Convergence and Dynamical Behavior of the ADAM Algorithm for Non Convex Stochastic Optimization

Anas Barakat, Pascal Bianchi

LTCI, Télécom Paris, Institut polytechnique de Paris

JDSE September 12th 2019

Optimization in Deep Learning

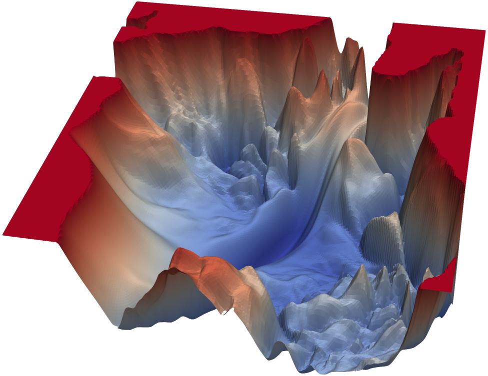


Figure 1: Visualization of a loss landscape (VGG-56 on CIFAR-10) https://www.cs.umd.edu/ tomg/projects/landscapes/

Li et al., Visualizing the Loss Landscape of Neural Nets, NeurIPS 2018

Problem statement

Problem

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

Assumptions

- $f(\,.\,,\xi)$: **nonconvex** differentiable function
- $(\xi_n : n \geq 1)$: iid copies of r.v ξ revealed online

Solution ?

Stochastic Gradient Descent (SGD)

$$x_{n+1} = x_n - \gamma_n \nabla f(x_n, \xi_{n+1})$$

- Limitations
  - learning rate choice
  - common learning rate for all the coordinates

Adaptive Algorithms

standard SGD

$$x_{n+1,i} = x_{n,i} - \gamma_n \nabla f(x_n, \xi_{n+1})_i$$

$$\gamma_n := \gamma \quad \text{ou} \quad \gamma_n := \frac{1}{\sqrt{n}}, n \geq 1$$

Adaptive Algorithms

$$x_{n+1,i} = x_{n,i} - \gamma_{n,i}\, g_{n,i}$$

$$\gamma_{n,i} := \Psi(\nabla f(x_p, \xi_{p+1})_i, p \leq n)$$

ADAM Algorithm

[Kingma and Ba, 2015]

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

$x_0 \in \mathbb{R}^d, m_0 = 0,\ v_0 = 0,\ \gamma > 0,\ \varepsilon > 0,\ (\alpha, \beta) \in [0,1)^2.$

for $n \geq 1$ **do**

$\quad m_n = \alpha m_{n-1} + (1-\alpha)\nabla f(x_{n-1}, \xi_n)$

$\quad v_n = \beta v_{n-1} + (1-\beta)\nabla f(x_{n-1}, \xi_n)^2$

$\quad \hat{m}_n = \frac{m_n}{1-\alpha^n}$

$\quad \hat{v}_n = \frac{v_n}{1-\beta^n}$

$\quad x_n = x_{n-1} - \frac{\gamma}{\varepsilon + \sqrt{\hat{v}_n}} \hat{m}_n$

end for

Assumptions and asymptotic regime

- Regime : **constant step size** $\gamma > 0$.

Assumptions on f

- regularity assumptions on f.
- $F : x \mapsto \mathbb{E}(f(x, \xi))$ coercive.

- Assumptions on hyperparameters: compatible with practical implementation.

From Discrete to Continuous Time

Continous Time : Dynamical System Analysis

Discrete Time : Convergence of ADAM

From Discrete to Continuous Time

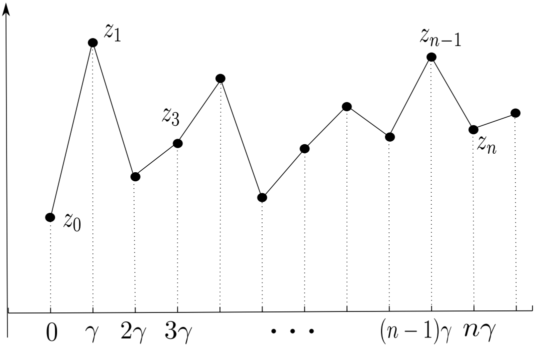
Continous Time : Dynamical System Analysis

Discrete Time : Convergence of ADAM

The ODE method

[Ljung, 1977, Kushner and Yin, 2003]

$\mathsf{z}^\gamma(t)$ interpolated from $z_n^\gamma = (x_n^\gamma, m_n^\gamma, v_n^\gamma)$



Towards Continuous Time

$$z_n^\gamma := z_{n-1}^\gamma + \gamma H_\gamma(n, z_{n-1}^\gamma, \xi_n)\,,$$

For all $\gamma > 0$, for all z,

$$\begin{aligned} h_\gamma(n,z) &:= \mathbb{E}(H_\gamma(n, z_{n-1}^\gamma, \xi_n)|\mathcal{F}_{n-1}) \\ \Delta_n^\gamma &:= H_\gamma(n, z_{n-1}^\gamma, \xi_n) - h_\gamma(n, z_{n-1}^\gamma) \end{aligned}$$

Decomposition in mean field + martingale noise

$$\text{For}\, \gamma > 0, \quad z_n^\gamma = z_{n-1}^\gamma + \gamma h_\gamma(n, z_{n-1}^\gamma) + \gamma \Delta_n^\gamma\,,$$

$$\frac{z_n^\gamma - z_{n-1}^\gamma}{\gamma} = h_\gamma(n, z_{n-1}^\gamma) + \Delta_n^\gamma$$

$\mathcal{F}_n = \sigma(\xi_1, \ldots, \xi_n)$

From Discrete to Continuous Time

Continous Time : Dynamical System Analysis

Discrete Time : Convergence of ADAM

Continuous Time System

similar approach to [Su et al., 2016]

Non autonomous ODE

Si $z(t) = (x(t), m(t), v(t))$,

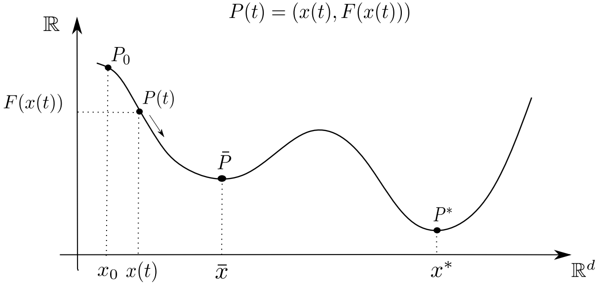
$$\dot{z}(t) = h(t, z(t)) \tag{ODE}$$

Theorem

Existence, uniqueness and boundedness of a global solution to the ODE from $(x_0, 0, 0)$.

Mechanical Interpretation - Heavy Ball with Friction

[Attouch et al., 2000, Cabot et al., 2009, Gadat et al., 2018]



- Gravity force (potentiel F).
- Force of friction of viscous type: $-\lambda \dot{x}(t)$ (damping).
- Reaction of the surface $\Sigma = Graph(F)$.

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

ADAM as a Heavy Ball with Friction (HBF)

"Generalized" HBF

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

- **Generalized HBF :**
  - Time dependent particle mass
  - Time dependent viscosity
- **Why HBF ?**
  - 2nd vs 1st order: acceleration (even if oscillations).
  - Escaping local traps (saddle points)

Convergence to stationary points

Theorem (Convergence)

$$\lim_{t\to\infty} \mathrm{d}(x(t), \nabla F^{-1}(\{0\})) = 0\,.$$

Key argument : Lyapunov function for the ODE

- Definition :

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

- Interpretation : mechanical energy of the dynamical system
- Lemma : $t \mapsto V(t, z(t))$ is decreasing on $(0, +\infty)$.

From Discrete to Continuous Time

Continous Time : Dynamical System Analysis

Discrete Time : Convergence of ADAM

Weak convergence of the interpolated process towards the ODE solution

Techniques [Benaïm and Schreiber, 2000]

Moment assumption - Noise control

For every compact set $K \subset \mathbb{R}^d$, there exists $r_K > 0$ s.t.

$$\sup_{x \in K} \mathbb{E}(\|\nabla f(x,\xi)\|^{2+r_K}) < \infty \, .$$

Theorem

Under previous assumptions and the **moment assumption**,

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

Simulations

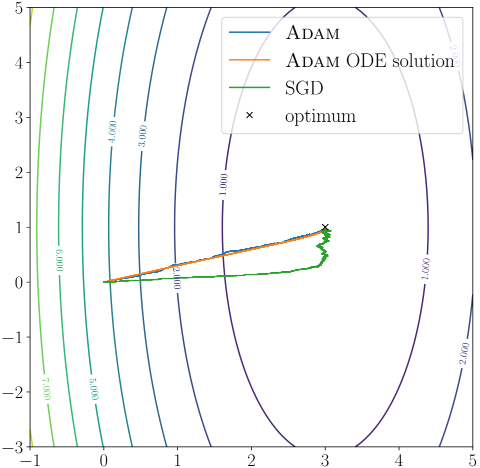


Figure 2: Convergence of ADAM and the ODE solution towards the optimum for a 2D linear regression

Long run convergence of the ADAM iterates

Techniques [Fort and Pagès, 1999, Bianchi et al., 2019]

- No a.s convergence : regime $n \to \infty$ then $\gamma \to 0$

Theorem (ergodic convergence of the ADAM iterates)

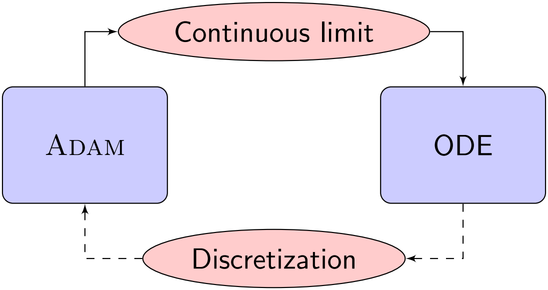
Let $x_0 \in \mathbb{R}^d$, $\gamma > 0$, $(z_n^\gamma : n \in \mathbb{N})$, $z_0^\gamma = (x_0, 0, 0)$. Under the same assumptions and :

- **Stability assumption:** $\sup_{n,\gamma} \mathbb{E}\|z_n^\gamma\| < \infty\,.$

Then, for all $\delta > 0$,

$$\lim_{\gamma \downarrow 0} \limsup_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{P}(\mathrm{d}(x_n^\gamma, \nabla F^{-1}(\{0\})) > \delta) = 0\,. \tag{1}$$

Conclusion



Thank you for your attention

For more details: submitted article, available on arXiv.

AB, P. Bianchi. *Convergence and Dynamical Behavior of the ADAM Algorithm for Non Convex Stochastic Optimization.*

Mean Field

where for all $t > 0$, all $z = (x, m, v)$:

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

$S : x \mapsto \mathbb{E}(\nabla f(x,\xi)^2)$ s.t. $\forall x \in \mathbb{R}^d$, $S(x) > 0$.

Simulations

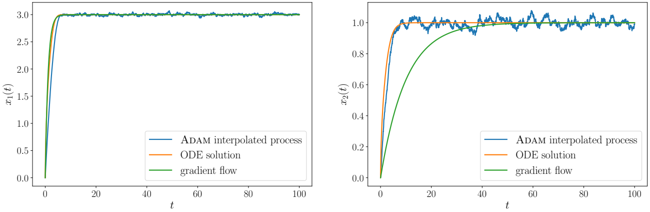


Figure 3: ADAM: interpolated process and solution to the ODE for a 2D linear regression.

2D linear regression

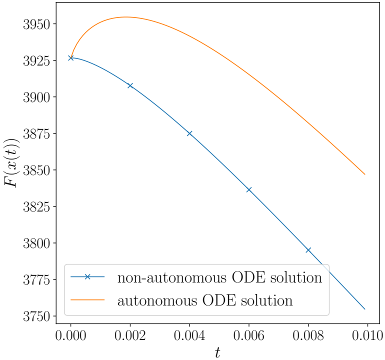
$Y = X\, x_1^* + (1 - X)\, x_2^* + \epsilon$ with $(x_1^*, x_2^*) = (3, 1)$.

$\xi = (X, Y)$ with $X \sim \mathcal{B}(p)$, $p \in (0, 1)$.

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

Biased vs Unbiased ADAM

With debiasing steps, $F(x(t)) \leq F(x_0)$.



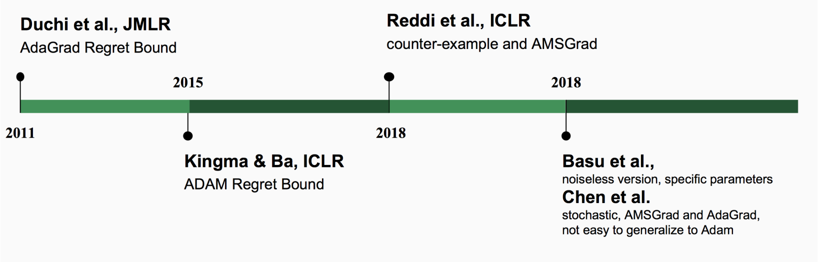
Algorithm 2 ADAM $(\gamma, \alpha, \beta, \varepsilon)$

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: $\quad m_n = \alpha m_{n-1} + (1-\alpha)\nabla f(x_{n-1}, \xi_n)$
4: $\quad v_n = \beta v_{n-1} + (1-\beta)\nabla f(x_{n-1}, \xi_n)^2$
5: $\quad \hat{m}_n = \frac{m_n}{1-\alpha^n}$
6: $\quad \hat{v}_n = \frac{v_n}{1-\beta^n}$
7: $\quad x_n = x_{n-1} - \frac{\gamma}{\varepsilon + \sqrt{\hat{v}_n}} \hat{m}_n$
8: **end for**

Autonomous/Non autonomous ODE solutions for a 100-dimensional Stochastic Quadratic Problem

Literature review

ADAM: Theoretical results



Literature review

ADAM: theoretical results

- $\mathcal{O}(\frac{1}{\sqrt{T}})$ average regret bound in nonconvex setting.
- counter-example: average regret does not converge to 0.
- AmsGrad: variant of ADAM
- noiseless version of ADAM (deterministic f):
  - small gradient norm for some upperbounded unknown instant
  - specific values of the ADAM hyperparameters
- similar result in the stochastic setting for a general class of adaptive algorithms
  - results stated for AmsGrad and AdaGrad
  - generalization to ADAM subject to conditions which are not easy to verify.

References I

Attouch, H., Goudou, X., and Redont, P. (2000).
The heavy ball with friction method, i. the continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system.
Communications in Contemporary Mathematics, 2(01):1–34.

Benaïm, M. and Schreiber, S. J. (2000).
Ergodic properties of weak asymptotic pseudotrajectories for semiflows.
J. Dynam. Differential Equations, 12(3):579–598.

Bianchi, P., Hachem, W., and Salim, A. (2019).
Constant step stochastic approximations involving differential inclusions: Stability, long-run convergence and applications.
Stochastics, 91(2):288–320.

References II

Cabot, A., Engler, H., and Gadat, S. (2009).
On the long time behavior of second order differential equations with asymptotically small dissipation.
Transactions of the American Mathematical Society, 361(11):5983–6017.

Fort, J.-C. and Pagès, G. (1999).
Asymptotic behavior of a Markovian stochastic algorithm with constant step.
SIAM J. Control Optim., 37(5):1456–1482 (electronic).

Gadat, S., Panloup, F., and Saadane, S. (2018).
Stochastic heavy ball.
Electronic Journal of Statistics, 12(1):461–529.

References III

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

Kushner, H. J. and Yin, G. G. (2003).
Stochastic approximation and recursive algorithms and applications, volume 35 of *Applications of Mathematics (New York)*.
Springer-Verlag, New York, second edition.
Stochastic Modelling and Applied Probability.

Ljung, L. (1977).
Analysis of recursive stochastic algorithms.
IEEE transactions on automatic control, 22(4):551–575.

References IV

Su, W., Boyd, S., and Candès, E. J. (2016).
A differential equation for modeling nesterov's accelerated gradient method: Theory and insights.
Journal of Machine Learning Research, 17(153):1–43.