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

Anas Barakat   Pascal Bianchi

LTCI, Télécom Paris, Institut polytechnique de Paris

Journées SMAI MODE 7-9 Septembre 2020

Outline

ADAM algorithm

From Discrete to Continuous-Time ADAM

Continuous-Time ADAM

- Existence, uniqueness, convergence
- Convergence

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

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

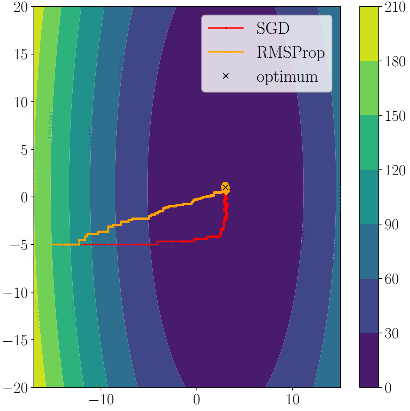
$$\gamma_{n,i} := \Psi(\nabla f(x_p, \xi_{p+1})_i, p \leq n)$$

RMSProp : coordinatewise stepsize

RMSProp

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

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

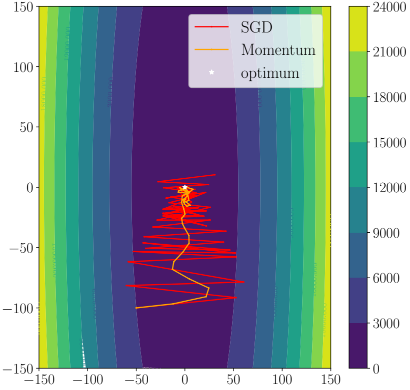


Momentum : (hoping) for acceleration

Momentum (aka Heavy Ball)

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

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



ADAM Algorithm

[Kingma and Ba, 2015]

- 51109 citations !

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

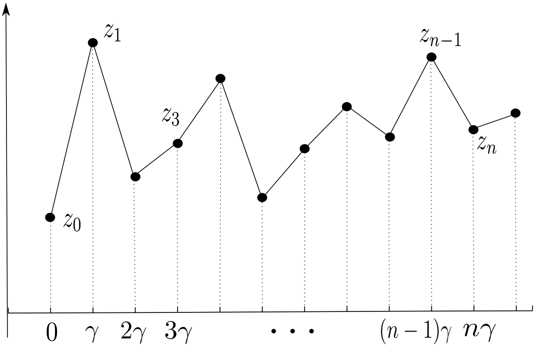
$x_0 \in \mathbb{R}^d$, $m_0 = 0$, $v_0 = 0$, $\gamma > 0$, $\varepsilon > 0$, $(\alpha, \beta) \in [0, 1)^2$.
for $n \geq 1$ **do**
  $m_n = \alpha m_{n-1} + (1 - \alpha)\nabla f(x_{n-1}, \xi_n)$
  $v_n = \beta v_{n-1} + (1 - \beta)\nabla f(x_{n-1}, \xi_n)^{\odot 2}$
  $\hat{m}_n = \frac{m_n}{1-\alpha^n}$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

I. From Discrete to Continuous-Time 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,

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

Decomposition in mean field + martingale noise

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

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

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).

II. Continuous-Time ADAM

Continuous Time System

Non autonomous ODE

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

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

Theorem

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

Remark : does not stem from off-the-shelf theorems.

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

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

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

- Łojasiewicz convergence rates inspired by [Haraux and Jendoubi, 2015].

III. ADAM with decreasing stepsizes

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

In this Talk, ADAM with decreasing stepsizes

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

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 :

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

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,

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

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

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

CLT using the SDE method

[Pelletier, 1998, Duflo, 1996]

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

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

Under some assumptions (on (η_n) and $(\bar{b}_n)$) a.s on $\Omega_0 \in \mathcal{F}_\infty$,

$$\text{given } \Omega_0, \quad Z_n \implies \mu \,,$$

unique stationary distribution of the process

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

$$\mu \sim \mathcal{N}(0, \bar{\Sigma}) \quad \text{with } \bar{H}\bar{\Sigma} + \bar{\Sigma}\bar{H}^T = -Q \,,$$

where (B_t) Brownian, Q omitted here.

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

- Trigonalization of

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

- Influence of b, v_n? Σ_1 = limiting covariance of:

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

i.e. preconditioned stochastic heavy ball.

Contributions and future work

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

2. 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?

More on : **anasbarakat.github.io**

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.

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. (1996).
Algorithmes stochastiques.
Springer.

References II

Duflo, M. (1997).
Random iterative models, volume 34 of *Applications of Mathematics (New York)*.
Springer-Verlag, Berlin.

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.

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.

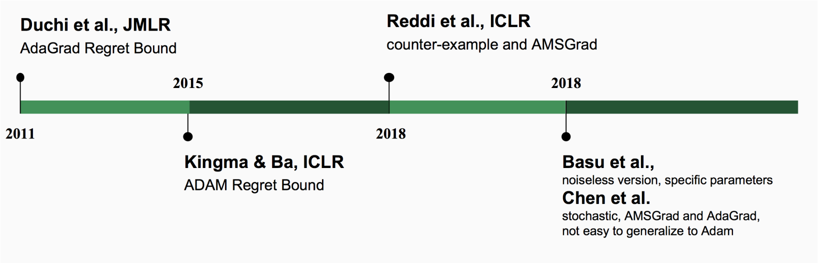
Pelletier, M. (1998).
Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing.
Annals of Applied Probability, pages 10–44.

References III

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



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

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

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

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.

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

- satisfied by broad class of functions : semialgebraic functions.

Convergence rates under Łojasiewicz property

inspired by [Haraux and Jendoubi, 2015]

Theorem (Convergence rates)

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

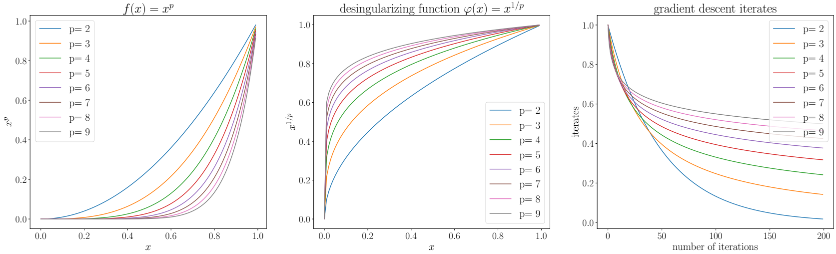
- $\exists x^* \in \nabla F^{-1}(\{0\})$ s.t. $\lim_{t\to\infty} x(t) = x^*$.
- if $\theta \in (0, \frac{1}{2}]$ is a Łojasiewicz exponent of f at x^*, $\exists C > 0$ s.t.

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

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

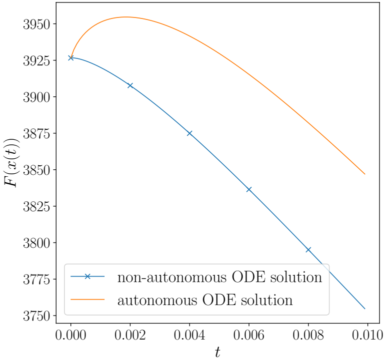
Łojasiewicz exponent and speed of convergence

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



Biased vs Unbiased ADAM

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



Algorithm 3 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: $x_n = x_{n-1} - \frac{\gamma}{\varepsilon + \sqrt{\hat{v}_n}} \hat{m}_n$
8: **end for**

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