Convergence de l'algorithme ADAM du point de vue des systèmes dynamiques

Anas Barakat, Pascal Bianchi

LTCI, Télécom Paris, Institut polytechnique de Paris

GRETSI 29 Août 2019

Optimisation des poids des réseaux de neurones

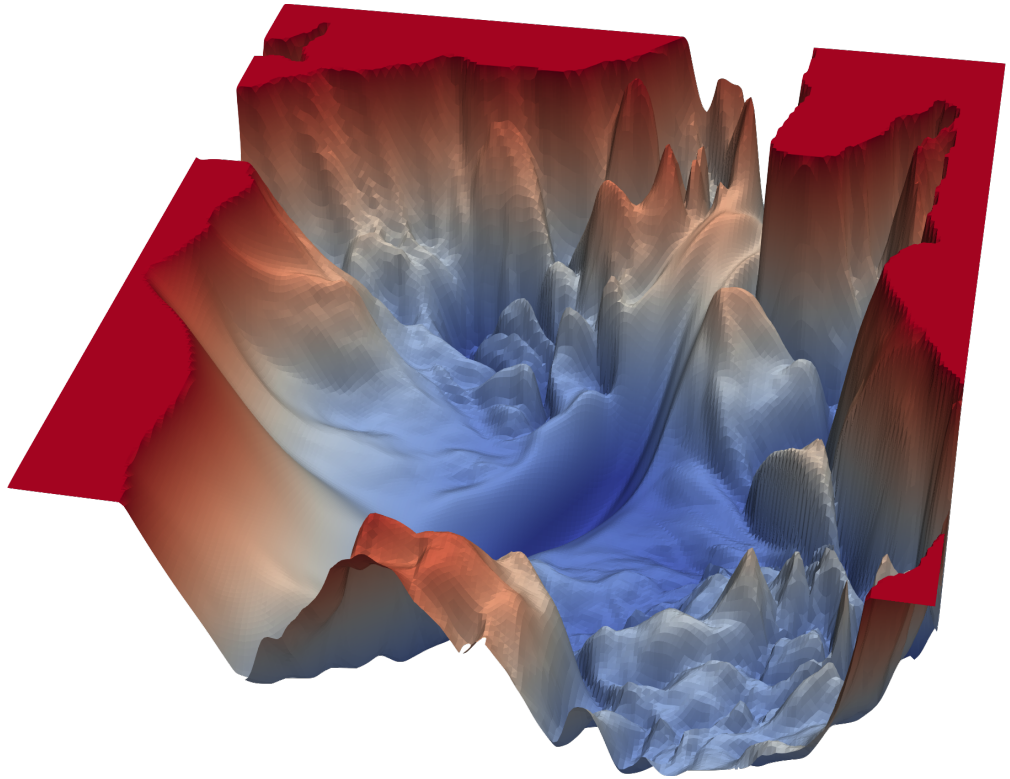


Figure 1: Visualisation d'une fonction de perte (VGG-56 sur CIFAR-10)
Crédit image: https://www.cs.umd.edu/ tomg/projects/landscapes/

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

Position du problème

Problème

$$\min_x F(x) := \mathbb{E}(f(x,\xi)) \quad \text{w.r.t.} \quad x \in \mathbb{R}^d$$

Hypothèses

- $f(\,.\,,\xi)$: fonction **non convexe**, différentiable
- $(\xi_n : n \geq 1)$: copies iid d'une v.a. ξ révélée en ligne

Solution ?

Descente de Gradient Stochastique (SGD)

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

- Limitations
  - choix du pas
  - pas commun à toutes les coordonnées du gradient

Algorithmes Adaptatifs

SGD standard

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

$$\gamma_n := \gamma \quad \text{ou} \quad \gamma_n := \frac{1}{\sqrt{n}}, n \geq 1$$

Algorithmes Adaptatifs

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

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

Algorithme ADAM

[Kingma and Ba, 2015]

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**

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

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

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

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

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

end for

Hypothèses et régime asymptotique

- Régime : **pas constant** $\gamma > 0$.

Hypothèses sur f

- hypothèses de régularité sur f .
- $F : x \mapsto \mathbb{E}(f(x, \xi))$ coercive.
- $S : x \mapsto \mathbb{E}(\nabla f(x, \xi)^2)$ t.q. $\forall x \in \mathbb{R}^d$, $S(x) > 0$.

Conséquence : $\nabla F^{-1}(\{0\}) \neq \emptyset$.

- Hypothèses sur les hyperparamètres: compatibles avec la pratique.

Du Temps Discret au Temps Continu

Temps Continu : Analyse du Système Dynamique

Temps Discret : Convergence d'ADAM

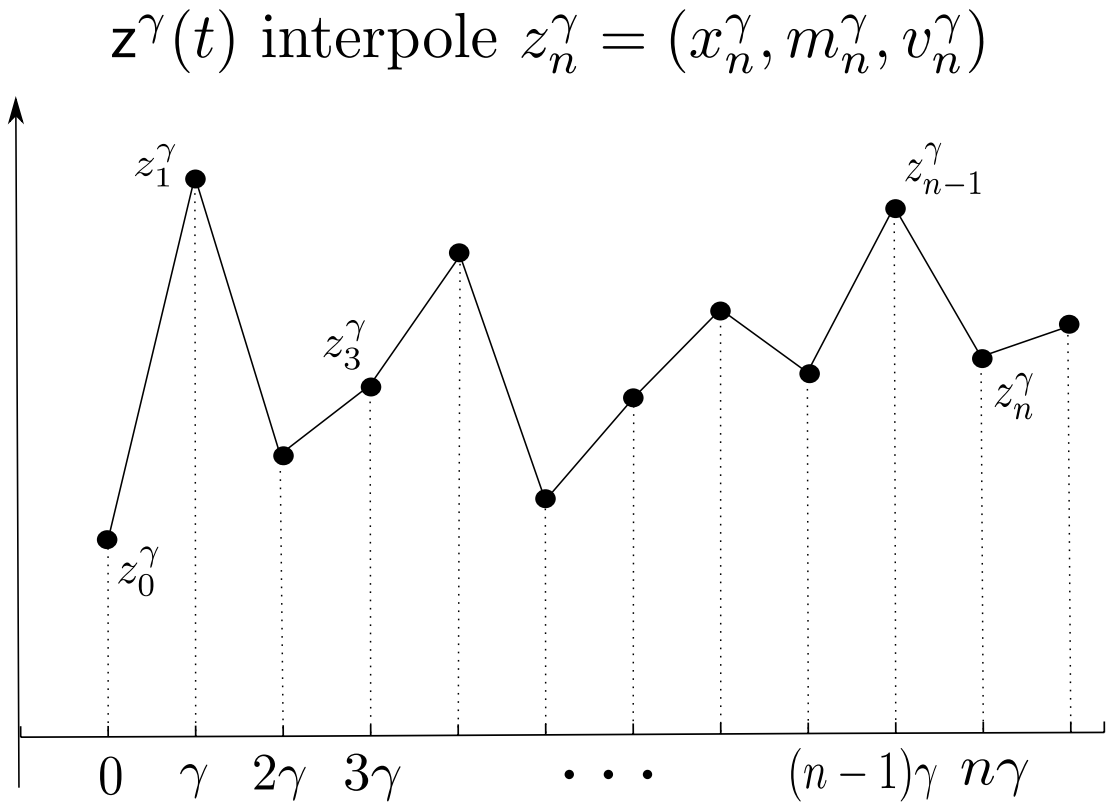
Du Temps Discret au Temps Continu

Temps Continu : Analyse du Système Dynamique

Temps Discret : Convergence d'ADAM

La méthode de l'EDO

[Ljung, 1977, Kushner and Yin, 2003]



Passage au Temps Continu

$$z_n^\gamma := z_{n-1}^\gamma + \gamma H_\gamma(n, z_{n-1}^\gamma, \xi_n)\,,$$

Pour tout $\gamma > 0$, pour tout z,

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

Décomposition champ moyen + bruit martingale

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

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

Du Temps Discret au Temps Continu

Temps Continu : Analyse du Système Dynamique

Temps Discret : Convergence d'ADAM

Système à Temps Continu

approche similaire à [Su et al., 2016]

EDO non autonome

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

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

où pour tout $t > 0$, tout $z = (x, m, v)$:

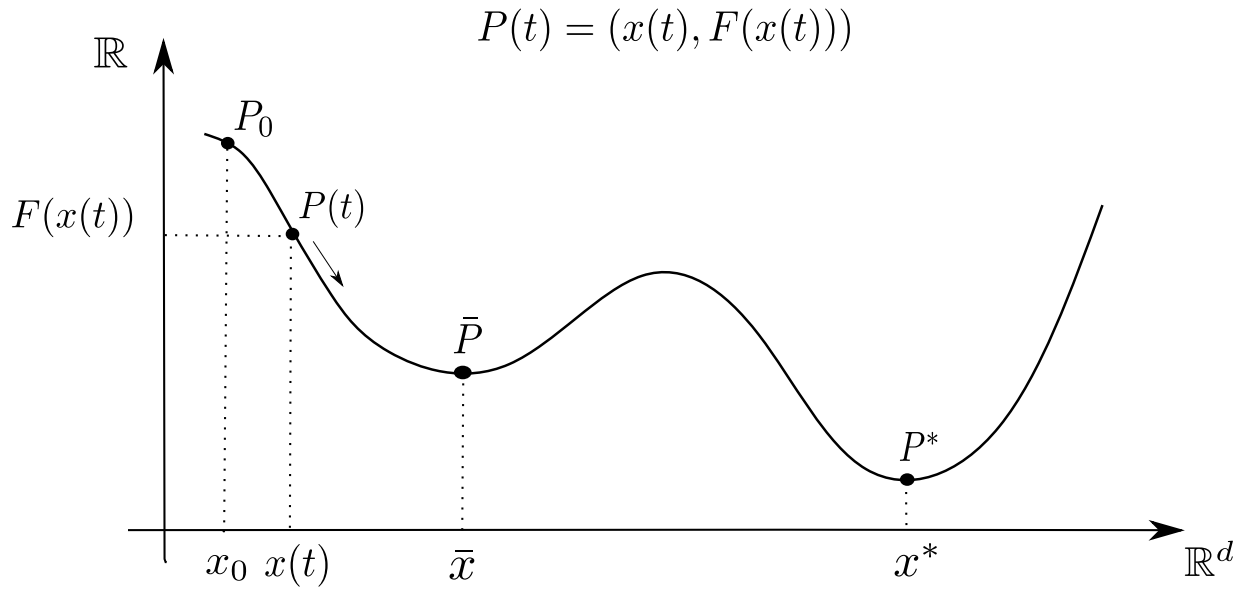
$$h(t, z) = \begin{pmatrix} -\frac{(1-e^{-at})^{-1} m}{\varepsilon + \sqrt{(1-e^{-bt})^{-1} v}} \\ a(\nabla F(x) - m) \\ b(S(x) - v) \end{pmatrix}$$

Théorème

Existence, unicité et bornitude d'une solution globale à l'EDO issue de $(x_0, 0, 0)$.

Interprétation mécanique - Heavy Ball with Friction

[Attouch et al., 2000, Cabot et al., 2009, Gadat et al., 2018]



- Force gravitationnelle (potentiel F).
- Force de frottement visqueux: $-\lambda \dot{x}(t)$ (amortissement).
- Réaction du support $\Sigma = \mathit{Graph}(F)$.

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

ADAM vu comme Heavy Ball with Friction (HBF)

HBF "généralisé"

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

- **HBF généralisé :**
  - Masse de la particle dépendante du temps.
  - Viscosité dépendante du temps.

- **Intérêt de HBF :**
  - 2nd ordre vs 1er ordre: accélération (même si oscillations).
  - Capacité à s'échapper des points selle.

Convergence vers les points stationnaires

Théorème (Convergence)

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

Argument clé : une fonction de Lyapunov pour l'EDO

- Définition :

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

- Interprétation : énergie mécanique du système dynamique
- Lemme : $t \mapsto V(t, z(t))$ est décroissante sur $(0, +\infty)$.

Du Temps Discret au Temps Continu

Temps Continu : Analyse du Système Dynamique

Temps Discret : Convergence d'ADAM

Convergence faible du processus interpolé vers la solution de l'EDO

Techniques [Benaïm and Schreiber, 2000]

Hypothèse de moment - Contrôle du bruit

Pour tout compact $K \subset \mathbb{R}^d$, il existe $r_K > 0$ tel que

$$\sup_{x \in K} \mathbb{E}(\|\nabla f(x, \xi)\|^{2+r_K}) < \infty \,.$$

Théorème

Sous les hypothèses précédentes et **l'hypothèse de moment**,

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

Convergence en Temps Long des itérées d'ADAM

Techniques [Fort and Pagès, 1999, Bianchi et al., 2019]

- Pas de convergence p.s : régime $n \to \infty$ puis $\gamma \to 0$

Théorème (convergence ergodique des itérées d'ADAM)

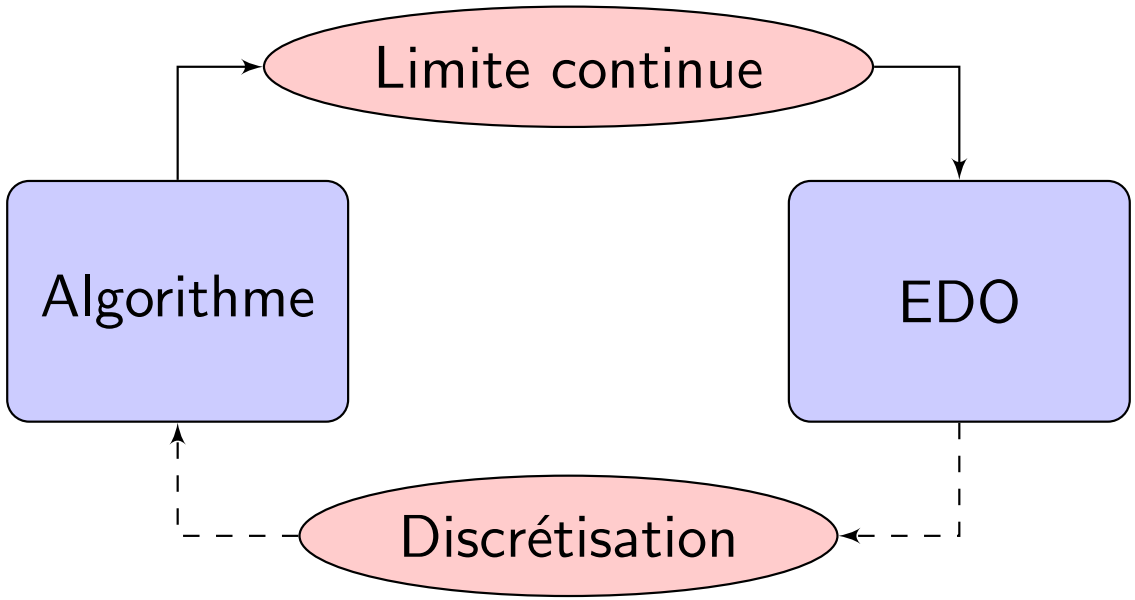
Soient $x_0 \in \mathbb{R}^d$, $\gamma > 0$, $(z_n^\gamma : n \in \mathbb{N})$, $z_0^\gamma = (x_0, 0, 0)$. Sous les mêmes hypothèses et :

- **Hypothèse de stabilité des itérées:** $\sup_{n,\gamma} \mathbb{E}\|z_n^\gamma\| < \infty$.

Alors, pour tout $\delta > 0$,

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

Conclusion



1. Introduction d'une version à temps continu d'ADAM.
   - Existence, unicité et bornitude de la solution.
   - Convergence vers les points stationnaires de F.
2. Convergence faible du processus interpolé vers la solution de l'EDO.
3. Convergence en temps long des itérées.

Merci de votre attention

Pour plus de détails : article soumis, disponible sur arXiv.

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

Simulations

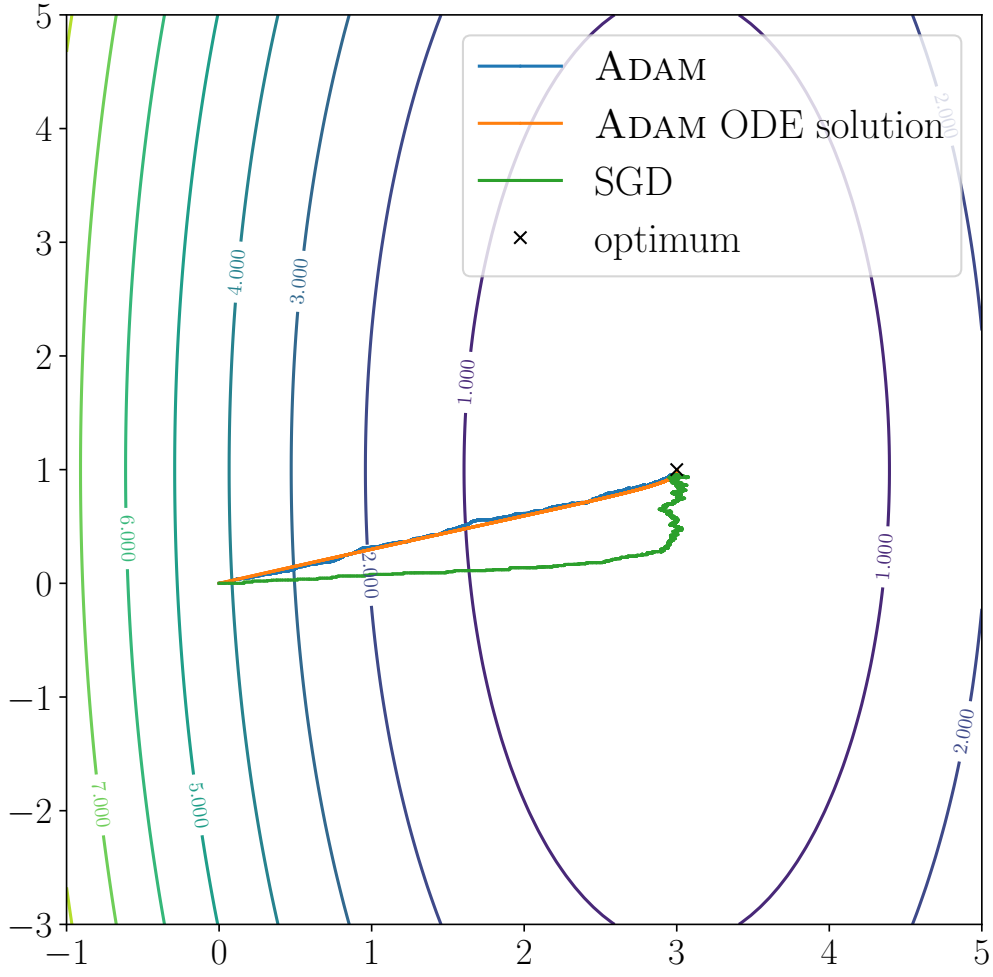


Figure 2: Convergence d'ADAM et de la solution de l'EDO vers l'optimum pour une régression linéaire 2D

Simulations

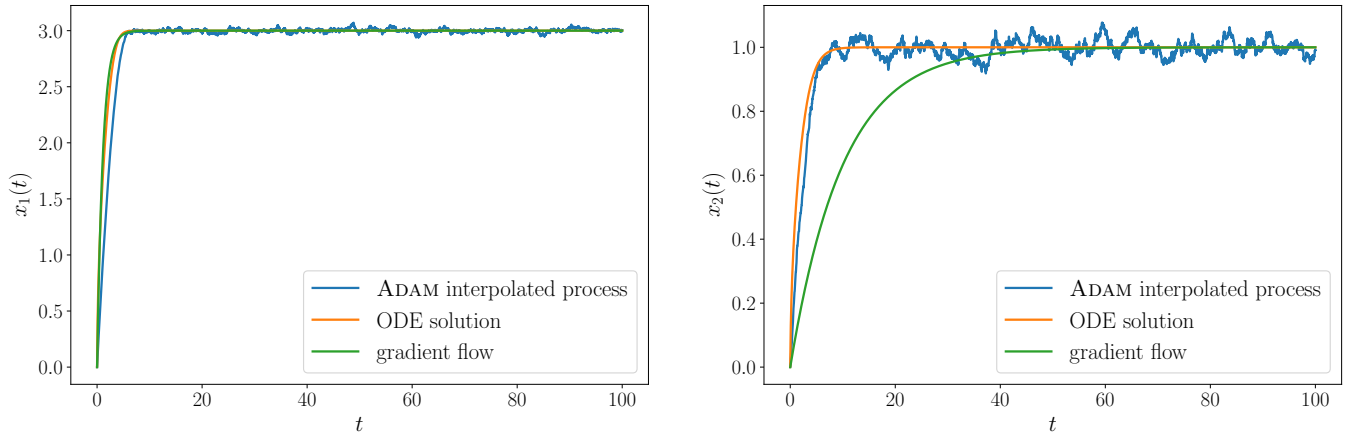


Figure 3: ADAM: interpolated process and solution to the ODE for a 2D linear regression.

Régression linéaire 2D

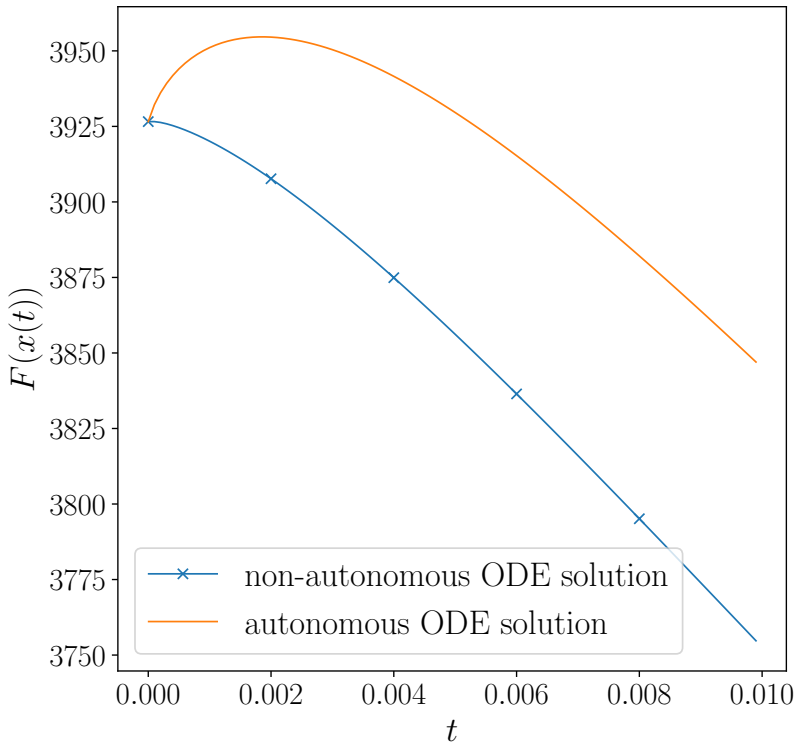
$$Y = X\,x_1^* + (1 - X)\,x_2^* + \epsilon \text{ avec } (x_1^*, x_2^*) = (3, 1).$$

$$\xi = (X, Y) \text{ avec } X \sim \mathcal{B}(p),\ p \in (0, 1).$$

$$f(\,.\,, \xi) := \frac{1}{2}\left(\left\langle \begin{pmatrix} X \\ 1 - X \end{pmatrix}, \cdot \right\rangle - Y\right)^2.$$

Intérêt du débiaisage dans ADAM

Avec le débiaisage, $F(x(t)) \leq F(x_0)$.



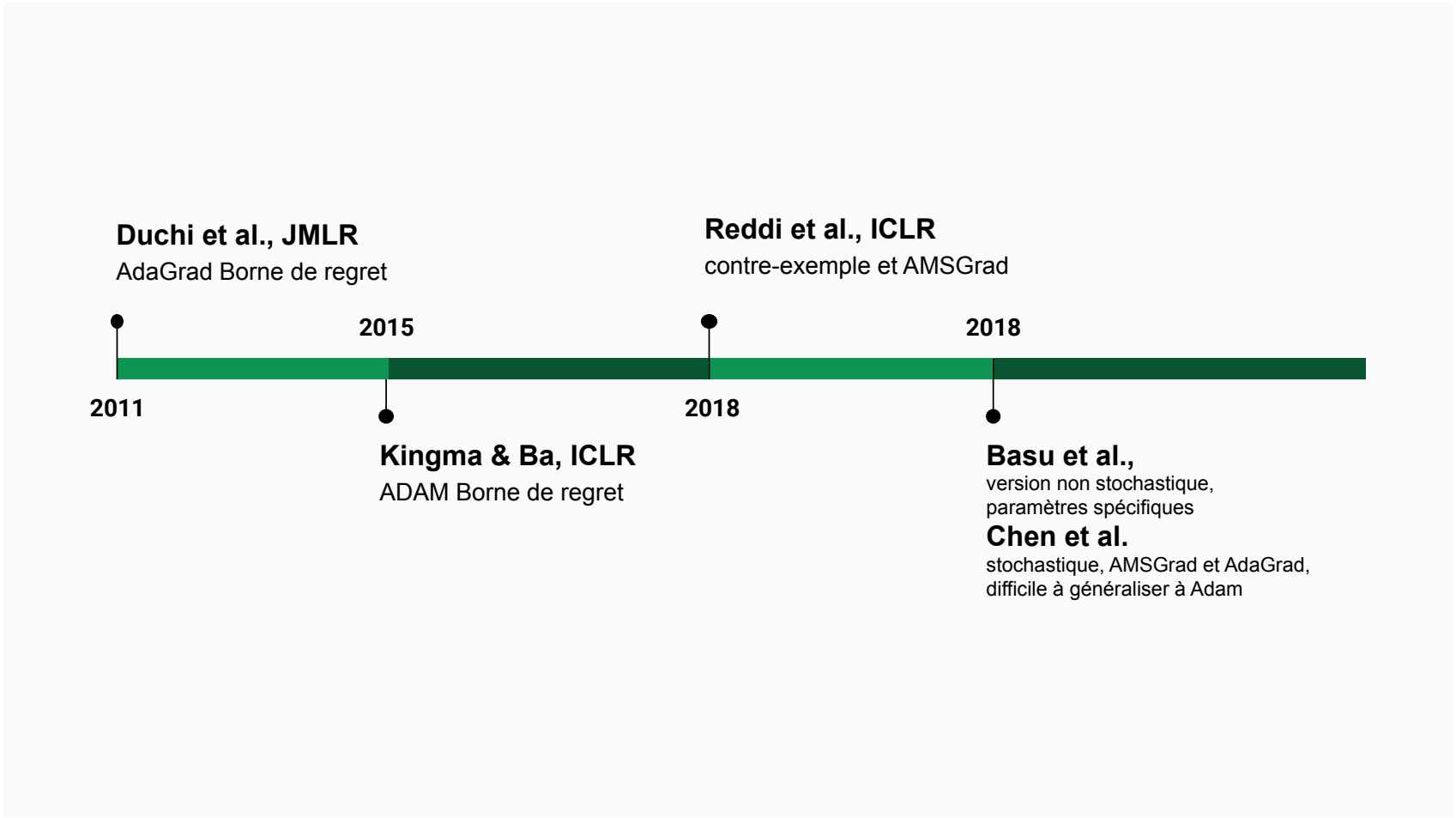
Algorithm 2 ADAM $(\gamma, \alpha, \beta, \varepsilon)$

$x_0 \in \mathbb{R}^d$, $m_0 = 0$, $v_0 = 0$, $\gamma > 0$, $\varepsilon > 0$, $(\alpha, \beta) \in [0,1)^2$.
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

Solutions de l'EDO ADAM autonome/ non autonome pour un problème stochastique quadratique en dimension 100

Revue de la littérature

ADAM: Résultats théoriques



Revue de la littérature

ADAM: résultats théoriques

- $\mathcal{O}(\frac{1}{\sqrt{T}})$ borne sur le regret moyen en non convexe.
- contre-exemple: regret moyen ne tend pas vers 0.
- AMSGRAD: variante d'ADAM
- version non bruitée d' ADAM (f n'est pas aléatoire):
  - norme du gradient petite pour un temps inconnu mais borné.
  - valeurs spécifiques des hyperparamètres d'ADAM
- Résultat similaire dans le cas stochastique pour une classe générale d'algorithmes adaptatifs
  - résultats énoncés pour AMSGRAD et ADAGRAD
  - Généralisation à ADAM sujette à des conditions difficiles à vérifier.

References I


Attouch, H., Goudou, X., and Redont, P. (2000).
The heavy ball with friction method, i. the continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system.
Communications in Contemporary Mathematics, 2(01):1–34.


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.

References II


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.


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., Panloup, F., and Saadane, S. (2018).
Stochastic heavy ball.
Electronic Journal of Statistics, 12(1):461–529.

References III


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.

References IV


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.