Convergence of the ADAM Algorithm from a Dynamical Systems Viewpoint

Anas Barakat and Pascal Bianchi

LTCI, Télécom Paris, Institut polytechnique de Paris, France

Problem

$$\min_{x} F(x) := \mathbb{E}(f(x,\xi)) \quad \text{w.r.t.} \quad x \in \mathbb{R}^d$$

- $f(\,.\,,\xi)$: **non-convex** differentiable
- ξ: r.v. with unknown distribution
- $(\xi_n : n \geq 1)$: iid copies of the r.v. ξ revealed online

The ADAM algorithm [1]

- Very popular in deep learning.
- Adaptive method.
- Less stepsize tuning needed.

Algorithm 1 ADAM $(\gamma, \alpha, \beta, \varepsilon)$

1: $x_0 \in \mathbb{R}^d, m_0 = 0, v_0 = 0, \gamma > 0, \varepsilon > 0, (\alpha,\beta) \in [0,1)^2$.
2: **for** $n \geq 1$ **do**
3: $m_n = \alpha m_{n-1} + (1-\alpha)\nabla f(x_{n-1}, \xi_n)$
4: $v_n = \beta v_{n-1} + (1-\beta)\nabla f(x_{n-1}, \xi_n)^2$
5: $\hat{m}_n = \frac{m_n}{1-\alpha^n}$
6: $\hat{v}_n = \frac{v_n}{1-\beta^n}$
7: $x_n = x_{n-1} - \gamma \frac{\hat{m}_n}{\varepsilon + \sqrt{\hat{v}_n}}$
8: **end for**

ODE method

Constant step $\gamma > 0$: no a.s convergence, stochastic approximation technique [2].

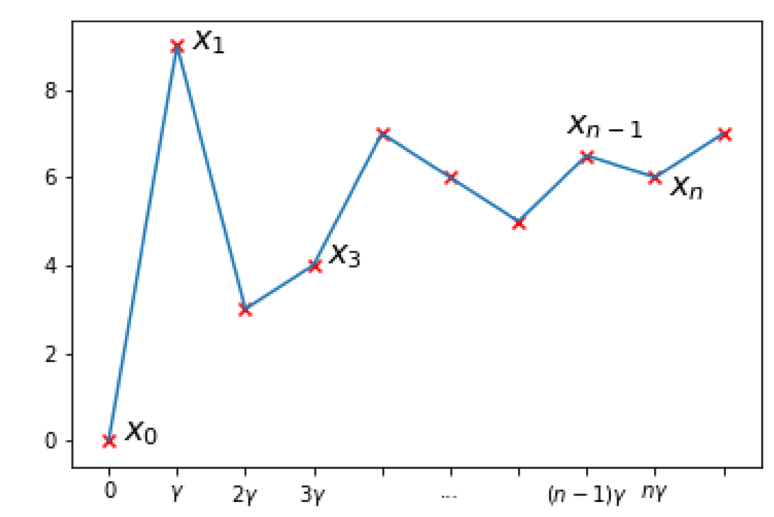
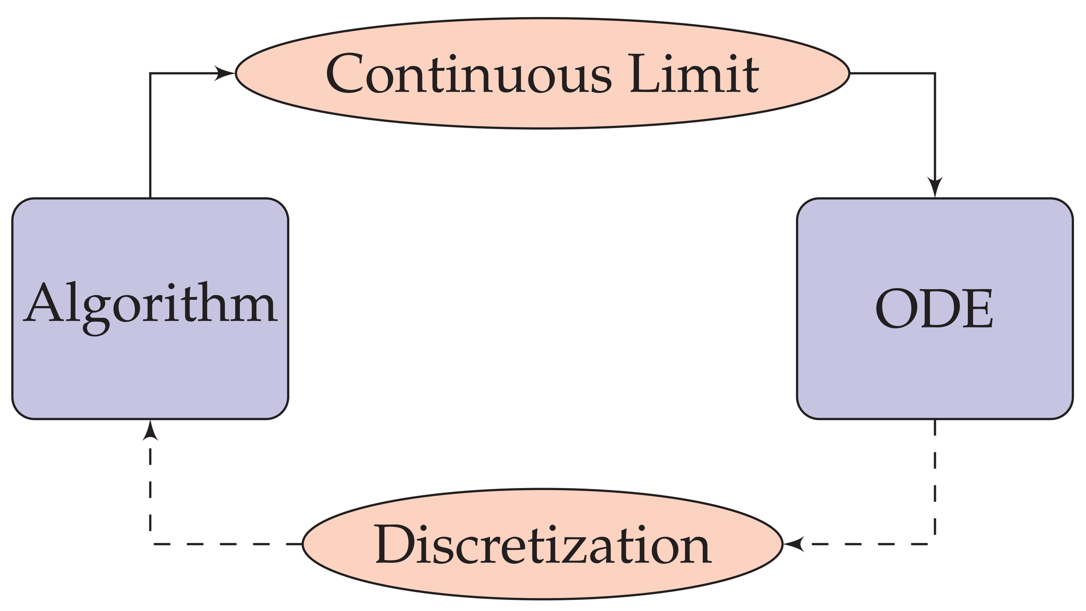


Figure 1: Piecewise linear interpolated process from ADAM iterates.

Piecewise linear interpolated process:

$$\mathsf{z}^\gamma(t) := z_n^\gamma + (z_{n+1}^\gamma - z_n^\gamma)\left(\frac{t - n\gamma}{\gamma}\right)$$

Approximation of a discrete time stochastic system by a deterministic one (ODE):



Contact Information

Email {anas.barakat,pascal.bianchi}@telecom-paristech.fr

Continuous-Time System

$$\dot{z}(t) = h(t, z(t)) \qquad \text{(ODE)}$$

where $h : (0, +\infty) \times \mathcal{Z}_+ \to \mathcal{Z}$ defined for all $t > 0$, all $z = (x, m, v)$ in $\mathcal{Z}_+$ by:

$$h(t,z) = \begin{pmatrix} -\frac{(1-e^{-at})^{-1} m}{\varepsilon + \sqrt{(1-e^{-bt})^{-1} v}} \\ a(\nabla F(x) - m) \\ b(S(x) - v) \end{pmatrix}$$

ADAM as a Heavy Ball with Friction (HBF).

$$c_1(t)\,\ddot{x}(t) + c_2(t)\dot{x}(t) + \nabla F(x(t)) = 0\,,$$

Particle mass and viscosity depend on time. 2nd order vs 1st order: faster convergence (acceleration), reduced oscillations, can go up and down along the graph of F.

ODE Analysis

- Existence, uniqueness and boundedness of a global ODE solution from $(x_0, 0, 0)$.
- Convergence of the solution to the stationary points of F.
- Key argument: Lyapunov function.

$$V(t,z) := F(x) + \frac{1}{2}\|m\|^2_{U(t,v)^{-1}}$$

Long run behavior

Theorem $\left(\mathsf{z}^\gamma \xrightarrow[\gamma \to 0]{weakly} z\right)$

Under mild assumptions, $\forall T > 0$, $\forall \delta > 0$,

$$\lim_{\gamma \downarrow 0} \mathbb{P}\left(\sup_{t \in [0,T]} \|\mathsf{z}^\gamma(t) - z(t)\| > \delta\right) = 0\,.$$

$$\lim_{\gamma \downarrow 0} \limsup_{n \to \infty} \mathbb{P}\left(\mathsf{d}(x_n^\gamma, \nabla F^{-1}(0)) > \delta\right) = 0\,.$$

Biased vs Unbiased ADAM

Only when unbiased, $F(x(t)) \leq F(x_0)$.

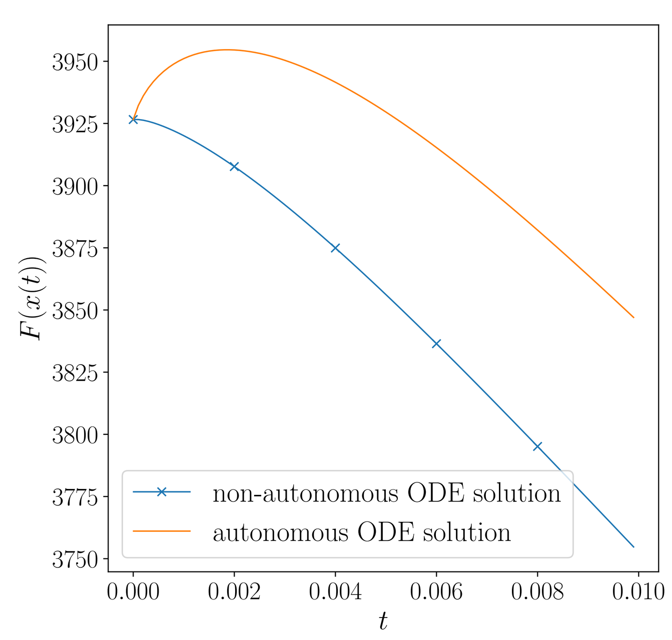


Figure 2: ADAM ODE solution vs autonomous ADAM ODE solution on a 100−dimensional Stochastic Quadratic Problem.

Numerical examples

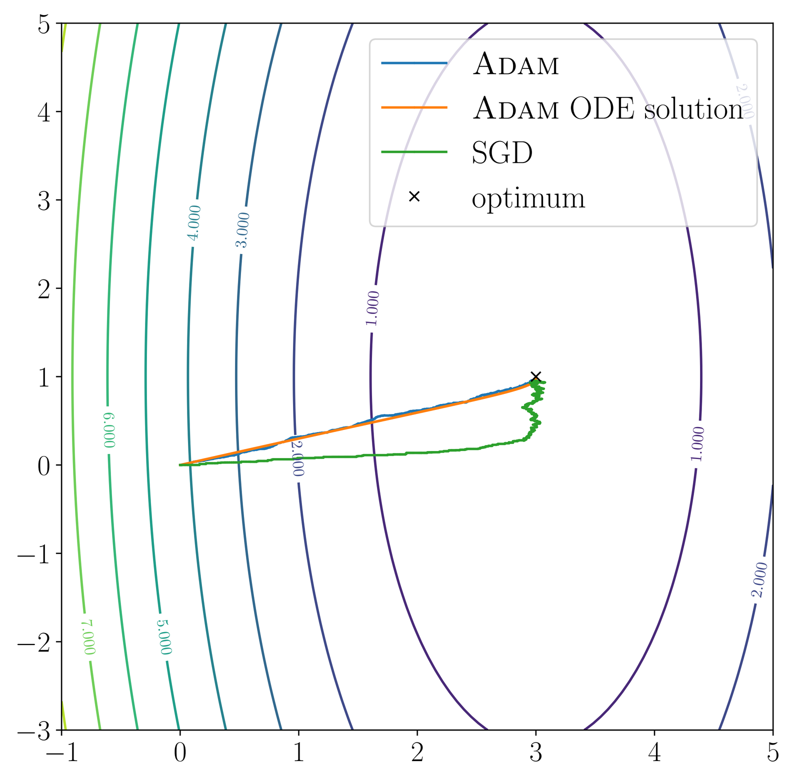


Figure 3: Convergence of ADAM and ODE solution to the optimum for a 2D linear regression.

Explicit Euler discretization scheme for ODEs.

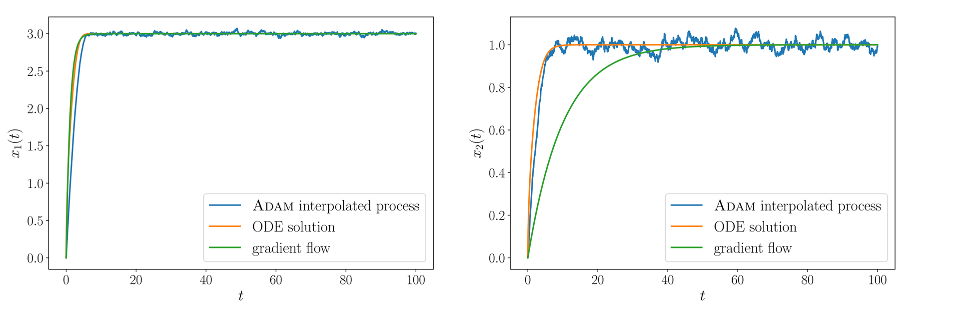


Figure 4: ADAM: interpolated process and solution to the ODE for a 2D linear regression.

Setting: 2D linear regression

$$Y = X x_1^* + (1 - X) x_2^* + \epsilon$$

where $(x_1^*, x_2^*) = (3, 1)$, $X \sim \mathcal{B}(p)$, $p \in (0,1)$

$$\xi = (X, Y)$$

$$f(\,.\,,\xi) := \tfrac{1}{2}\left(\left\langle \begin{pmatrix} X \\ 1 - X \end{pmatrix}, \cdot \right\rangle - Y\right)^2.$$

Conclusion and future work

1. Introduction of a continuous-time version of ADAM (non-autonomous ODE).
2. Existence, uniqueness and boundedness of the solution.
3. Weak convergence of the interpolated process to the ODE solution.
4. Convergence in the long run to the stationary points of the objective function.

Future works:

1. Stability of the ADAM Markov chain.
2. Rate of convergence of ADAM.

References

[1] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[2] H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 2003.

[3] 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*, 2016.

[4] P. Bianchi, W. Hachem, and A. Salim. Constant step stochastic approximations involving differential inclusions: Stability, long-run convergence and applications. *arXiv preprint*, 2016.