# A Qualitative Study of the Dynamic Behavior of Adaptive Gradient Algorithms

**Chao Ma, Lei Wu, Weinan E**

Mathematical and Scientific Machine Learning 2021

July 29, 2021

# Adaptive gradient method and observations

RMSprop:

$$v_t = \alpha v_{t-1} + (1-\alpha) g_t^2$$
$$\theta_{t+1} = \theta_t - \eta \frac{g_t}{\sqrt{v_t + \epsilon}}$$

Adam:

$$v_t = \alpha v_{t-1} + (1-\alpha) g_t^2$$
$$m_t = \beta m_{t-1} + (1-\beta) g_t$$
$$\theta_{t+1} = \theta_t - \eta \frac{m_t/(1-\beta^t)}{\sqrt{v_t/(1-\alpha^t)} + \epsilon}$$


1. **Fast initial convergence**
2. **Small oscillations**
3. **Large spikes**

# Fast initial convergence: perspective from signGD

**Continuous limits:**

$\alpha, \beta$ fixed, $\eta \to 0$:

$$\dot{\boldsymbol{x}} = -\frac{\nabla f(\boldsymbol{x})}{|\nabla f(\boldsymbol{x})| + \epsilon}.$$

Sign GD when $\epsilon = 0$

$\alpha = 1 - a\eta$, $\beta = 1 - b\eta$:

$$\dot{\boldsymbol{v}} = a(\nabla f(\boldsymbol{x})^2 - \boldsymbol{v})$$
$$\dot{\boldsymbol{m}} = b(\nabla f(\boldsymbol{x}) - \boldsymbol{m})$$
$$\dot{\boldsymbol{x}} = -\frac{(1 - e^{-bt})^{-1}\boldsymbol{m}}{\sqrt{(1 - e^{-at})^{-1}\boldsymbol{v}} + \epsilon}$$

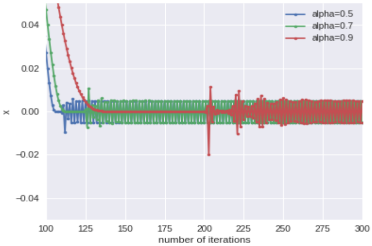
### Theorem

*Assume the objective function satisfies the Polyak-Lojasiewicz (PL) condition: $\|\nabla f(\boldsymbol{x})\|_2^2 \geq \mu f(\boldsymbol{x})$ for any $\boldsymbol{x}$. Then, for $\boldsymbol{x}(\cdot)$ given by the continuous-time signGD dynamics $\dot{\boldsymbol{x}} = -sign(\nabla f(\boldsymbol{x}))$, we have*

$$f(\boldsymbol{x}(t)) \leq \left(\sqrt{f(\boldsymbol{x}_0)} - \frac{\sqrt{\mu}}{2}t\right)^2.$$

# Small oscillations: insights from linearization

(RMSprop) falls into $2$-periodic solution for simple objective functions. For $f(x) = \frac{1}{2}x^2$, RMSprop oscillates at $-\frac{\eta}{2}$ and $\frac{\eta}{2}$.




Continuous RMSprop:

$$\dot{\boldsymbol{v}} = a(\nabla f(\boldsymbol{x})^2 - \boldsymbol{v})$$
$$\dot{\boldsymbol{x}} = -\frac{\nabla f(\boldsymbol{x})}{\sqrt{\boldsymbol{v}} + \epsilon}$$

Linearization around stationary point $(x^*, 0)$:

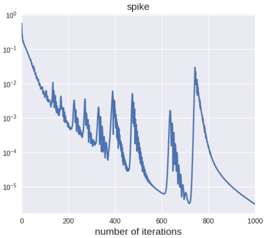
$$\dot{\boldsymbol{x}} = -\frac{\nabla^2 f(\boldsymbol{x}^*)}{\epsilon}(\boldsymbol{x} - \boldsymbol{x}^*),$$
$$\dot{\boldsymbol{v}} = -a\boldsymbol{v}.$$

Jacobian matrix:

$$\begin{bmatrix} -\frac{\nabla^2 f(\boldsymbol{x}^*)}{\epsilon} & 0 \\ 0 & -aI \end{bmatrix},$$

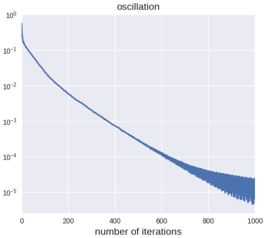
# Influence of hyper-parameters for Adam

**The spike regime**
When $b$ is sufficiently larger than $a$. The optimization process is unstable.



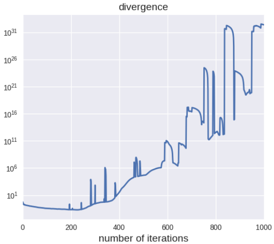

a=1, b=100

**The oscillation regime**
When $a$ and $b$ are in the same order. Small loss and stable loss curve can be achieved.




a=10, b=10

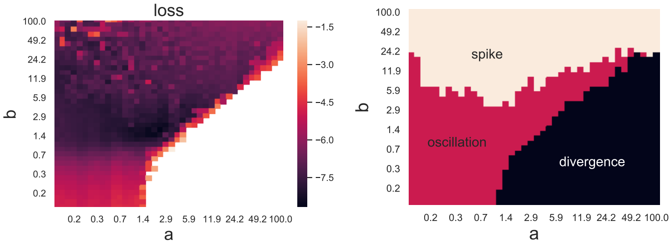
**The divergence regime**
When $a$ is sufficiently larger than $b$. Unstable and may diverges after a period of training.




a=100, b=1

# Influence of hyper-parameters for Adam

The distribution of the three regimes on the diagram of $a$ and $b$:




**Summary:**

- Qualitative behaviors of Adam and RMSprop are studied.
- Three typical features are summarized on the loss curve of adaptive gradient algorithms.
- Three behavior patterns are identified for Adam with different hyper-parameters. Observations show that small and stable loss curve can be achieved in the oscillation regime (where $a \approx b$).