Contributions to Non-Convex Stochastic Optimization and Reinforcement Learning

Anas Barakat, PhD Defense

LTCI, Télécom Paris, Institut Polytechnique de Paris

December 7th 2021

Jury:

Prof. Sébastien GADAT	President, Referee
Prof. Vivek S. BORKAR	Referee
Prof. Robert M. GOWER	Examiner
Prof. Niao HE	Examiner
Prof. Edouard PAUWELS	Examiner
Prof. Pascal BIANCHI	Supervisor
Prof. Walid HACHEM	Supervisor

Outline

- Introduction

PART I

1. Convergence analysis of Adam
2. Generalization to stochastic momentum algorithms
3. Some non-asymptotic results

PART II

4. Actor-critic with target network and linear FA for RL

- Conclusion and Perspectives

Guiding principle: ODE method

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

Algorithm

$$\begin{aligned} z_{n+1} &= z_n + \gamma_{n+1} H(n, z_n, \xi_{n+1}) \\ &= z_n + \gamma_{n+1} h(n, z_n) + \gamma_{n+1}\eta_{n+1}\,. \end{aligned}$$

where $h(n,z) := \mathbb{E}[H(n, z_n, \xi_{n+1})|\mathcal{F}_n], \mathcal{F}_n := \sigma(z_0, \xi_1, \cdots, \xi_n)$.

noisy discretization of

ODE

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

- Constant/decreasing stepsizes.
- Autonomous/non-autonomous.
- Stochastic optimization/RL.

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?

[Robbins and Monro, 1951]

Stochastic Gradient Descent (SGD)

$$\begin{aligned} x_{n+1} &= x_n - \gamma_n \nabla f(x_n, \xi_{n+1}) \\ &= x_n - \gamma_n \nabla F(x_n) + \gamma_n \eta_{n+1} \,. \end{aligned}$$

$$\dot{x}(t) = -\nabla F(x(t)) \qquad \text{(ODE)}$$

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

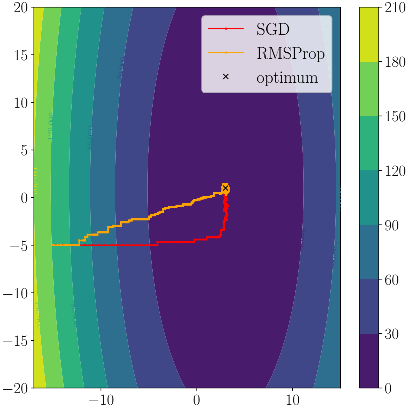
RMSProp : coordinatewise stepsize

[Tieleman and Hinton, 2012]

RMSProp

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

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

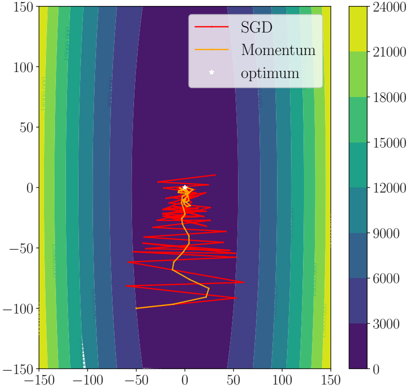


Momentum : (hoping) for acceleration

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]

- > 90000 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**

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

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

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

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

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

end for

- **Hyperparameters**: in practice α, β close to 1.

Related Work

- Existing theoretical guarantees
  - Regret bounds in the *convex* setting for variants of ADAM [Kingma and Ba, 2015, Reddi et al., 2018, Alacaoglu et al., 2020b].
  - Control of $\min_{0\leq k\leq N} \mathbb{E}[\|\nabla F(x_k)\|^2]$. [Zaheer et al., 2018, Basu et al., 2018, Chen et al., 2019, Zou et al., 2019, Alacaoglu et al., 2020a]

What about the convergence of the iterates?

1. Convergence analysis of ADAM

A. B. & Pascal Bianchi (2021). Convergence and Dynamical Behavior of the ADAM Algorithm for Non-Convex Stochastic Optimization.
In: *SIAM Journal on Optimization, 31 (1), 244-274.*

Continuous Time System

Non autonomous ODE

If $z(t) = (x(t), m(t), v(t))$, $z(0) = (x_0, 0, 0)$,

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

$$h(t, \underbrace{z}_{(x,m,v)}) = \begin{pmatrix} -\frac{(1-e^{-at})^{-1}m}{\varepsilon + \sqrt{(1-e^{-bt})^{-1}v}} \\ a(\nabla F(x) - m) \\ b(\mathbb{E}(\nabla f(x,\xi)^{\odot 2}) - v) \end{pmatrix}, a, b \text{ constants}$$

Theorem

Under regularity assumptions on f, coercivity of F and '$\alpha, \beta \sim 1$',

there exists a unique bounded global solution to ODE.

Convergence to stationary points

Theorem

Under same assumptions,

$$\lim_{t\to\infty} \mathrm{d}(x(t), \underbrace{\text{zeros}\,\nabla F}_{\text{critical points}}) = 0\,.$$

Key argument : Lyapunov function for the ODE

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

- + Convergence rates under Łojasiewicz property.

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

Under some standard assumptions, a moment assumption and:

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

Then, for all $\delta > 0$,

$$\lim_{\gamma \downarrow 0} \limsup_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{P}(\mathrm{d}(x_n^\gamma, \underbrace{\text{zeros}\, \nabla F}_{\text{critical points}}) > \delta) = 0\,.$$

Novel ADAM with decreasing stepsizes

Algorithm 2 ADAM $(\gamma_n, \alpha_n, \beta_n, \varepsilon)$.

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**
    $m_n = \alpha_n m_{n-1} + (1 - \alpha_n)\nabla f(x_{n-1}, \xi_n)$
    $v_n = \beta_n v_{n-1} + (1 - \beta_n)\nabla f(x_{n-1}, \xi_n)^{\odot 2}$
    $r_n = \alpha_n r_{n-1} + (1 - \alpha_n)$
    $\bar{r}_n = \beta_n \bar{r}_{n-1} + (1 - \beta_n)$
    $\hat{m}_n = m_n / r_n$ {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

Almost sure convergence

- ODE method: $h_\infty(z) = \lim_{t\to\infty} h(t,z)$

$$z_n = (x_n, m_n, v_n)$$

$$z_{n+1} = z_n + \gamma_{n+1} \underbrace{h_\infty}_{\text{mean field}} (z_n) + \gamma_{n+1} \underbrace{\eta_{n+1}}_{\text{noise}} + \gamma_{n+1} \underbrace{b_{n+1}}_{\text{bias}\to 0\ a.s.} ,$$

Theorem

Under some regularity and moment assumptions and if $\sum_n \gamma_n = +\infty$ and $\sum_n \gamma_n^2 < +\infty$, then, w.p.1,

$$\lim_{n\to\infty} \mathrm{d}(x_n, \underbrace{\text{zeros}\,\nabla F}_{\text{critical points}}) = 0\,.$$

Fluctuations

Theorem (conditional CLT)

Under some assumptions, given the event $\{z_n \to z^*\}$,

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

with Σ solution to Lyapunov equation (closed formula).

2. Generalization to stochastic momentum algorithms

A. B., Pascal Bianchi, Walid Hachem & Sholom Schechtman (2021). Stochastic optimization with momentum: convergence, fluctuations, and traps avoidance. In: *Electronic Journal of Statistics 15 (2), 3892-3947.*

A General Dynamical System

including ADAM and many others

- **Non-autonomous ODE** [Belotto da Silva and Gazeau, 2020]

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

$$\dot{z}(t) = h(t, z(t)) \iff \begin{cases} \dot{\mathrm{v}}(t) & = \mathrm{p}(t)S(\mathrm{x}(t)) - \mathrm{q}(t)\mathrm{v}(t) \\ \dot{\mathrm{m}}(t) & = \mathrm{h}(t)\nabla F(\mathrm{x}(t)) - \mathrm{r}(t)\mathrm{m}(t) \\ \dot{\mathrm{x}}(t) & = -\mathrm{m}(t)/\sqrt{\mathrm{v}(t) + \varepsilon} \end{cases}$$

- $h_\infty(z) = \lim_{t\to\infty} h(t, z)$.

Theorem

$$\lim_{t\to\infty} \mathrm{d}(z(t), \mathrm{zeros}\, h_\infty) = 0\,,$$

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

A particular case: Nesterov

Theorem (Nesterov ODE)

Let F be possibly **nonconvex**, then

$$\lim_{t\to\infty} \mathrm{d}(x(t), \mathrm{zeros}\,\nabla F) = 0\,,$$

where the prior ODE amounts to $\ddot{x}(t) + \frac{3}{t}\dot{x}(t) + \nabla F(x(t)) = 0\,.$

- [Su et al., 2016] (convex setting) and [Cabot et al., 2009]

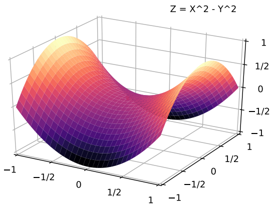
General Algorithm

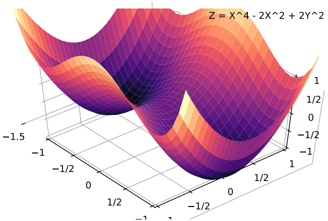
Stochastic algorithm

$$\begin{cases} v_{n+1} & = (1-\gamma_{n+1} q_n) v_n + \gamma_{n+1} p_n \nabla f(x_n, \xi_{n+1})^{\odot 2} \\ m_{n+1} & = (1-\gamma_{n+1} r_n) m_n + \gamma_{n+1} h_n \nabla f(x_n, \xi_{n+1}) \\ x_{n+1} & = x_n - \gamma_{n+1} m_{n+1} / \sqrt{v_{n+1} + \varepsilon} \end{cases}$$

- Generalization of ADAM results: a.s. convergence, CLT.
- [Gadat and Gavra, 2020] ADAGRAD and RMSPROP with the possibility to use mini-batches but without momentum.

Avoidance of trap problem





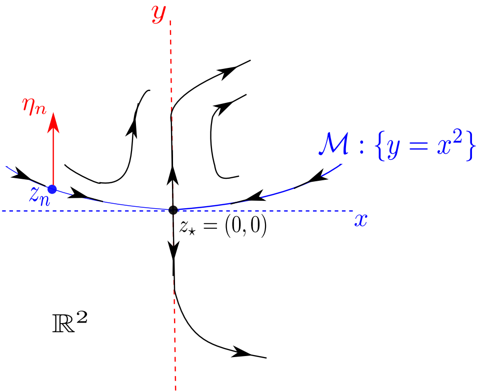
- Points where $\nabla^2 F(x)$ is not positive semidefinite: e.g., saddle points, local maxima.

Do the algorithms converge toward these undesirable points?

The invariant manifold approach

[Pemantle, 1990, Brandière and Duflo, 1996, Benaïm, 1999]

$$\dot{z}(t) = h(z(t)) \quad \text{with} \quad h(z) = h((x,y)) = (-x+(x^2-y)^4, y-3x^2)\,.$$



$$z_{n+1} = z_n + \gamma_n h(z_n) + \gamma_n \eta_{n+1}\,.$$

Our general avoidance of traps result

Non-autonomous invariant manifold [Pötzsche and Rasmussen, 2006]

There exist an invariant manifold for $\dot{z}(t) = h(t, z(t))$:

$$\mathcal{M} = \left\{ \left(t, \begin{bmatrix} z^- \\ w(z^-, t) \end{bmatrix} \right) \in I \times \mathbb{R}^d \ : \ z^- \in \mathbb{R}^{d^-} \right\}$$

where $d^+ = \dim(Eigen(\nabla h(z_\star) : Re(\lambda) > 0))$.

Theorem

$$z_{n+1} = z_n + \gamma_{n+1} h(n, z_n) + \gamma_{n+1}\eta_{n+1} + \gamma_{n+1} b_{n+1}$$

Assume $h(t, z) = \nabla h_\infty(z_\star)(z - z_\star) + e(t, z)$ close to $z_\star \in \text{zeros}\, h_\infty$ and

$$\liminf \mathbb{E}[\|P_+(\eta_{n+1})\|^2 \,|\, \mathcal{F}_n] \geq c^2 > 0\,,$$

where $P_+(\eta_n)$ projection on $Eigen(\nabla h_\infty(z_\star))$ s.t.$Re(\lambda) > 0$. Under assumptions on e , b_n , η_n , γ_n, $\mathbb{P}([z_n \to z_\star]) = 0$.

Application to stochastic algorithms

Proposition

Let $z_\star = (x_\star, m_\star, v_\star) \in \operatorname{zeros} h_\infty$ and write:

$$h(t,z) = \nabla h_\infty(z_\star)(z - z_\star) + e(z,t)\,.$$

$$\dim(Eigen(\nabla h_\infty(z_\star) : Re(\lambda) > 0)) = \dim(Eigen(\nabla^2 F(x_\star) : Re(\lambda) < 0))\,.$$

Eg. Trap avoidance for S-NAG

Let $x_\star \in \operatorname{zeros} \nabla F$ s.t. $\nabla^2 F(x_\star)$ has a negative eigenvalue. If

$$\Pi_{\mathrm{u}} \mathbb{E}_\xi (\nabla f(x_\star, \xi) - \nabla F(x_\star))(\nabla f(x_\star, \xi) - \nabla F(x_\star))^T \Pi_{\mathrm{u}} \neq 0\,,$$

where Π_{u} orthogonal projector on $Eigen(\nabla^2 F(x_\star))$ s.t. $Re(\lambda) < 0$.

Then, $\mathbb{P}([x_n \to x_\star]) = 0$.

3. Some non-asymptotic results

A. B. & Pascal Bianchi (2020). Convergence Rates of a Momentum Algorithm with Bounded Adaptive Stepsize for Non-Convex Optimization. In: *Asian Conference on Machine Learning 2020, PMLR, 129, 225-240.*

A Momentum Algorithm with Adaptive Stepsize

Algorithm

$$\begin{cases} m_{n+1} = m_n + b\left(\nabla f(x_n, \xi_{n+1}) - m_n\right) \\ x_{n+1} = x_n - \underbrace{a_{n+1}}_{\in \mathbb{R}_+^d} m_{n+1} \end{cases}$$

- recovers SGD, Heavy Ball, AdaGrad, ADAM ....

Theorem (stochastic)

Under standard regularity assumptions, if :

$$0 < \delta \leq a_{n+1} \leq a_{\sup}(L) \simeq \frac{2}{L},$$

then,

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}[\|\nabla F(x_k)\|^2] = O\left(\frac{1}{n}\right) + \underbrace{O(\sigma^2)}_{\forall x, \mathbb{V}(\nabla f(x,\xi)) \leq \sigma^2}.$$

- Limitations: clipping and same as SGD.

Convergence rates under the KŁ property

- local prop. satisfied by semialgebraic funs, even NNs (ReLU).

Theorem (deterministic)

Under the same assumptions on F and a_n, if:

- F is a KŁ function with KŁ exponent θ,

then, $\lim_k F(x_k) = F(x_*)$ for some critical point x^* and

$$F(x_k) - F(x_*) = \begin{cases} O(q^k) & \text{for } q \in (0,1) \text{ if } 1/2 \leq \theta < 1 \\ O(k^{\frac{1}{2\theta-1}}) & \text{if } 0 < \theta < 1/2 \end{cases}$$

- [Bolte et al., 2018] for gradient-like descent sequences.

4. Actor-critic with target network and linear FA for RL

A. B, Pascal Bianchi & Julien Lehmann (2021). Analysis of a Target-Based Actor-Critic Algorithm with Linear Function Approximation.
ArXiv Preprint: arXiv:2106.07472.

(Preliminary) Motivation

- Stochastic approximation and ODE method.
- Actor-critic: popular methods in deep RL.

Outline

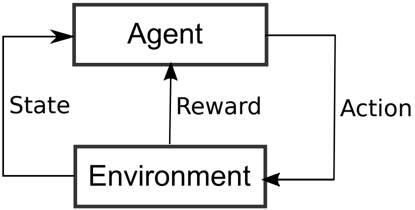
a. **Standard Actor-Critic**

b. **Actor-Critic with target network**

c. **Critic analysis**

d. **Actor analysis**

Reinforcement Learning

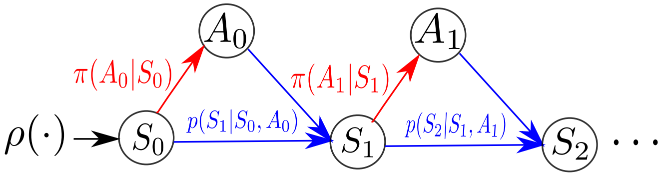


Goal

Maximize long-term rewards

Markov Decision Process and RL problem

- Environment → MDP $(\mathcal{S}, \mathcal{A}, p, R, \rho, \gamma)$.
- Agent → Policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$.



Problem

$$\max_\pi J(\pi) := \mathbb{E}_{\rho,\pi}\left[\sum_{t=0}^{+\infty} \gamma^t R_{t+1}\right]$$

Policy Gradient framework

- Policy parameterization: $\max_{\theta \in \mathbb{R}^d} J(\theta) := J(\pi_\theta)$.
- (Stochastic) Gradient Ascent:

$$\theta_{t+1} = \theta_t + \alpha_t \widehat{\nabla J(\theta_t)}\,.$$

Policy Gradient Theorem [Sutton et al., 1999, Konda, 2002]

Under some regularity conditions on $\theta \mapsto \pi_\theta$,

$$\nabla J(\theta) = \mathbb{E}_{(S,A)\sim\mu_{\rho,\theta}}\Big[\underbrace{\Delta_{\pi_\theta}(S,A)}_{\text{advantage function}} \nabla \ln \pi_\theta(A|S)\Big]\,,$$

where $\mu_{\rho,\theta}$ is the discounted state-action visitation distribution.

Policy evaluation

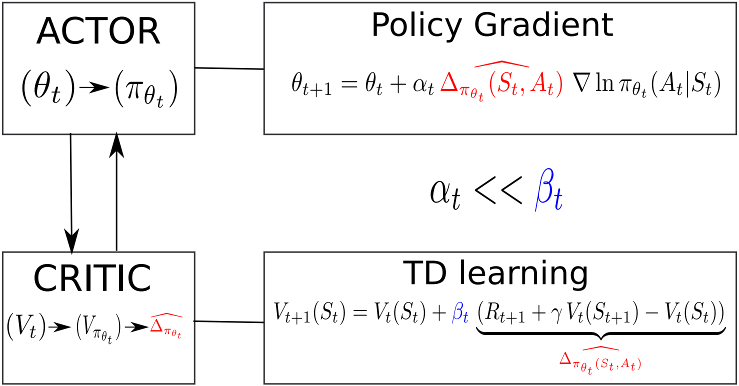
- Given π, to estimate Δ_π, estimate V_π where:

$$V_\pi(s) := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | S_0 = s \right] .$$

- Temporal Difference (TD) learning algorithm:

$$V_{t+1}(S_t) = V_t(S_t) + \beta_t \underbrace{\left(R_{t+1} + \gamma\, V_t(S_{t+1}) - V_t(S_t)\right)}_{\widehat{\Delta_\pi(S_t, A_t)}} .$$

(Standard) Actor-Critic



Critic with function approximation

Huge state space → use FA: $V_{\pi_\theta}(s) \approx V_\omega(s)$.

$$\omega_{t+1} = \omega_t + \beta_t(R_{t+1} + \gamma V_{\omega_t}(S_{t+1}) - V_{\omega_t}(S_t))\nabla_\omega V_{\omega_t}(S_t)\,.$$

- **Linear FA:** $V_\omega(s) = \omega^T \phi(s)$ where $\omega \in \mathbb{R}^m$ for $m \ll |\mathcal{S}|$.
- **Nonlinear FA:** $V_\omega(s) = NN_\omega(s)$.
  → INSTABILITY [Tsitsiklis and Van Roy, 1997]

Experimental trick: using a target network

Standard critic with FA:

TD learning with FA

$$\omega_{t+1} = \omega_t + \beta_t(R_{t+1} + \gamma V_{\omega_t}(S_{t+1}) - V_{\omega_t}(S_t))\nabla_\omega V_{\omega_t}(S_t)$$

Now:

(Target) - TD learning

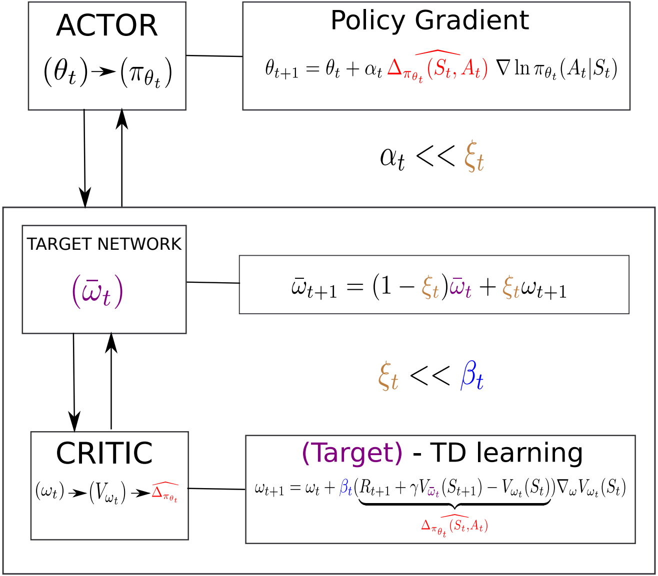
$$\omega_{t+1} = \omega_t + \beta_t(R_{t+1} + \gamma V_{\bar{\omega}_t}(S_{t+1}) - V_{\omega_t}(S_t))\nabla_\omega V_{\omega_t}(S_t)$$

$\xi_t << \beta_t$

Target Network

$$\bar{\omega}_{t+1} = (1 - \xi_t)\bar{\omega}_t + \xi_t \omega_{t+1}$$

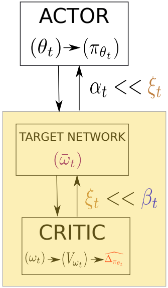
Target-based actor-critic



Motivation: few remarks

- Trick was proposed in [Mnih et al., 2013] for DQN and analyzed in [Avrachenkov et al., 2021].
- Several deep RL **actor-critic** use this trick.
  Is this theoretically sound?
- Here, we look at linear FA to pave the way for nonlinear FA.
  - even linear setting not understood for AC,
    [Lee and He, 2019] single timescale target-TD,
    [Zhang et al., 2021] value-based methods .

Critic analysis



Convergence analysis (Critic)

- Multi-timescales SA
  [Borkar, 1997, Borkar, 2008, Karmakar and Bhatnagar, 2018]

Theorem

Under standard assumptions (Markov chain ergodicity, stepsizes, independence of the features), if $\frac{\alpha_t}{\xi_t} \to 0$ and $\frac{\xi_t}{\beta_t} \to 0$,

$$\lim_t \|\omega_t - \omega_*(\theta_t)\| = 0 \ w.p.1 .$$

where $\omega_*(\theta)$ solution to some linear system $\forall \theta$.

- same interpretation than TD-like solution with linear FA
  [Tsitsiklis and Van Roy, 1997]

Finite-time analysis (Critic)

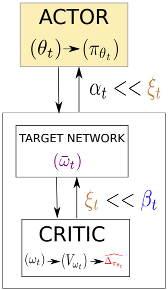
Theorem

Let $0 < \beta < \xi < \alpha < 1$. Set $\alpha_t = \frac{c_1}{t^\alpha}$, $\xi_t = \frac{c_2}{t^\xi}$, $\beta_t = \frac{c_3}{t^\beta}$. Then,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\omega_t - \omega_*(\theta_t)\|^2] = \mathcal{O}\left(\frac{1}{T^{1-\xi}}\right) + \mathcal{O}\left(\frac{\log T}{T^\beta}\right) + \mathcal{O}\left(\frac{1}{T^{2(\alpha-\xi)}}\right) + \mathcal{O}\left(\frac{1}{T^{2(\xi-\beta)}}\right).$$

- $\alpha > \xi$ and $\xi > \beta$.

Actor analysis



Convergence analysis (Actor)

Theorem

Under same assumptions, if $\frac{\alpha_t}{\xi_t} \to 0$ and $\frac{\xi_t}{\beta_t} \to 0$,

$$\liminf_t \left(\|\nabla J(\theta_t)\| - \underbrace{\|b(\theta_t)\|}_{\text{bias due to linear FA}} \right) \leq 0\,, w.p.1$$

Finite-time analysis (Actor)

Preliminary result

Set $\alpha_t = \frac{c_1}{t^\alpha}$, $\xi_t = \frac{c_2}{t^\xi}$, $\beta_t = \frac{c_3}{t^\beta}$ with $0 < \beta < \xi < \alpha < 1$. Then,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2] = \mathcal{O}\left(\frac{1}{T^{1-\alpha}}\right) + \mathcal{O}\left(\frac{\log^2 T}{T^\alpha}\right) + \mathcal{O}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\omega_t - \omega_*(\theta_t)\|^2]\right) + \mathcal{O}(\epsilon_{\mathsf{FA}}) .$$

Theorem (Actor with tuned stepsizes)

Set $\alpha_t = \frac{c_1}{t^{2/3}}$, $\xi_t = \frac{c_2}{t^{1/2}}$, $\beta_t = \frac{c_3}{t^{1/3}}$. Then,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla J(\theta_t)\|^2] = \mathcal{O}\left(\frac{\log T}{T^{1/3}}\right) + \mathcal{O}(\epsilon_{\mathsf{FA}}) .$$

Contributions and perspectives

About AC methods with target networks for RL

- **Contributions**
  - Convergence analysis: critic and actor.
  - Finite-time analysis: average expected gradient norm.
- **Perspectives**
  - Nonlinear FA for deep RL.
  - Off-policy learning.

Contributions of this thesis
Non-convex stochastic optimization

- ADAM.
  - ODE analysis,
  - constant stepsize,
  - decreasing stepsizes.
- Generalization beyond ADAM.
  - Avoidance of traps: general non-autonomous result.
- Non-asymptotic results.

Perspectives

Non-convex stochastic optimization

- Non-asymptotic results: extension of the KL analysis to the stochastic setting?
- Constrained optimization: proximal variants.
- Nonsmoothness/non-differentiability.
  [Davis et al., 2020, Bolte and Pauwels, 2019]

Possibility of bridging both parts: momentum and RL.

Acknowledgements

I would like to thank:

- My supervisors Prof. Bianchi and Prof. Hachem and other collaborators Sholom Schechtman and Julien Lehmann.
- All the members of the jury.

References I

Alacaoglu, A., Malitsky, Y., and Cevher, V. (2020a).
Convergence of adaptive algorithms for weakly convex constrained optimization.
arXiv preprint arXiv:2006.06650.

Alacaoglu, A., Malitsky, Y., Mertikopoulos, P., and Cevher, V. (2020b).
A new regret analysis for adam-type algorithms.
arXiv preprint arXiv:2003.09729.

Avrachenkov, K. E., Borkar, V. S., Dolhare, H. P., and Patil, K. (2021).
Full gradient dqn reinforcement learning: a provably convergent scheme.
In *Modern Trends in Controlled Stochastic Processes:*, pages 192–220. Springer.

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.

Belotto da Silva, A. and Gazeau, M. (2020).
A general system of differential equations to model first-order adaptive algorithms.
Journal of Machine Learning Research, 21(129):1–42.

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.

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

Bolte, J. and Pauwels, E. (2019).
Conservative set valued fields, automatic differentiation, stochastic gradient method and deep learning.
arXiv preprint arXiv:1909.10300.

Bolte, J., Sabach, S., Teboulle, M., and Vaisbourd, Y. (2018).
First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems.
SIAM Journal on Optimization, 28(3):2131–2151.

Borkar, V. S. (1997).
Stochastic approximation with two time scales.
Systems & Control Letters, 29(5):291–294.

Borkar, V. S. (2008).
Stochastic approximation.
Cambridge University Press, Cambridge; Hindustan Book Agency, New Delhi.
A dynamical systems viewpoint.

Brandière, O. and Duflo, M. (1996).
Les algorithmes stochastiques contournent-ils les pièges?
Ann. Inst. H. Poincaré Probab. Statist., 32(3):395–427.

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.

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 III

Davis, D., Drusvyatskiy, D., Kakade, S., and Lee, J. (2020).
Stochastic subgradient method converges on tame functions.
Foundations of Computational Mathematics, 20(1):119–154.

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).

Gadat, S. and Gavra, I. (2020).
Asymptotic study of stochastic adaptive algorithm in non-convex landscape.
arXiv preprint arXiv:2012.05640.

Karmakar, P. and Bhatnagar, S. (2018).
Two time-scale stochastic approximation with controlled Markov noise and off-policy temporal-difference learning.
Math. Oper. Res., 43(1):130–151.

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

Konda, V. R. (2002).
Actor-Critic Algorithms.
PhD thesis, USA.
AAI0804543.

References IV

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.

Lee, D. and He, N. (2019).
Target-based temporal-difference learning.
In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3713–3722. PMLR.

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

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013).
Playing atari with deep reinforcement learning.
arXiv preprint arXiv:1312.5602.

Pemantle, R. (1990).
Nonconvergence to unstable points in urn models and stochastic approximations.
Ann. Probab., 18(2):698–712.

Pötzsche, C. and Rasmussen, M. (2006).
Taylor approximation of integral manifolds.
J. Dynam. Differential Equations, 18(2):427–460.

Reddi, S. J., Kale, S., and Kumar, S. (2018).
On the convergence of adam and beyond.
In *International Conference on Learning Representations*.

References V

Robbins, H. and Monro, S. (1951).
A stochastic approximation method.
The annals of mathematical statistics, pages 400–407.

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.

Sutton, R. S., Mcallester, D., Singh, S., and Mansour, Y. (1999).
Policy gradient methods for reinforcement learning with function approximation.
In *Advances in Neural Information Processing Systems 12*, volume 99, pages 1057–1063. MIT Press.

Tieleman, T. and Hinton, G. (2012).
Lecture 6.e-rmsprop: Divide the gradient by a running average of its recent magnitude.
Coursera: Neural networks for machine learning, pages 26–31.

Tsitsiklis, J. N. and Van Roy, B. (1997).
An analysis of temporal-difference learning with function approximation.
IEEE transactions on automatic control, 42(5):674–690.

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.

Zhang, S., Yao, H., and Whiteson, S. (2021).
Breaking the deadly triad with a target network.
ICML 2021, arXiv preprint arXiv:2101.08862.

Zou, F., Shen, L., Jie, Z., Zhang, W., and Liu, W. (2019).
A sufficient condition for convergences of adam and rmsprop.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11127–11135.

References VI