theta} + \frac{p+1}{t}\dot{\theta} = -Cp^2t^{p-2}\phi e_y$	$\ddot{\theta} + \left[\beta\mathcal{N}_t - \frac{\dot{\mathcal{N}}_t}{\mathcal{N}_t}\right]\dot{\theta} = -\gamma\beta\mathcal{N}_t\phi e^TPb$

- Natural parameterization 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 variation
- 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

Momentum Stability Analysis

Lyapunov Function Comparison

Lyapunov Function in [3] For Time-Varying Regression

$$V = \frac{1}{2}\|\tilde{\theta} + \frac{1}{\beta(1+\mu\phi^T\phi)}\dot{\theta}\|^2 + \frac{\gamma}{\beta(1+\mu\phi^T\phi)}\frac{1}{2}e_y^2$$

$$\dot{V} = -\gamma e_y^2\left(1 + \frac{\mu\phi^T\dot{\phi}}{\beta(1+\mu\phi^T\phi)^2}\right) + \frac{\gamma}{\beta(1+\mu\phi^T\phi)}e_y\tilde{\theta}^T\dot{\phi}$$

Our Control Inspired Lyapunov Function [4]

$$V = \frac{1}{\gamma}\|\vartheta - \theta^*\|^2 + \frac{1}{\gamma}\|\theta - \vartheta\|^2 + e^TPe$$

Gradient Step Error — Mixing Step Error — Model Prediction Error

$$\dot{V} \leq -\frac{2\beta}{\gamma}\|\theta - \vartheta\|^2 - \|e\|^2 - [\|e\| - 2\|Pb\|\|\theta - \vartheta\|\|\phi\|]^2 \leq 0$$

$$\|\theta - \vartheta\|_{\mathcal{L}_2}^2 \leq \frac{\gamma V(t_0)}{2\beta} \qquad \text{as } \beta \to \infty,\ \|\theta - \vartheta\|_{\mathcal{L}_2}^2 \to 0$$

- With bounded feature magnitude and time derivative:

$$\lim_{t\to\infty} e(t) = 0,\ \lim_{t\to\infty}(\theta(t) - \vartheta(t)) = 0,\ \lim_{t\to\infty}\dot{\vartheta}(t) = 0,\ \lim_{t\to\infty}\dot{\theta}(t) = 0$$

- Regret bounded/constant:

$$\text{Regret}_{\text{continuous}} := \int_0^T \|e(\tau)\|^2 d\tau = \mathcal{O}(1)$$

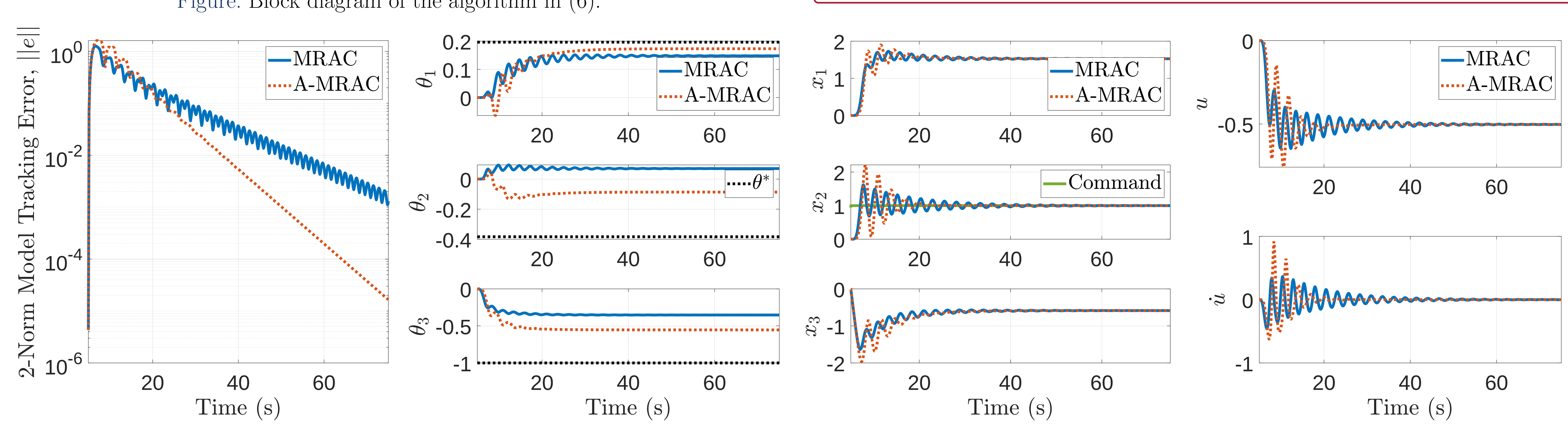


Figure: State Feedback adaptive control - step response.

Algorithm 2

Parameter Update with Adaptive Rate

$$\dot{\theta}(t) = -\gamma\Gamma(t)\phi(t)e^T(t)Pb \quad (7)$$

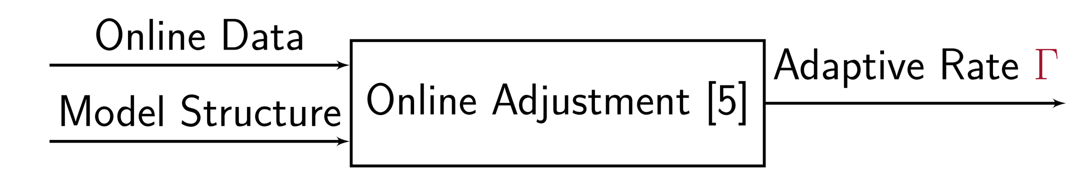


Figure: Adaptive rate block diagram.

- For static parameters, primary and secondary goals achieved: exponential parameter convergence $\theta \to \theta^*$ and model tracking error convergence $e \to 0$
- Track time-varying parameters $\theta^*(t)$ with bound proportional to:

$$\|\tilde{\theta}(t)\| \propto \|\dot{\theta}^*(t)\|$$

- Less restrictive finite excitation properties as compared to persistent excitation. Holds onto 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

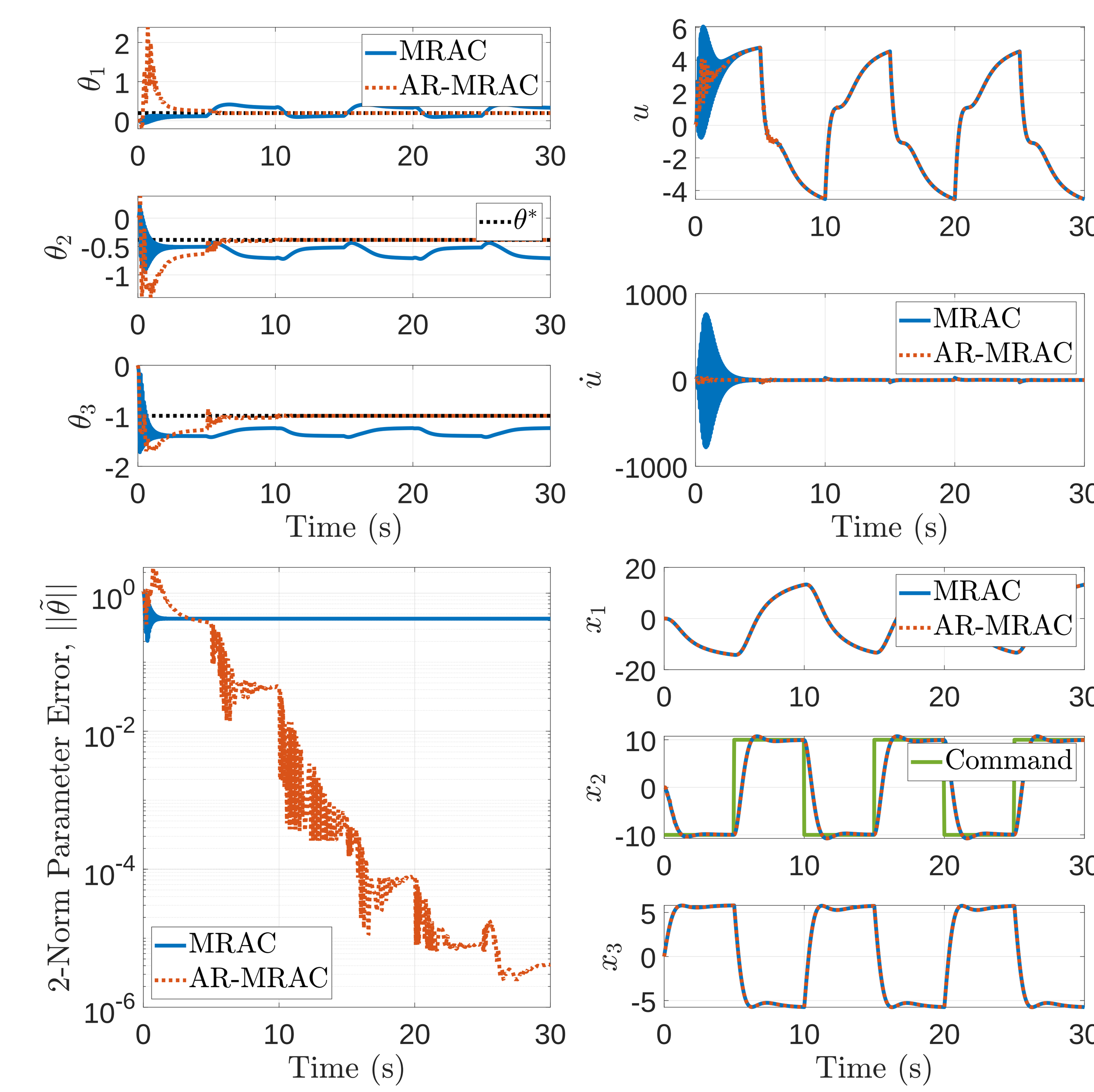


Figure: Adaptive rate parameter convergence

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 [6] for many other areas of opportunity for combining insights from both fields to solve new problems