A general system of differential equations to model first order adaptive algorithms

Application to ADAM

André Belotto da Silva* & Maxime Gazeau**

∗ Aix-Marseille Université, Institut de Mathématiques de Marseille

∗∗ Borealis AI; Current: LG Electronics

andre-ricardo.belotto-da-silva@univ-amu.fr

Introduction

Optimization is at the core of many machine learning problems. Estimating the model parameters can often be formulated in terms of an unconstrained optimization problem of the form

$$\min_{\theta\in\mathbb{R}^d} f(\theta) \qquad \text{where } f:\mathbb{R}^d\to\mathbb{R} \text{ is differentiable.} \tag{1}$$

- Emergence of adaptive algorithms ADAM, RMSPROP, AMSGRAD, ADAGRAD in machine learning.
- Commonly observed that the value of the training loss decays faster than for stochastic gradient descent. Became the default method of choice for training feed-forward and recurrent neural networks.
- Can we provide a theoretical framework to study adaptive algorithms? Can we obtain conditions on the hyper-parameters that guarantee convergence of trajectories?
- What properties make them so well suited for deep learning? Is it the right class of algorithms to optimize the loss surface given by deep neural networks?

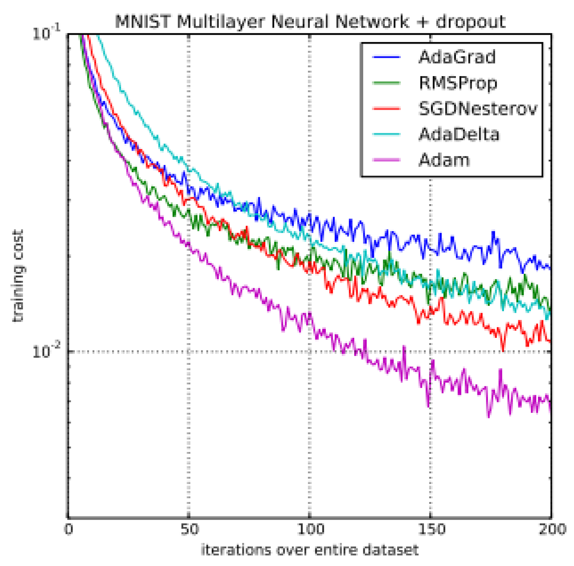

Figure 1: Training of multilayer neural networks on MNIST images using dropout stochastic regularization [3].

Dynamics of first order optimization algorithms

We analyze discrete *adaptive* optimization algorithms by introducing their continuous time counterparts, with a focus on ADAM. The connection between difference equations and continuous differential equations [4] is an active area of research in both the deterministic and stochastic setting [2].

Continuous equation	Discrete optimizer
Gradient flow	Gradient descent Proximal method
Second order eq. [5, 6]	Heavy ball Nesterov
?	Adaptive algorithms

In our work [1], we study the following general system of differential equations:

$$\begin{cases} \dot\theta(t) = -m(t)/\sqrt{v(t)+\varepsilon} \\ \dot m(t) = h(t)\nabla f(\theta(t)) - r(t)m(t) \\ \dot v(t) = p(t)\left[\nabla f(\theta(t))\right]^2 - q(t)v(t), \end{cases} \tag{2}$$

Which allow us to recover several optimization algorithms, such as:

1. **Heavy Ball**: $h(t)\equiv 1, \quad r(t)\equiv\gamma, \quad$ and $\quad p(t)\equiv q(t)\equiv 0.$
2. **Nesterov**: $h(t)\equiv 1, \quad r(t)=r/t, \quad$ *and* $\quad p(t)\equiv q(t)\equiv 0.$
3. A modification of the equation gives **ADAGRAD ($q\equiv 0$) and RMSPROP ($p\equiv q\equiv \alpha_2$).**

In order to establish a relation between the continuous ODE and the optimization algorithms, we study the finite difference approximation of (2) by the forward Euler method

$$\begin{cases} \theta_{k+1} = \theta_k - s m_k/\sqrt{v_k+\varepsilon} \\ m_{k+1} = (1-sr(t_{k+1}))m_k + sh(t_{k+1})\nabla f(\theta_{k+1}) \\ v_{k+1} = (1-sq(t_{k+1}))v_k + sp(t_{k+1})\left[\nabla f(\theta_{k+1})\right]^2 \end{cases} \tag{3}$$

where $t_k = ks$. In [1], we address the following questions

- *Existence/Uniqueness*: Wellposedness of the Cauchy problem (2).
- *Convergence analysis*: Find sufficient conditions on the functions f and p, q, r, h in order for the solutions of equation (2) to converge to a critical value of f. We have four main lines of results:

(0) *Gradient convergence:* Sufficient conditions so that $\nabla f(\theta(t)) \to 0$ when $t\to\infty$.

(I) *Topological convergence:* Sufficient conditions so that $\theta(t)$ converges to a critical value of f.

(II) *Avoiding local maximum and saddles*: Sufficient conditions so that the dynamics avoid local maximum and saddle points and only converge to local minimum.

(III) *Rate of convergence*: Under the convexity assumption, find the rate of convergence.

Connection to existing optimization algorithms: ADAM

Iterative method generating a sequence $(\theta_k, m_k, v_k) \in \mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d_+$. The algorithm can be reformulated as follows: for any constants $\beta_1,\beta_2\in(0,1)$, $\varepsilon>0$ and initial vectors $\theta_0\in\mathbb{R}^d$, $m_0 = \nabla_\theta f(\theta_0)$, $v_0 = \nabla_\theta f(\theta_0)^2$ and for all $k\geq 0$

$$\begin{cases} \theta_{k+1} = \theta_k - s\, m_k/\sqrt{v_k+\varepsilon} \\ g_{k+1} = \nabla f(\theta_{k+1}) \\ m_{k+1} = \mu_{k+2}m_k + (1-\mu_{k+2})g_{k+1} \\ v_{k+1} = \nu_{k+2}v_k + (1-\nu_{k+2})g_{k+1}^2 \end{cases} \tag{4}$$

where the two parameters for the moving average, depending on the iterations, are given by $\mu_k = \beta_1(1-\beta_1^{k-1})/(1-\beta_1^k)$ and $\nu_k = \beta_2(1-\beta_2^{k-1})/(1-\beta_2^k)$. Consider now the family of differential equations (2) where the coefficients are given by

$$h\equiv r\equiv g_1^A(t,\lambda,\alpha_1,\alpha_2), \qquad p\equiv q\equiv g_2^A(t,\lambda,\alpha_1,\alpha_2), \qquad g_i^A(t,\lambda,\alpha_1,\alpha_2) = \frac{1-e^{-\lambda/\alpha_i}}{\lambda\left(1-e^{-t/\alpha_i}\right)},$$

where $(\lambda,\alpha_1,\alpha_2)$ are positive real numbers. Note that both functions have a simple pole at $t=0$. Now, let us consider the associated discretization (3) with learning rate s and a sub-family of discrete models parametrized by $(\beta_1,\beta_2)\in(0,1)\times(0,1)$ which are given by $\lambda = s$ and $\beta_i = e^{-\lambda/\alpha_i}$. It easily follows that for $i=1,2$

$$sg_i^A((k+1)s,\lambda,\alpha_1,\alpha_2) = 1-\beta_1\frac{1-\beta_1^k}{1-\beta_1^{k+1}} = 1-\mu_{k+1},$$

which recovers ADAM's discrete system. We can now present a simplified version of our results:

Theorem 1 (Convergence of ADAM)**.** *Suppose that f is a C^2 and coercive function, $\varepsilon>0$ and*

$$3+\beta_2 > 4\beta_1, \qquad \textit{where} \qquad \beta_i = \exp(-\lambda/\alpha_i), \quad i=1,2.$$

(0) Convergence of the gradient: *Suppose that the loss function f is bounded from below and its gradient ∇f is globally Lipschitz and bounded. Then $\nabla f(\theta(t))\to 0$ when $t\to\infty$.*

(I) Topological convergence: *We have that $f(\theta(t))\to f_\star$, $m(t)\to 0$ and $v(t)\to 0$ when $t\to\infty$, where $f_\star$ is a critical value of f.*

(II) Non-local minimum avoidance: *Suppose that assumptions f is Morse. Fix $t_0>0$ and denote by S_{t_0} the set of initial conditions $(\theta_0,m_0,v_0)\in\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d_{\geq 0}$ such that $\theta_\star\in\omega(\theta(t))$, where $\theta_\star$ is not a local-minimum of f. Then S_{t_0} has Lebesgue measure zero.*

(III) Rate of convergence: *Suppose that f is convex. There exists a constant $\mathcal{K}>0$ which depends on f, θ_0 and v_0, so that:*

$$\lim_{t\to\infty} f(\theta(t)) - f(\theta_\star) < \mathcal{K}\frac{1-e^{-\lambda/\alpha_2}}{\alpha_1(1-e^{-\lambda/\alpha_1})} = \mathcal{K}\ln(1/\beta_1)\frac{1-\beta_2}{s(1-\beta_1)}.$$

The rate of convergence to this neighbourhood, furthermore, is of order $\mathcal{O}(1/t)$.

Empirical observations

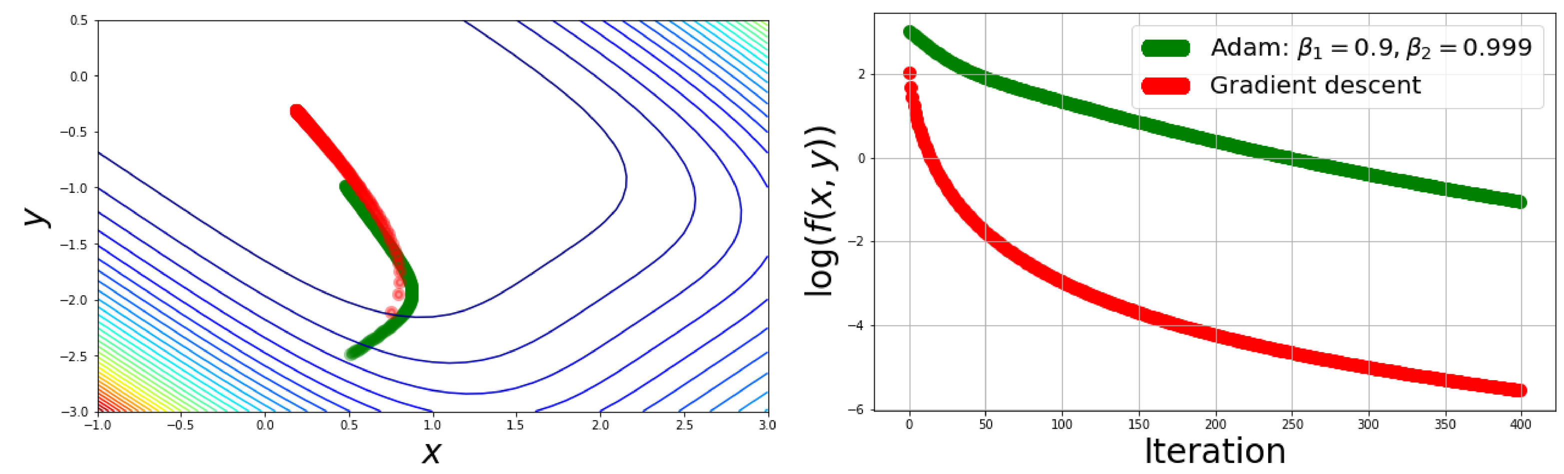

Figure 2: Comparison between gradient descent and ADAM for $f(x,y) = (x+y)^4 + (x/2-y/2)^4$. Gradient Descent outperforms ADAM in this example because β_1,β_2 are large and ADAM keeps memory of the past large gradients. Both trajectories start from the point $(0.5,-2.5)$.

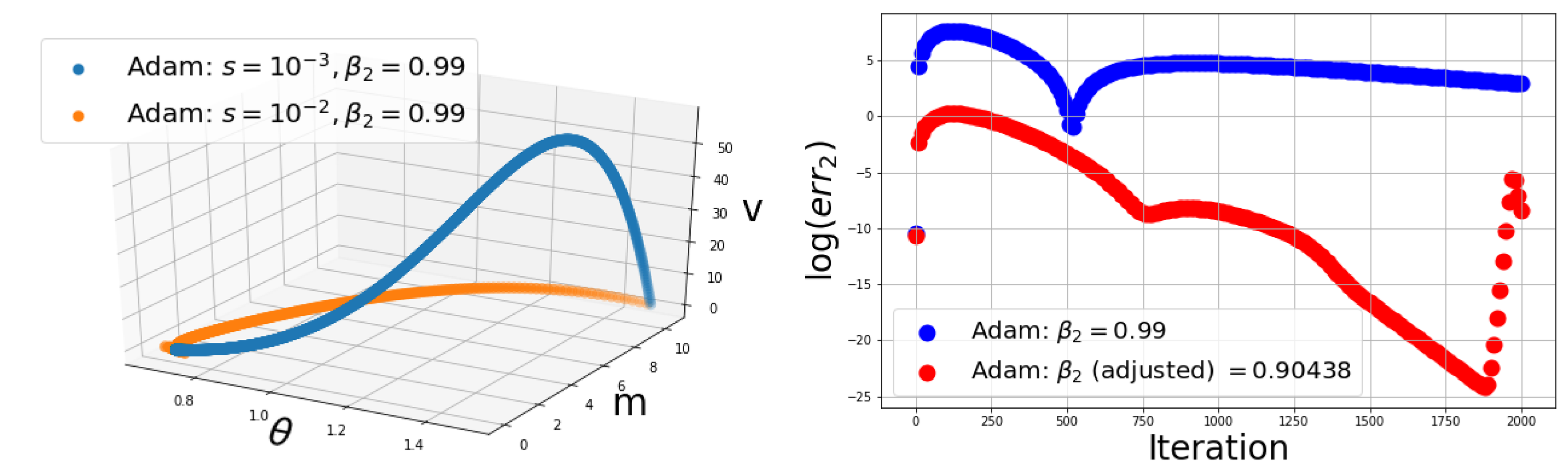

Figure 3: Fixing β_2 and changing the learning rate s lead to different dynamics. **right)** Trajectories of ADAM when only the learning rate is changed and β_1,β_2 are fixed. **left)** Comparison of the error between different trajectories.

Conclusions and Forthcoming Research

Conclusions:

- The convergence rate is nonlinear –in the sense that it depends on the variables– and depends on the history of the dynamics.
- With the standard choices of hyperparameters, adaptivity degrades the rate of convergence to the global minimum of a convex function compared to gradient descent.

Questions:

1. Does adaptivity reduces the variance (compared to SGD) and speed up the training for convex functional?
2. Is the fast training observed in deep learning induced by the specificity of the loss surface and common initialization scheme for the weights?

References

[1] Andre Belotto da Silva and Maxime Gazeau. A general system of differential equations to model first-order adaptive algorithms. *Journal of Machine Learning Research*, 21(129):1–42, 2020.

[2] S. Gadat, F. Panloup, and S. Saadane. Stochastic Heavy Ball. *ArXiv e-prints*, September 2016.

[3] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[4] D. Scieur, V. Roulet, F. Bach, and A. d'Aspremont. Integration methods and optimization algorithms. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1109–1118. Curran Associates, Inc., 2017.

[5] Bin Shi, Simon S. Du, Michael I. Jordan, and Weijie J. Su. Understanding the Acceleration Phenomenon via High-Resolution Differential Equations. *ArXiv e-prints*, October 2018.

[6] W. Su, S. Boyd, and E. J. Candes. A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights. *Journal of Machine Learning Research*, 17(153), 2016.