Exponential weight averaging as damped harmonic motion

Jonathan Patsenker*[1], Henry Li*[1], Yuval Kluger[1]

[1]Applied Mathematics Program, Yale University

*Equal contribution; order decided by coin toss

ICML International Conference On Machine Learning

Yale

Motivations

- The exponential moving average (EMA) of neural network weights is a commonly used in deep learning optimization, especially in generative models
- EMA improves the stability of the inference model during and after training.
- Benefits *after* training have been studied
- Benefits *during* training not well understood.

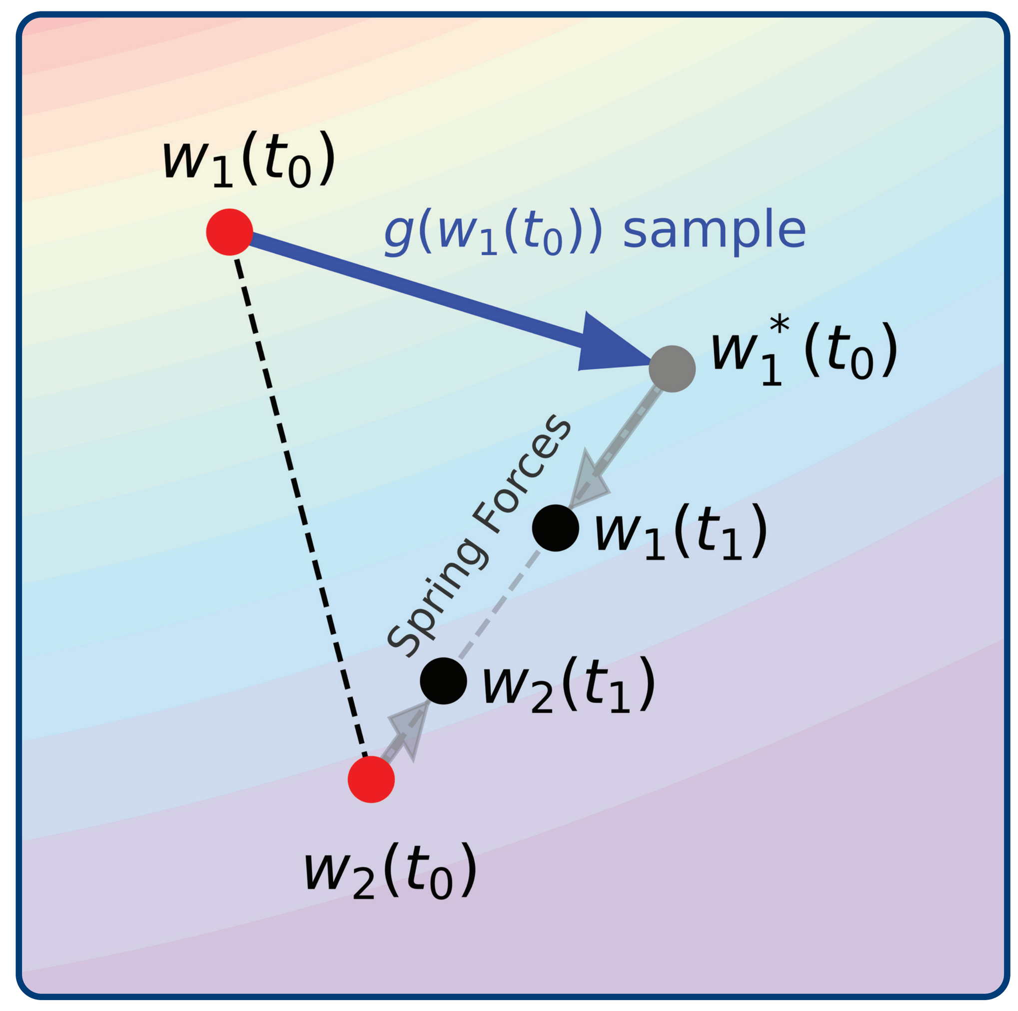


Fig 1. A visualization of the **BELAY** update step. The background color corresponds to the true full-batch loss function, and g is sampled using an optimizer on a minibatch.

BELAY: Physical EMA

Let $\mathbf{w}_1$, $\mathbf{w}_2$ represent point-particles with masses m_1, m_2, attached by a 0-length spring with spring constant k. The particles are subject to damping with constants c_1, c_2 respectively. External forces notated by $f(\mathbf{w}_1,t)$ are exerted upon $\mathbf{w}_1$ but not $\mathbf{w}_2$. We break down the total forces ($\mathbf{F}_1$,$\mathbf{F}_2$) exerted on $\mathbf{w}_1$, $\mathbf{w}_2$.

$$\boldsymbol{F_1} = \underbrace{k(\boldsymbol{w}_2 - \boldsymbol{w}_1)}_{\text{Hookean}} - \underbrace{c_1\dot{\boldsymbol{w}}_1}_{\text{Damping}} + \underbrace{f(\boldsymbol{w}_1, t)}_{\text{External}} = \underbrace{m_1\ddot{\boldsymbol{w}}_1}_{\text{Newton's 2nd Law}}$$

$$\boldsymbol{F_2} = k(\boldsymbol{w}_1 - \boldsymbol{w}_2) - c_2\dot{\boldsymbol{w}}_2$$

$$\ddot{\boldsymbol{w}}_1 = \frac{k}{m_1}(\boldsymbol{w}_2 - \boldsymbol{w}_1) - \frac{c_1}{m_1}\dot{\boldsymbol{w}}_1 + \frac{1}{m_1}f(\boldsymbol{w}_1, t)$$

$$\ddot{\boldsymbol{w}}_2 = \frac{k}{m_2}(\boldsymbol{w}_1 - \boldsymbol{w}_2) - \frac{c_2}{m_2}\dot{\boldsymbol{w}}_2$$

Harmonic Oscillator: motion of spring system

Discretization with Kinematics

$$\begin{aligned}\boldsymbol{w}_1(t+1) &= \boldsymbol{w}_1(t) + \dot{\boldsymbol{w}}_1(t) + \frac{k}{2m_1}(\boldsymbol{w}_2(t) - \boldsymbol{w}_1(t)) - \frac{c_1}{2m_1}\dot{\boldsymbol{w}}_1 + \frac{1}{2m_1}f(\boldsymbol{w}_1, t)\\ &= (1-\beta)\underbrace{\boldsymbol{w}_1^*(t)}_{\boldsymbol{w}_1(t)+\eta f(\boldsymbol{w}_1,t)} + \beta\boldsymbol{w}_2(t) \overset{\beta\to 0}{=} \boldsymbol{w}_1^*(t)\end{aligned}$$

Optimizer Update

$$\begin{aligned}\boldsymbol{w}_2(t+1) &= \boldsymbol{w}_2(t) + \dot{\boldsymbol{w}}_2(t) + \frac{k}{2m_2}(\boldsymbol{w}_1(t) - \boldsymbol{w}_2(t)) - \frac{c_2}{2m_2}\dot{\boldsymbol{w}}_2\\ &= (1-\alpha)\boldsymbol{w}_2(t) + \alpha\boldsymbol{w}_1(t) \rightarrow \mathbf{EMA}(\boldsymbol{w}_1)\end{aligned}$$

When $c_1 = 2m_1$, $c_2 = 2m_2$, for constants α, β, η.

BELAY: *modified EMA as Harmonic Oscillator*

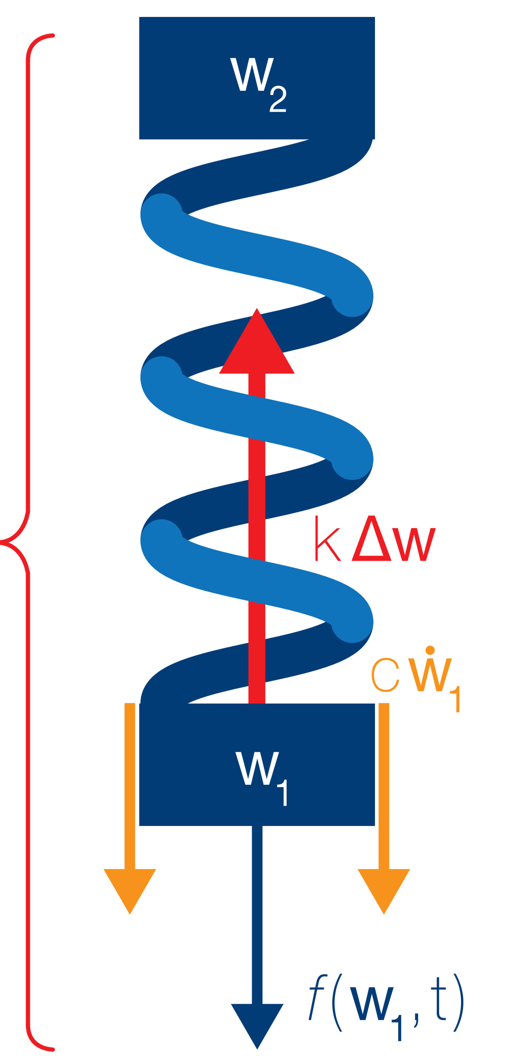


Fig P2. An illustration of a classical Hookean spring system in the scenario described. Forces illustrated on $\mathbf{w}_1$. Forces on $\mathbf{w}_2$ are not pictured, but would be mirror images, with no external force applied.

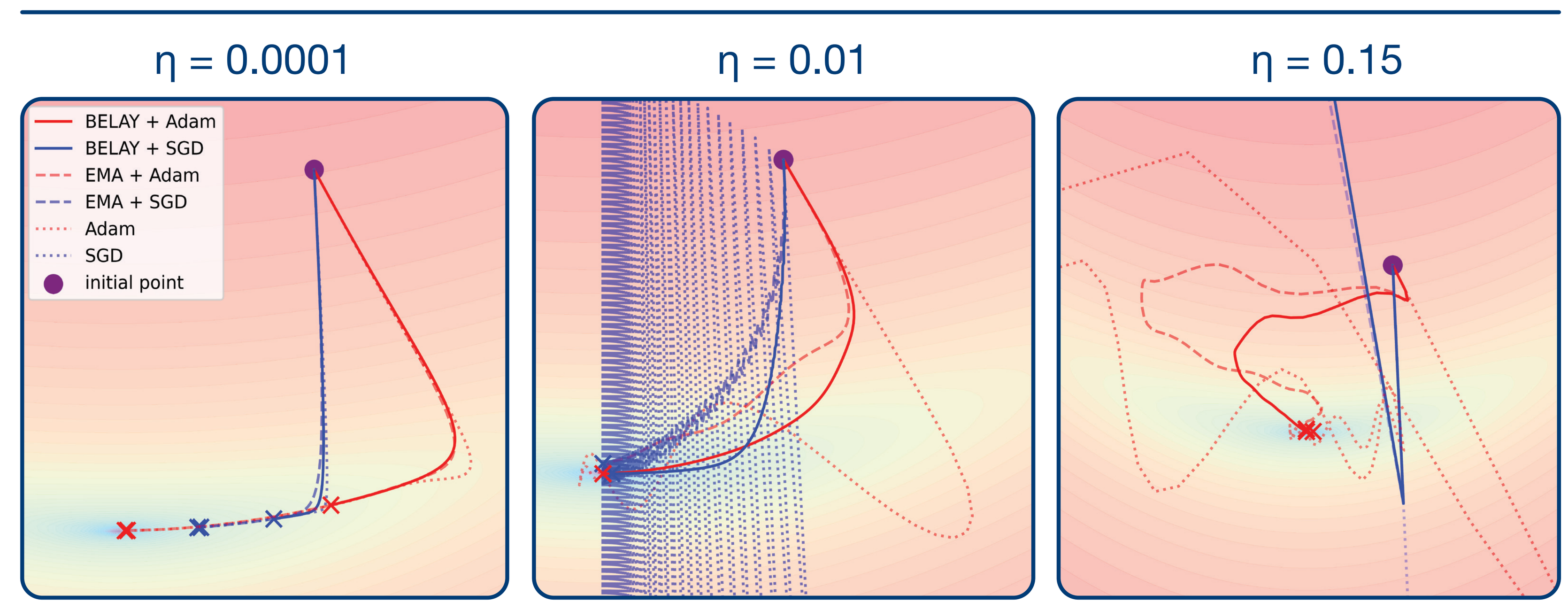


Fig 2. Comparison of **BELAY** against EMA and a control, using both Adam, and SGD on the Rosenbrock function across learning rates. Robustness to learning rate (η) is related to robustness across varying function smoothness.

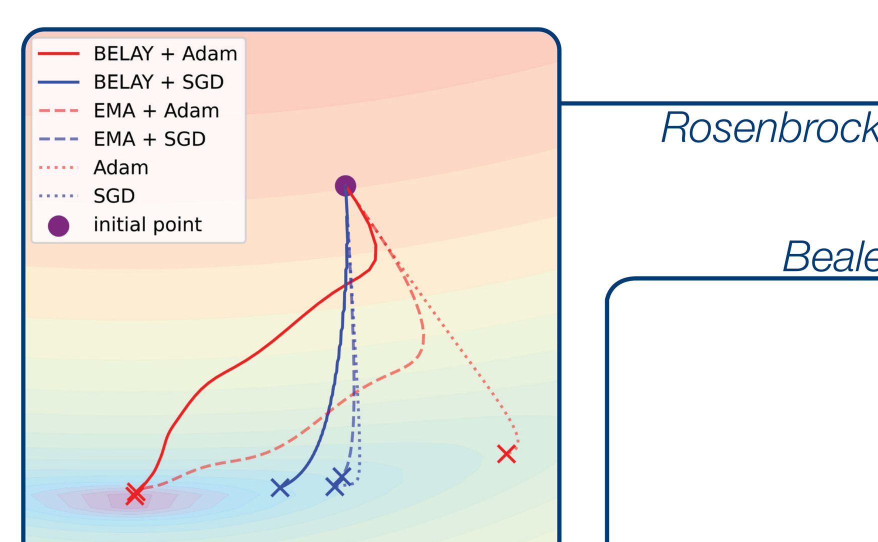
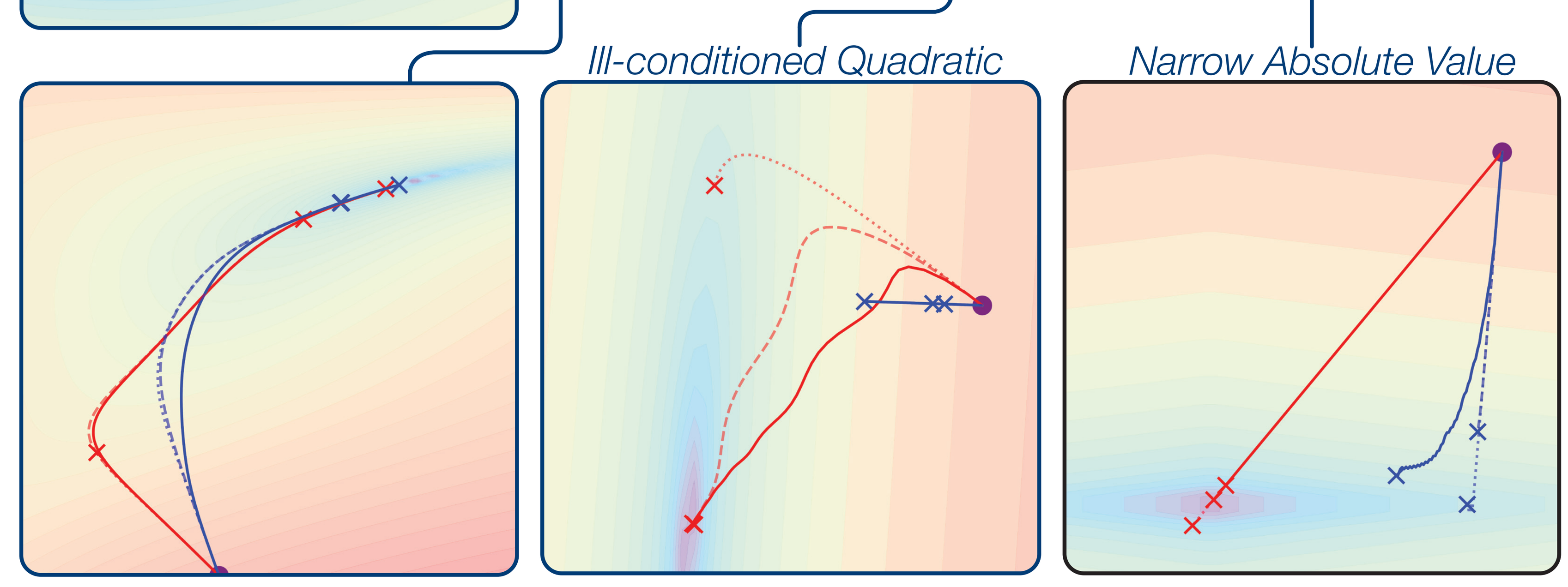


Fig 3. Comparison of **BELAY** against EMA and a control, using both Adam, and SGD with top performing parameters across various functions. Experiments stopped once one method crossed a convergence threshold.



Further Insights

Connections to Momentum-based Methods

Momentum

$$\boldsymbol{v}(t) = \lambda g(\boldsymbol{w}(t)) + (1-\lambda)\boldsymbol{v}(t-1) = (1-\lambda)^s\lambda g(\boldsymbol{w}(t-s))$$

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + \alpha\boldsymbol{v}(t) = \boldsymbol{w}(t) + \alpha\sum_{s=0}^{t} a_s g(\boldsymbol{w}(t-s))$$

Linear g

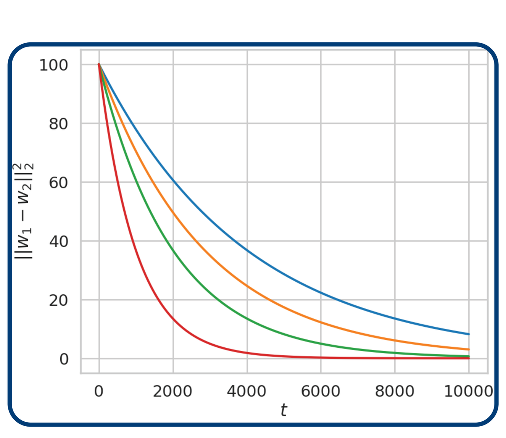
$$= \boldsymbol{w}(t) + \alpha g\left(\sum_{s=0}^{t} a_s \boldsymbol{w}(t-1)\right) = \boldsymbol{w}(t) + \alpha g(\boldsymbol{w}^{EMA}(t))$$

BELAY

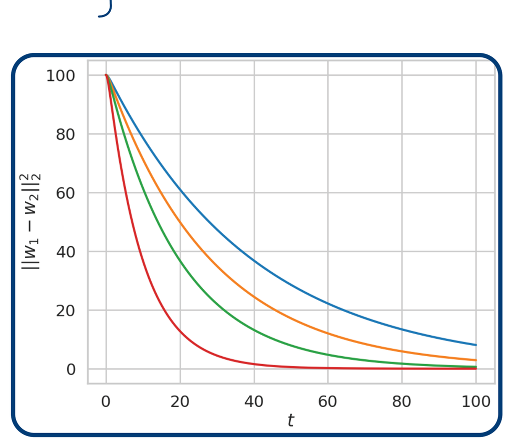
Physically-based Spring Parameterization

$$\boldsymbol{w}(t) = C_1 e^{\left(-\delta+\sqrt{\delta^2-\frac{k}{m}}\right)t} + C_2 e^{\left(-\delta-\sqrt{\delta^2-\frac{k}{m}}\right)t}$$

Spring System Solution



Time-invariant Dynamics



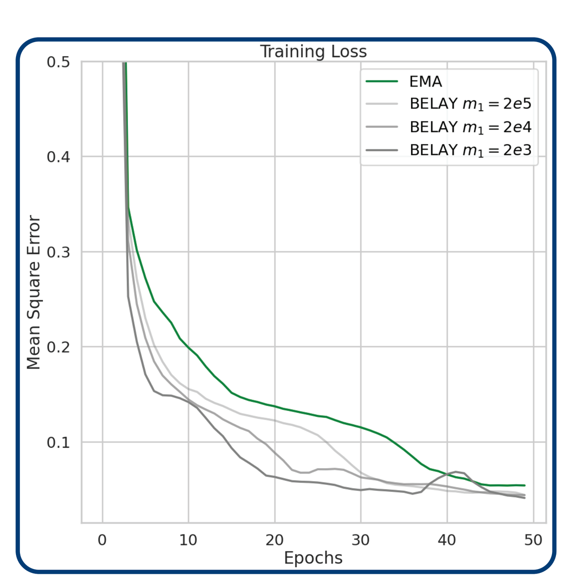


Fig 4. Comparison of **BELAY** against EMA on the MNIST dataset. The standard EMA algorithm is compared against **BELAY** with various settings of the model mass $\mathbf{m}_2$.