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

Anas Barakat

Joint work with Pascal Bianchi

LTCI, Télécom Paris, Institut polytechnique de Paris

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

August-September 2020

Outline

ADAM algorithm

Continuous-Time ADAM

- From discrete to continuous time
- Existence, uniqueness, convergence
- Łojasiewicz convergence rates

ADAM with constant stepsize

ADAM with decreasing stepsizes

- Almost sure convergence
- Stability
- Central Limit Theorem

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

Solution ?

Stochastic Gradient Descent (SGD)

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

- Limitations
  - learning rate tuning
  - 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

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

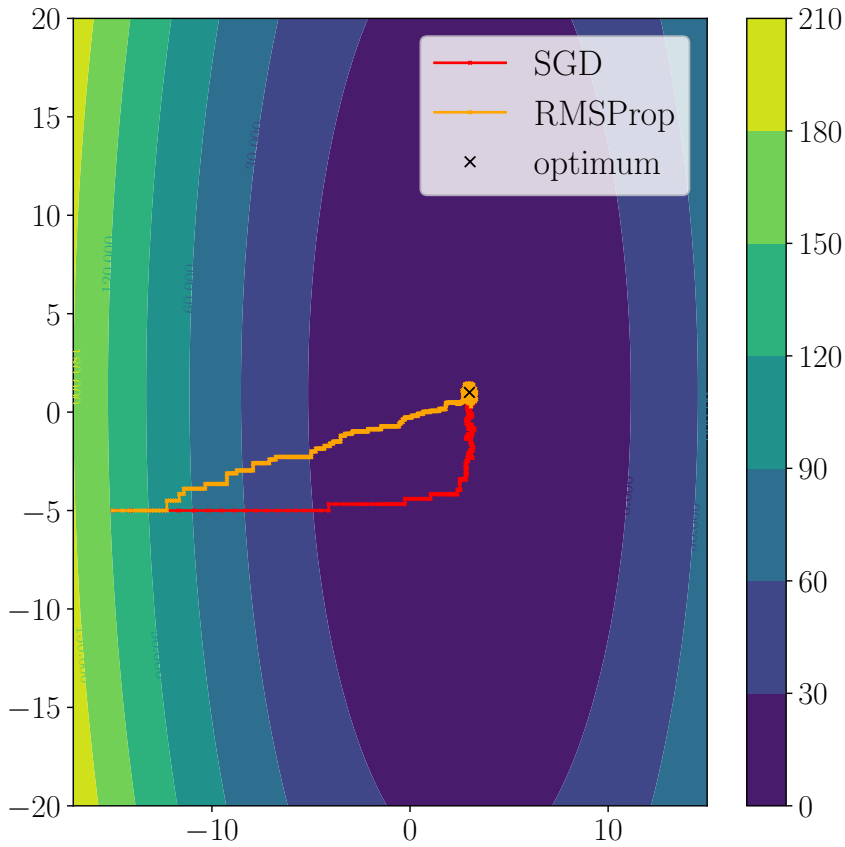
$$\gamma_{n,i} := \Psi(\nabla f(x_p, \xi_{p+1})_i, p \leq n)$$

RMSProp : coordinatewise stepsize

RMSProp

$$x_{n+1,i} = x_{n,i} - \frac{\gamma_0}{\varepsilon+\sqrt{v_{n,i}}} \nabla f(x_n, \xi_{n+1})_i$$

$$\begin{cases} x_{n+1} &= x_n - \frac{\gamma_0}{\varepsilon+\sqrt{v_n}} \nabla f(x_n, \xi_{n+1}) \\ v_{n+1} &= \beta v_n + (1-\beta) \nabla f(x_n, \xi_{n+1})^{\odot 2} \end{cases}$$

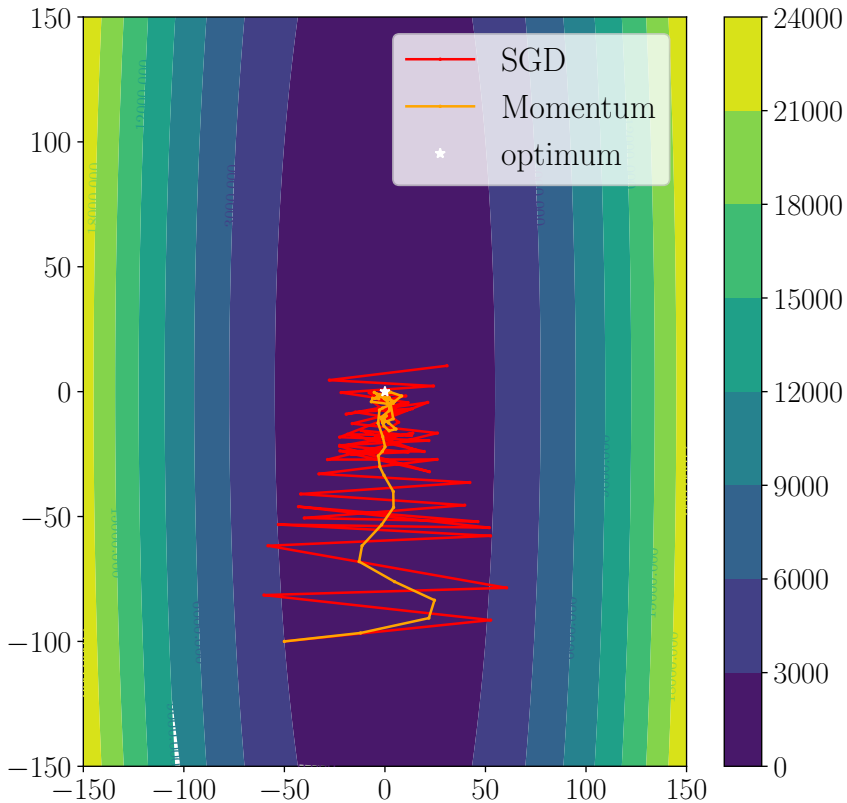


Momentum : (hoping) for acceleration

Momentum (aka Heavy Ball)

$$\begin{cases} m_n & = \alpha m_{n-1} + (1-\alpha)\nabla f(x_{n-1}, \xi_n) \\ x_{n+1} & = x_n - \gamma m_n \end{cases}$$

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



Debiasing

- During first iterations, m_n is a biased estimate of $\nabla F(x_0)$

$$\begin{aligned} m_n &= \alpha m_{n-1} + (1-\alpha)\nabla f(x_{n-1}, \xi_n) \\ &= (1-\alpha)\sum_{k=1}^{n} \alpha^{n-k} \nabla F(x_{k-1}, \xi_k) + \text{noise} \\ &\simeq (1-\alpha^n)\nabla F(x_0) + \text{noise} \end{aligned}$$

Debiasing step

$$\hat{m}_n = \frac{m_n}{1-\alpha^n} = \frac{1-\alpha}{1-\alpha^n} \sum_{k=0}^{n-1} \alpha^k \nabla f(x_k, \xi_{k+1})\,.$$

ADAM Algorithm

[Kingma and Ba, 2015]

- 51109 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**

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

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

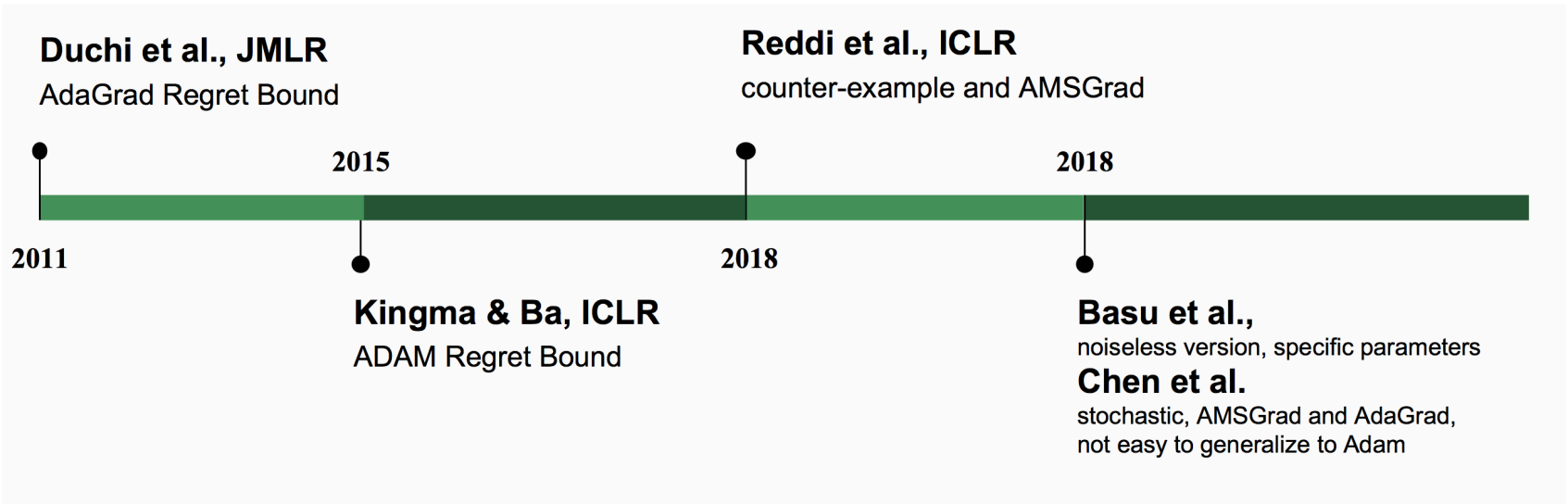
  $\hat{m}_n = \frac{m_n}{1-\alpha^n}$

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

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

end for

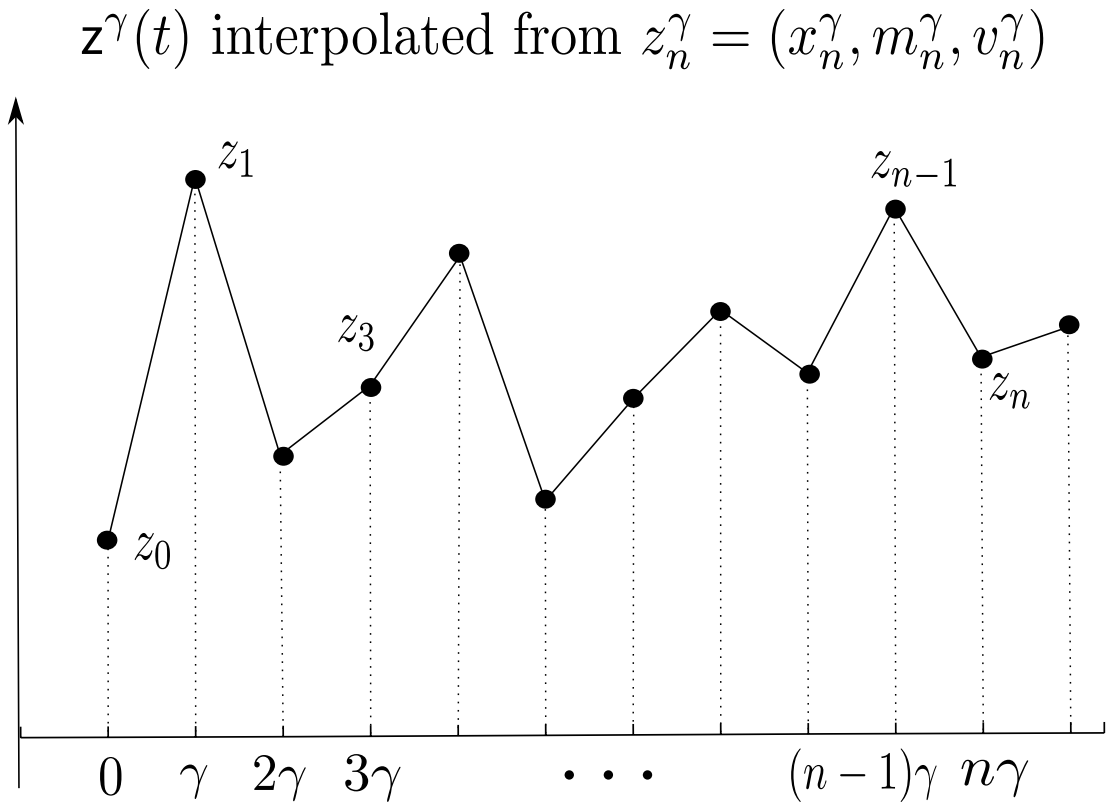
Related Work



I. 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

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

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

$$\begin{cases} \dfrac{m_n - m_{n-1}}{\gamma} & = \underbrace{\dfrac{1-\alpha(\gamma)}{\gamma}}_{HYP:\to a \text{ as } \gamma\to 0} (\nabla F(x_{n-1}) - m_{n-1}) + \frac{1-\alpha(\gamma)}{\gamma}(\nabla f(x_{n-1},\xi_n) - \nabla F(x_{n-1})) \\ \dfrac{v_n - v_{n-1}}{\gamma} & = \frac{1-\beta(\gamma)}{\gamma}(S(x_{n-1}) - v_{n-1}) + \frac{1-\beta(\gamma)}{\gamma}(\nabla f(x_{n-1},\xi_n)^{\odot 2} - S(x_{n-1})) \\ \dfrac{x_n - x_{n-1}}{\gamma} & = -\frac{(1-\alpha^n)^{-1} m_n}{\varepsilon + \sqrt{(1-\beta^n)^{-1} v_n}} \end{cases}$$

$$\begin{pmatrix} \dot{x} \\ \dot{m} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} -\dfrac{(1-e^{-at})^{-1} m}{\varepsilon + \sqrt{(1-e^{-bt})^{-1} v}} \\ a(\nabla F(x) - m) \\ b(S(x) - v) \end{pmatrix} := h(t,z) \quad \text{for } t > 0, z = (x,m,v)\,.$$

$$n = \left\lfloor \frac{t}{\gamma} \right\rfloor; \qquad S : x \mapsto \mathbb{E}(\nabla f(x,\xi)^{\odot 2})$$

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).

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

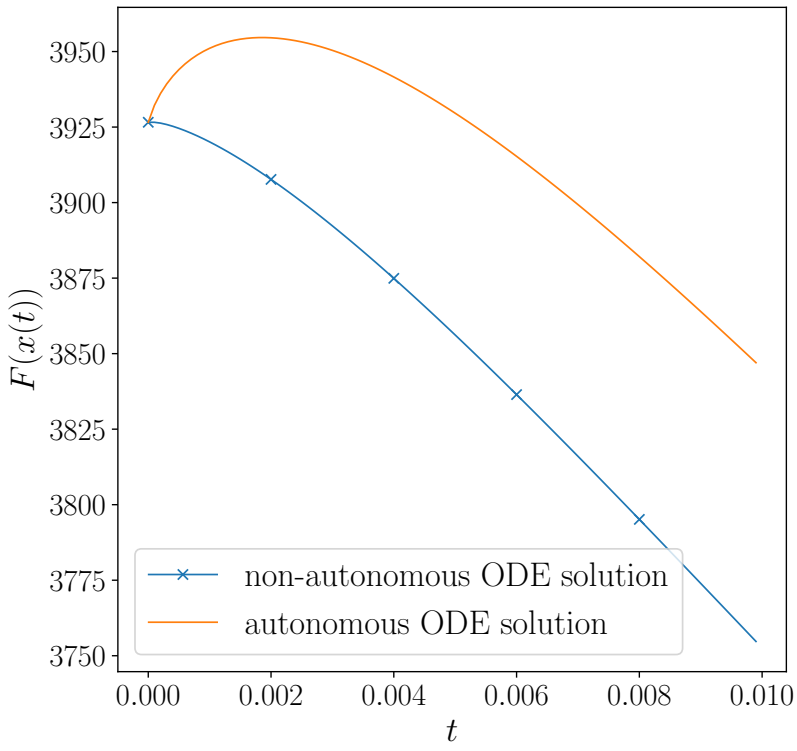
Existence, uniqueness and boundedness of a global solution to the ODE from $(x_0, 0, 0)$ under $(\mathcal{H}_{model})$.

Remark :

- $h(\cdot, z)$ undefined at 0
- $h(t, \cdot)$ not locally Lipschitz continuous.

Biased vs Unbiased ADAM

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



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

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

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

- Definition :

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

- Lemma : $t \mapsto V(t, z(t))$ is nonincreasing on $(0, +\infty)$.

Łojasiewicz inequality

[Łojasiewicz, 1963, Attouch and Bolte, 2009]

Assumption (Łojasiewicz property)

$\forall x^* \in \nabla F^{-1}(\{0\})$, $\exists c > 0\,, \sigma > 0\,, \theta \in (0, \frac{1}{2}]$ s.t.

$$\forall x \in \mathbb{R}^d \text{ s.t } \|x - x^*\| \leq \sigma\,, \quad \|\nabla F(x)\| \geq c|F(x) - F(x^*)|^{1-\theta}\,.$$

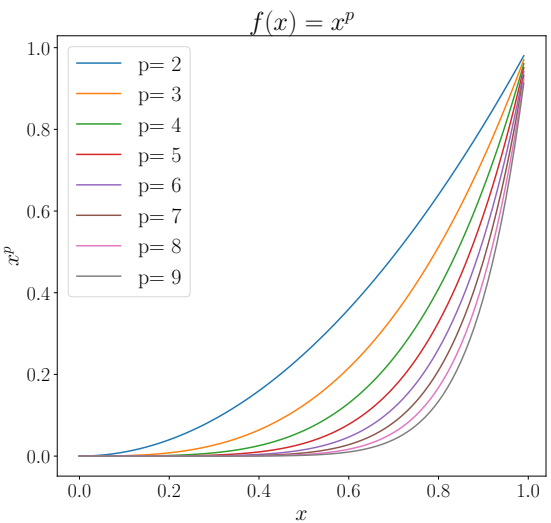
- F becomes **sharp** under a reparameterization of its values w.l.o.g. if $F(x^*) = 0$, for suitable x^* points, if $\theta \in (0, 1]$, $\varphi(s) = \frac{1}{c\theta} s^\theta$

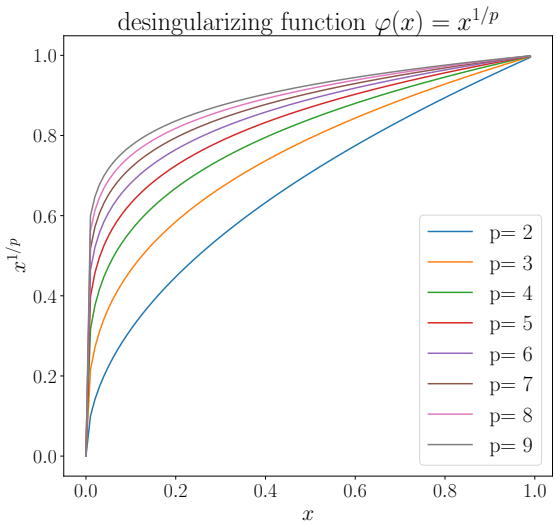
$$\|\nabla(\varphi \circ F)(x)\| \geq 1.$$

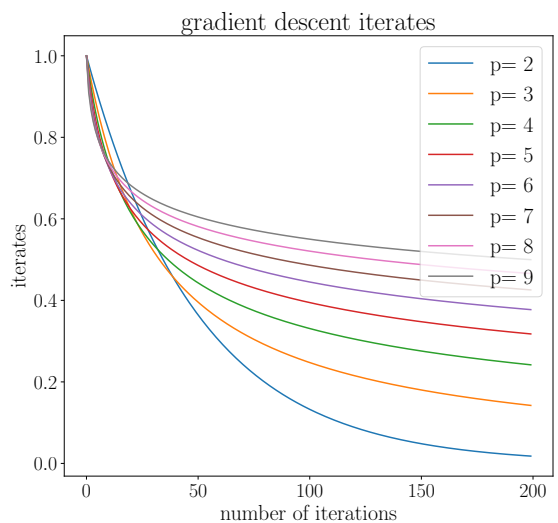
- satisfied by broad class of functions : semialgebraic functions; KŁ ineq. satisfied by NNs with ReLU activation functions.

Łojasiewicz exponent and speed of convergence

- Łojasiewicz exponent : $\theta \in (0, 1]$ when $\varphi(s) = \frac{c}{\theta} s^{\theta}$.
- Basic example : slower convergence for smaller exponent $1/p$.







Convergence rates under Łojasiewicz property

inspired by [Haraux and Jendoubi, 2015]

Theorem (Convergence rates)

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

- $\exists x^* \in \nabla F^{-1}(\{0\})$ s.t. $\lim_{t\to\infty} x(t) = x^*$.
- if $\theta \in (0, \frac{1}{2}]$ is a Łojasiewicz exponent of f at x^*, $\exists C > 0$ s.t.

$$\|x(t) - x^*\| \leq Ct^{-\frac{\theta}{1-2\theta}}, \quad \text{if } 0 < \theta < \frac{1}{2},$$

$$\|x(t) - x^*\| \leq Ce^{-\delta t}, \quad \text{for some } \delta > 0 \text{ if } \theta = \frac{1}{2}.$$

II. ADAM with constant stepsize

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).

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

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 $(\mathcal{H}_{model})+$ **moment assumption**,

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

Simulations

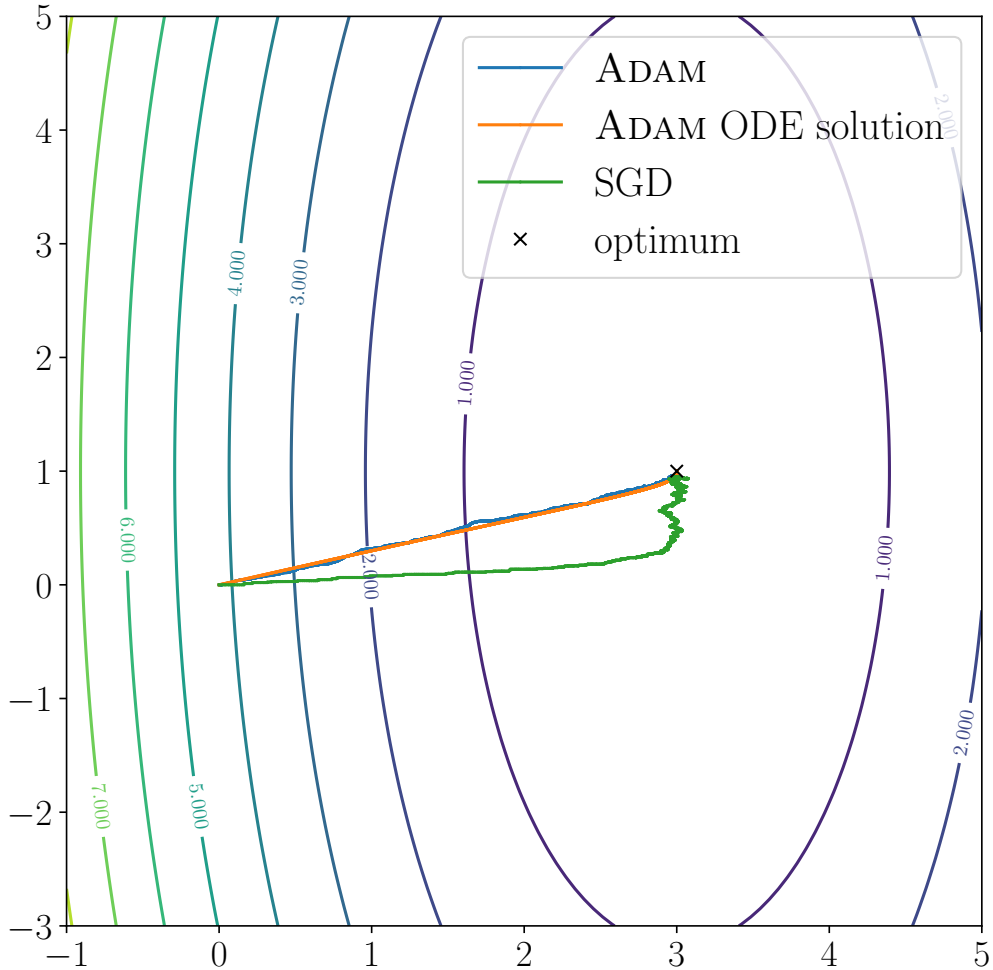


Figure 1: Convergence of ADAM and the ODE solution towards the optimum for a 2D linear regression

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 $(\mathcal{H}_{model})$ + moment assumption + :

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

Then, for all $\delta > 0$,

$$\lim_{\gamma \downarrow 0} \limsup_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{P}(\mathrm{d}(x_n^\gamma, \nabla F^{-1}(\{0\})) > \delta) = 0 \,. \tag{2}$$

III. ADAM with decreasing stepsizes

ADAM with decreasing stepsizes

Algorithm 3 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 $\frac{1-\beta_n}{\gamma_n} \to b$.

From Robbins Monro scheme to flows

RM scheme

$$\mathbf{y_{n+1}} = \mathbf{y_n} + \gamma_{\mathbf{n+1}}[\mathbf{g}(\mathbf{y_n}) + \mathbf{M_{n+1}} + \mathbf{b_{n+1}}]$$

- (γ_n) det. stepsizes s.t. $\gamma_n \geq 0, \gamma_n \to 0$ and $\sum_k \gamma_k = \infty$.
- (M_n) adapted to $\mathcal{F}_n$ and $\mathbb{E}[M_{n+1}|\mathcal{F}_n] = 0$.

- Noisy discretization of $\dot{\mathbf{y}}(\mathbf{t}) = \mathbf{g}(\mathbf{y}(\mathbf{t}))$.

  $\to$ Time interpolation : $\tau_n = \sum_{k=1}^{n} \gamma_k$.

$\to$ Link between asymptotic behavior of (y_n) and ODE ?

ODE method in Stochastic Approximation

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

- Informally:
  if $\gamma_n \to 0$, noise washes out and " $\lim_{n\to\infty} y_n = \lim_{t\to\infty} y(t)$" .

- (A bit) more precisely: if $Y(t) = y_n + \frac{t-\tau_n}{\tau_{n+1}-\tau_n}(y_{n+1} - y_n)$,

$$\forall T > 0, \quad \lim_{t\to\infty} \sup_{s\in[0,T]} \|Y(t+s) - \Phi_s(Y(t))\| = 0 .$$

  where $\Phi_s(y)$: solution to ODE at time s issued from y.

  $\to$ needs : control of the noise + boundedness of (y_n) .

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.

Theorem (Almost Sure Convergence)

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

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

Convergence w.p.1 to unique point if $\nabla F^{-1}(\{0\})$ finite/countable.

Stability

Additional assumptions ($\mathcal{H}_{stab}$)

- ∇F is Lipschitz continuous.
- $\exists C > 0 \ \forall x \in \mathbb{R}^d$, $\mathbb{E}[\|\nabla f(x, \xi)\|^2] \leq C(1 + F(x))$.

Theorem (stability)

Under $(\mathcal{H}'_{model}) + (\mathcal{H}_{as-cv}) + (\mathcal{H}_{stab})$, $(z_n = (x_n, m_n, v_n))$ satisfies

$$\sup_{n \in \mathbb{N}} \|z_n\| < \infty \quad a.s.$$

- Proof (technical) : adapt Lyapunov function to discrete time.

Towards a Central Limit Theorem

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

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

$$Z_{n+1} = Z_n + \underbrace{\left(\sqrt{\frac{\gamma_n}{\gamma_{n+1}}} - 1\right)}_{(2)\,(\frac{1}{2\gamma_0}\mathbb{1}_{\kappa=1}+o(1))\gamma_n} Z_n + \sqrt{\gamma_{n+1}}\,[\underbrace{h_\infty(z_n)}_{(3)} + b_{n+1}] + \underbrace{\sqrt{\gamma_{n+1}}\eta_{n+1}}_{(1)\ \textit{strong noise}} .$$

(3) On a neighborhood of $z^* \in h_\infty^{-1}(\{0\})$:

$$h_\infty(z_n) = \underbrace{h_\infty(z^*)}_{=0} + \nabla h_\infty(z^*)(z_n - z^*) + O(\|z_n - z^*\|^2)\,.$$

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

CLT using the SDE method

[Pelletier, 1998, Duflo, 1996]

Informal Th. ([Pelletier, 1998] adapted)- Strongly disturbed algo.

$$Z_{n+1} = (I + \gamma_{n+1}\bar{H})Z_n + \gamma_{n+1}\bar{b}_{n+1} + \sqrt{\gamma_{n+1}}\eta_{n+1} \in \mathbb{R}^k$$

s.t. $\mathbb{E}(\|Z_0\|^2) < 0$, $\bar{H}$ is a $k \times k$ **stable matrix**. Let $\Omega_0 \in \mathcal{F}_\infty$ have a positive probability. Assume a.s on Ω_0:

- Some assumptions to control (η_n) and $(\bar{b}_n)$.
- $\mathbb{E}_n(\eta_{n+1}\eta_{n+1}^T) = Q + \Delta_n$ where $\mathbb{E}(\|\Delta_n\|\mathbb{1}_{\Omega_0}) \to 0$, Q PSD matrix.

Then, given Ω_0, $Z_n \implies \mu$ unique stationary distribution of the process

$$dX_t = \bar{H}X_t dt + \sqrt{Q}dB_t$$

where (B_t) Brownian and $\mu \sim \mathcal{N}(0, \bar{\Sigma})$ with $\bar{H}\bar{\Sigma} + \bar{\Sigma}\bar{H}^T = -Q$.

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 $\kappa = 1$.

Asymptotic variance formula

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

- Trigonalization of the matrix

$$H:=\nabla h_\infty(z^*)=\begin{pmatrix}0 & -D & 0\\ a\nabla^2 F(x^*) & -aI_d & 0\\ b\nabla S(x^*) & 0 & -bI_d\end{pmatrix}\;;D:=\mathrm{diag}\left((\varepsilon+\sqrt{S(x^*)})^{-1}\right)$$

Asymptotic Variance Formula

$$\Sigma_1=D^{1/2}P\left(\frac{C_{k,\ell}}{(1-\frac{2\zeta}{a})(\lambda_k+\lambda_\ell-2\zeta+\frac{2}{a}\zeta^2)+\frac{1}{2(a-2\zeta)}(\lambda_k-\lambda_\ell)^2}\right)_{k,\ell=1\dots d}P^{-1}D^{1/2}$$

$D^{1/2}\nabla^2F(x^*)D^{1/2}=P\mathrm{diag}(\lambda_1,\cdots,\lambda_d)P^{-1}$; C omitted here.

Asymptotic variance reduction?

$$\Sigma_1 = D^{1/2} P \left(\frac{C_{k,\ell}}{(1-\frac{2\zeta}{a})(\lambda_k+\lambda_\ell-2\zeta+\frac{2}{a}\zeta^2)+\frac{1}{2(a-2\zeta)}(\lambda_k-\lambda_\ell)^2} \right)_{k,\ell=1\dots d} P^{-1} D^{1/2}$$

$$D := \operatorname{diag}\left((\varepsilon + \sqrt{S(x^*)})^{-1} \right)$$

- Influence of b, v_n? Σ_1 coincides with limiting covariance of:

$$\begin{cases} x_{n+1} &= x_n - \gamma_{n+1} p_{n+1} \\ p_{n+1} &= p_n + a\gamma_{n+1}(D\nabla f(x_n, \xi_{n+1}) - p_n)\,, \end{cases}$$

  i.e. preconditioned stochastic heavy ball.

- Influence of a?

$$\Sigma_1 = \underbrace{\Sigma_1^{(0)}}_{\text{no momentum}} + \frac{1}{a} \underbrace{\Delta}_{\text{neither pos.def nor neg.def.}} + O\left(\frac{1}{a^2}\right)$$

Contributions and future work

1. Introduction and analysis of a continuous-time ADAM.
2. Long run convergence of ADAM with constant stepsize.
3. Almost sure convergence of ADAM with decreasing stepsizes.

Convergence rates in discrete time? $\rightarrow$ follow-up work
Avoidance of traps?
What about the non-differentiable case?

References I

Attouch, H. and Bolte, J. (2009).

On the convergence of the proximal algorithm for nonsmooth functions involving analytic features.
Mathematical Programming, 116(1-2):5–16.

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.

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.

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*.

References II

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. (1996).

Algorithmes stochastiques.
Springer.

Duflo, M. (1997).

Random iterative models, volume 34 of *Applications of Mathematics (New York)*.
Springer-Verlag, Berlin.

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).

Haraux, A. and Jendoubi, M. (2015).

The convergence problem for dissipative autonomous systems.
SpringerBriefs in Mathematics. Springer International Publishing.

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.

References III

Ljung, L. (1977).

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

Łojasiewicz, S. (1963).

Une propriété topologique des sous-ensembles analytiques réels.
Les équations aux dérivées partielles, 117:87–89.

Pelletier, M. (1998).

Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing.
Annals of Applied Probability, pages 10–44.

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.