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

Anas Barakat, Pascal Bianchi

LTCI, Télécom Paris, Institut polytechnique de Paris

Machine Learning in the Real World, October 2nd 2019

Optimization in Deep Learning

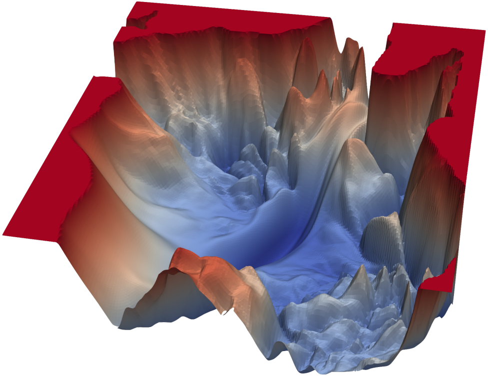


Figure 1: Visualization of a loss landscape (VGG-56 on CIFAR-10)
https://www.cs.umd.edu/ tomg/projects/landscapes/

Li et al., Visualizing the Loss Landscape of Neural Nets, NeurIPS 2018

Problem statement

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
- regularity assumptions on f (smoothness, coercivity of F, etc.)
- $(\xi_n : n \geq 1)$: iid copies of r.v ξ revealed online

ADAM : an adaptive algorithm
[Kingma and Ba, 2015]

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

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**

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

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

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

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

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

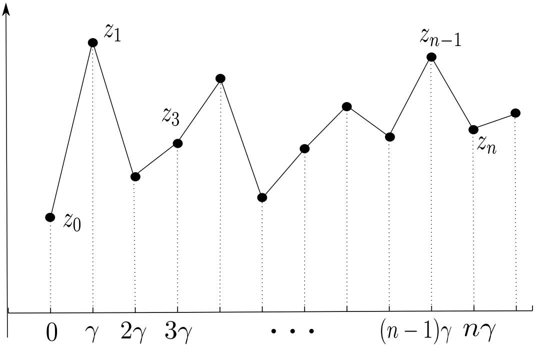
$\quad x_n = x_{n-1} - \gamma \nabla f(x_{n-1}, \xi_n)$ (SGD for comparison)

end for

From Discrete to Continuous Time

The ODE Method [Ljung, 1977, Kushner and Yin, 2003]

$\mathsf{z}^\gamma(t)$ interpolated from $z_n^\gamma = (x_n^\gamma, m_n^\gamma, v_n^\gamma)$



Continuous Time System

similar approach to [Su, Boyd and Candès, 2016]

Non autonomous ODE

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

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

Theorem (Convergence)

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

$$c_1(t)\,\ddot{x}(t) + c_2(t)\,\dot{x}(t) + \nabla F(x(t)) = 0\,,$$

- 2nd vs 1st order: acceleration (even if oscillations).
- Escaping local traps (saddle points)

Long run convergence of the ADAM iterates

- No a.s convergence : regime $n \to \infty$ then $\gamma \to 0$

Theorem (ergodic convergence of the ADAM iterates)

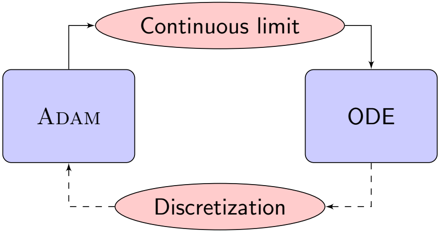
Let $x_0 \in \mathbb{R}^d$, $\gamma > 0$, $(z_n^\gamma : n \in \mathbb{N})$, $z_0^\gamma = (x_0, 0, 0)$. Under the same assumptions and :

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

Then, for all $\delta > 0$,

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

Thank you for your attention



For more details: submitted article, available on arXiv.

AB, P. Bianchi. *Convergence and Dynamical Behavior of the ADAM Algorithm for Non Convex Stochastic Optimization.*