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

Anas Barakat **Pascal Bianchi**

LTCI, Télécom Paris, Institut polytechnique de Paris

2^{nd} **Symposium on Machine Learning and Dynamical Systems, Fields Institute**

September 21^{st} 2020

Outline

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
  ($+$ some regularity assumptions to define F, ∇F)
- $(\xi_n : n \geq 1)$: iid copies of r.v ξ revealed online

ADAM Algorithm

[Kingma and Ba, 2015]

- More than 52000 citations !

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**

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

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

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

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

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

end for

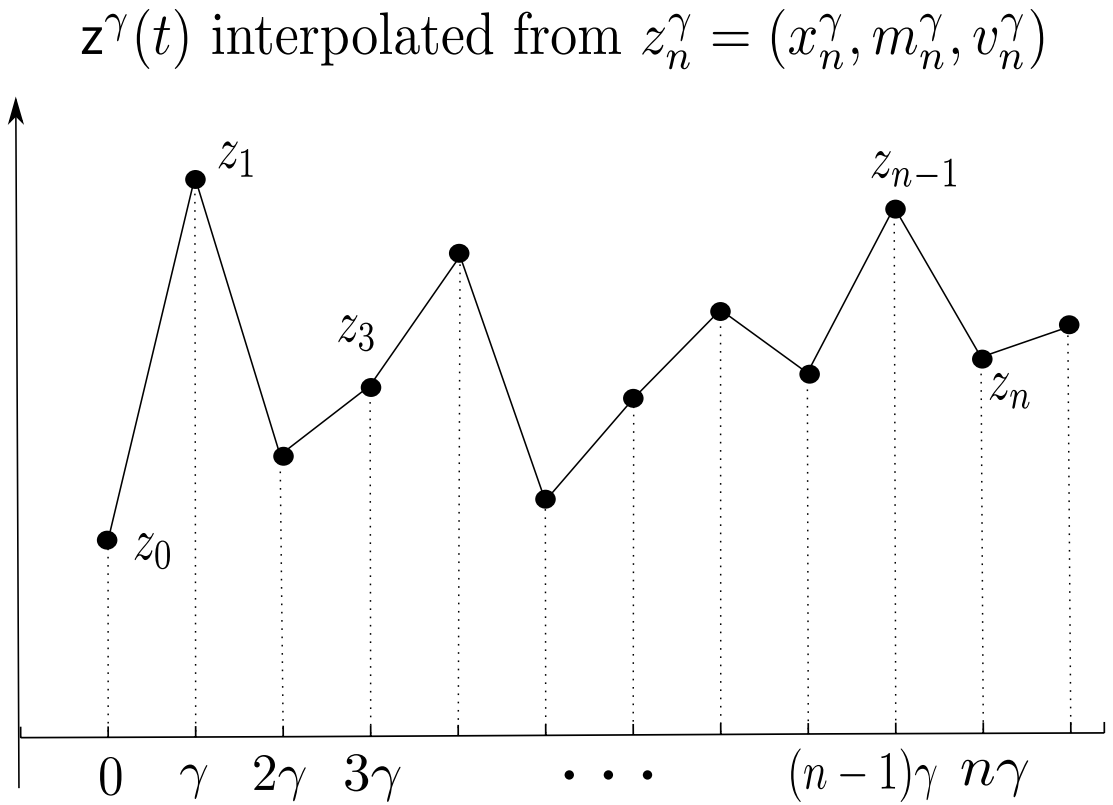
Set ($\mathcal{H}_{model}$) of Assumptions

- **Model**: regularity assumptions on f (non-convex, diff., ...).
- Coercivity : $F : x \mapsto \mathbb{E}(f(x, \xi))$ coercive.
- $\forall x \in \mathbb{R}^d$, $S(x) > 0$ where $S : x \mapsto \mathbb{E}(\nabla f(x, \xi)^{\odot 2})$.
- **Hyperparameters**: verified in practice ('α,β close to 1').
- **Noise:** iid sequence (ξ_n).

I. From Discrete to Continuous-Time ADAM

The ODE method

[Ljung, 1977, Kushner and Yin, 2003]



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,

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

Decomposition in mean field + martingale noise

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

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

II. Continuous-Time Adam

Continuous Time System

Non autonomous ODE

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

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

Theorem

Existence, uniqueness and boundedness of a global solution to the ODE from $(x_0, 0, 0)$ under $(\mathcal{H}_{model})$.

Convergence to stationary points

Theorem (Convergence)

Under $(\mathcal{H}_{model})$,

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

Key argument : Lyapunov function for the ODE

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

III. ADAM with decreasing stepsizes

$\rightarrow$ Link between asymptotic behavior of (z_n) and ODE ?

In this Talk, ADAM with decreasing stepsizes

Algorithm 2 ADAM $(((\gamma_n, \alpha_n, \beta_n) : n \in \mathbb{N}^*), \varepsilon)$.

Initialization: $x_0 \in \mathbb{R}^d$, $m_0 = 0$, $v_0 = 0$, $r_0 = \bar{r}_0 = 0$.
for $n = 1$ **to** n_{iter} **do**
  $m_n = \alpha_n m_{n-1} + (1 - \alpha_n) \nabla f(x_{n-1}, \xi_n)$
  $v_n = \beta_n v_{n-1} + (1 - \beta_n) \nabla f(x_{n-1}, \xi_n)^{\odot 2}$
  $r_n = \alpha_n r_{n-1} + (1 - \alpha_n) \longrightarrow r_n = 1 - \prod_{i=1}^{n} \alpha_i$
  $\bar{r}_n = \beta_n \bar{r}_{n-1} + (1 - \beta_n) \longrightarrow \bar{r}_n = 1 - \prod_{i=1}^{n} \beta_i$
  $\hat{m}_n = m_n / r_n$ {bias correction step}
  $\hat{v}_n = v_n / \bar{r}_n$ {bias correction step}
  $x_n = x_{n-1} - \frac{\gamma_n}{\varepsilon + \sqrt{\hat{v}_n}} \hat{m}_n$.
end for

- Define $(\mathcal{H}'_{model}) = (\mathcal{H}_{model})$ with α_n, β_n instead of α, β i.e. $\frac{1-\alpha_n}{\gamma_n} \to a$ and $\frac{1-\beta_n}{\gamma_n} \to b$.

ODE method in Stochastic Approximation

[Ljung, 1977, Kushner and Yin, 2003, Duflo, 1997, Benaïm, 1999, Borkar, 2008] ...

Robbins Monro scheme

$$\mathbf{z_{n+1}} = \mathbf{z_n} + \gamma_{\mathbf{n+1}} \underbrace{g}_{\textbf{mean field}}(\mathbf{z_n}) + \gamma_{\mathbf{n+1}} \underbrace{\eta_{n+1}}_{\textbf{noise}} + \gamma_{\mathbf{n+1}} \underbrace{b_{n+1}}_{\textbf{bias}},$$

- Noisy discretization of $\dot{\mathbf{z}}(\mathbf{t}) = \mathbf{g}(\mathbf{z}(\mathbf{t}))$.
- Informally:
  if $\gamma_n \to 0$, noise washes out and " $\lim_{n\to\infty} z_n = \lim_{t\to\infty} z(t)$".

Almost sure convergence

- RM algorithm :

$$z_{n+1} = z_n + \gamma_{n+1} \underbrace{h_\infty}_{\text{mean field}}(z_n) + \gamma_{n+1} \underbrace{\eta_{n+1}}_{\text{noise}} + \gamma_{n+1} \underbrace{b_{n+1}}_{\text{bias}},$$

Assumptions $(\mathcal{H}_{as-cv})$

- $\sum_n \gamma_n = +\infty$ and $\sum_n \gamma_n^2 < +\infty$,
- $\forall$ compact $K \subset \mathbb{R}^d$, $\sup_{x \in K} \mathbb{E}(\|\nabla f(x, \xi)\|^4) < \infty$.
- $\sup_{n \in \mathbb{N}} \|z_n\| < +\infty$ a.s. (proven separately)

Theorem (Almost Sure Convergence)

Under $(\mathcal{H}'_{model}) + (\mathcal{H}_{as-cv})$, w.p.1,

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

Towards a Central Limit Theorem

Our algorithm: $z_{n+1} = z_n + \gamma_{n+1} h_\infty(z_n) + \gamma_{n+1} b_{n+1} + \gamma_{n+1}\eta_{n+1}$,

Rescaled algorithm: $Z_{n+1} = \frac{z_{n+1}-z^*}{\sqrt{\gamma_{n+1}}}$; $\gamma_n = n^{-\kappa}, \kappa \in (0,1]$

Strongly disturbed algorithm:

$$Z_{n+1} = (I + \gamma_{n+1} \underbrace{\bar{H}}_{\frac{1}{2\gamma_0}\mathbb{1}_{\kappa=1} I + \nabla h_\infty(z^*)})Z_n + \gamma_{n+1}\bar{b}_{n+1} + \sqrt{\gamma_{n+1}}\eta_{n+1}$$

Assumptions ($\mathcal{H}_{CLT}$)

- Let $x^* \in \nabla F^{-1}(\{0\})$. $\exists$ neighborhood $\mathcal{V}$ of x^* s.t.
  - *i)* F is $\mathcal{C}^2$ on $\mathcal{V}$, and $\nabla^2 F(x^*)$ is positive definite.
  - *ii)* S is $\mathcal{C}^1$ on $\mathcal{V}$.

- $\exists \kappa \in (0, 1]$, $\gamma_0 > 0$, s.t. $\gamma_n = \gamma_0/(n+1)^\kappa$. If $\kappa = 1$, $\gamma_0 > \frac{1}{2L}$.

- $\forall$ compact $K \subset \mathbb{R}^d$, $\exists p_K > 4$, $\sup_{x \in K} \mathbb{E}(\|\nabla f(x, \xi)\|^{p_K}) < \infty$.

Theorem (CLT)

Assume $\mathbb{P}(z_n \to z^*) > 0$.
Under $(\mathcal{H}'_{model}) + (\mathcal{H}_{CLT})$, given the event $\{z_n \to z^*\}$,

$$\frac{z_n - z^*}{\sqrt{\gamma_n}} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \Sigma)\,.$$

(on $\mathbb{R}^{3d}$) with a covariance matrix Σ s.t.

$$(H + \zeta I_{3d})\,\Sigma + \Sigma\left(H^T + \zeta I_{3d}\right) = -Q\,.$$

where $\zeta := 0$ if $0 < \kappa < 1$ and $\zeta := \frac{1}{2\gamma_0}$ if $\kappa = 1$.

Contributions and future work

1. Introduction and analysis of a continuous-time ADAM.

2. Almost sure convergence of ADAM with decreasing stepsizes.

Avoidance of traps?
What about the non-differentiable case?

More on : **https://anasbarakat.github.io**, on arxiv

and on **my pre-recorded talk**

References I

Basu, A., De, S., Mukherjee, A., and Ullah, E. (2018).

Convergence guarantees for rmsprop and adam in non-convex optimization and their comparison to nesterov acceleration on autoencoders.

arXiv preprint arXiv:1807.06766.

Benaïm, M. (1999).

Dynamics of stochastic approximation algorithms.

In *Séminaire de Probabilités, XXXIII*, volume 1709 of *Lecture Notes in Math.*, pages 1–68. Springer, Berlin.

Borkar, V. S. (2008).

Stochastic approximation. A dynamical systems viewpoint.

Cambridge University Press, Cambridge; Hindustan Book Agency, New Delhi.

Chen, X., Liu, S., Sun, R., and Hong, M. (2019).

On the convergence of a class of adam-type algorithms for non-convex optimization.

In *International Conference on Learning Representations*.

Duchi, J., Hazan, E., and Singer, Y. (2011).

Adaptive subgradient methods for online learning and stochastic optimization.

Journal of Machine Learning Research, 12(Jul):2121–2159.

Duflo, M. (1997).

Random iterative models, volume 34 of *Applications of Mathematics (New York)*.

Springer-Verlag, Berlin.

Kingma, D. P. and Ba, J. (2015).

Adam: A method for stochastic optimization.

In *International Conference on Learning Representations*.

References II

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.

Reddi, S. J., Kale, S., and Kumar, S. (2018).

On the convergence of adam and beyond.
In *International Conference on Learning Representations*.

Zaheer, M., Reddi, S. J., Sachan, D., Kale, S., and Kumar, S. (2018).

Adaptive methods for nonconvex optimization.
In *Advances in Neural Information Processing Systems*, pages 9793–9803.

Related Work