Stochastic optimization with momentum: convergence, fluctuations, and traps avoidance

Anas Barakat

Télécom Paris, Institut Polytechnique de Paris

December 20th 2021, CMStatistics 2021

Outline

- **Introduction**

1. **Convergence analysis of Adam (as a motivation)**
2. **Generalization to stochastic momentum algorithms**

- **Conclusion and Perspectives**

Based on

- 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. B. & Pascal Bianchi (2021). Convergence and dynamical behavior of the ADAM algorithm for non-convex stochastic optimization. *SIAM Journal on Optimization, 31 (1), 244-274.*

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

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

- Autonomous/non-autonomous.

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

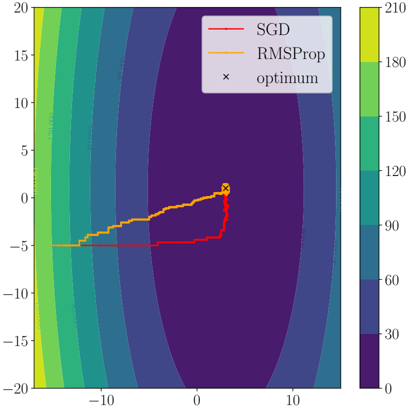
RMSProp : coordinatewise stepsize

[Tieleman and Hinton, 2012]

RMSProp

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

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

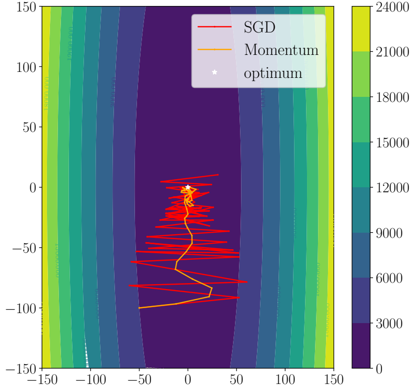


Momentum : (hoping) for acceleration

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 $(\gamma, \alpha, \beta, \varepsilon)$

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

- **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?

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

1. Convergence analysis of ADAM

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,

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

Key argument : Lyapunov function for the ODE

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

- + Convergence rates under Łojasiewicz property.

Almost sure convergence

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

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

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

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 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]

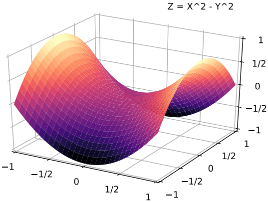
General Algorithm

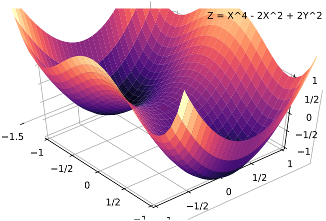
Stochastic algorithm

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

- 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





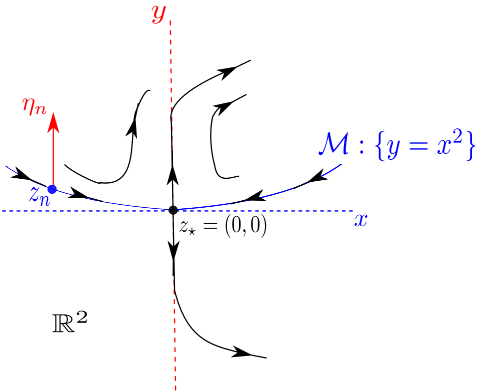
- Points where $\nabla^2 F(x)$ is not positive semidefinite: e.g., saddle points, local maxima.

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

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 $\mathit{Eigen}(\nabla^2 F(x_\star))$ s.t. $\mathit{Re}(\lambda) < 0$.

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

Summary and perspectives

- ADAM: ODE analysis, a.s. convergence and CLT.
- Generalization beyond ADAM.
  - Avoidance of traps: general non-autonomous result.

Perspectives

- Constrained optimization: proximal variants.
- Nonsmoothness/non-differentiability.
- Other problems (min-max optimization, sampling, OT)

Thank you for your attention

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.

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.

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

References II

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*.

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.

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

References III

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.

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*.

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.

References IV

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.

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.

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.

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 $\delta > 0$,

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

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