Garanties pour l'autodiff.

Travail en commun avec
J. Bolte & E. Pauwels

pas un talk sur des considérations pratiques
(contrainte de temps)

Focus : Pourquoi l'autodiff d'algorithme itératif fonctionne ?

1) Algorithme itératif paramétrique

$$\overset{\sim}{=} \; (*) \quad x_{k+1}(\theta) = F(x_k(\theta), \theta) = F_\theta(x(\theta))$$

F : opérateur "de point fixe"

ex. • $F(x, \theta) = x - \theta \nabla \phi(x)$. descente de gradient.

• $F(x, \theta) = \text{prox}_{\theta \psi}(x - \theta \nabla \phi(x))$ Forward Backward

condition typique : F ρ-Lip $\underline{\underline{\rho < 1}}$

$$x_k(\theta) \xrightarrow{k \to +\infty} \bar{x}(\theta)$$

$$\text{Fix } F_\theta = \bar{x}(\theta)$$

Cas de GD

$\phi \in C^{1,1}_\rho(\mathbb{R}^d)$

$\bar{x} = \text{argmin } \phi$

$x_k \to \bar{x}$.

2) Autodiff.

Si F lisse (C^1), modèle simple de différentiation automatique d'algo

→ règle de la chaîne appliqué à θ. "piggy back"

$$\frac{\partial x_{k+1}}{\partial \theta}(\theta) = A \, \frac{\partial x_k(\theta)}{\partial \theta} + B$$

$$\underbrace{\left[\partial_1 F(x_k(\theta), \theta), \quad \partial_2 F(x_k(\theta), \theta) \right]}_{\text{dérivée totale}}$$

($\simeq$ .backward de PyTorch).

$$\begin{tikzcd}
x_k(\theta) \arrow[r, "\text{autodiff.}"'] \arrow[d, "k \to \infty"] & \frac{\partial x_k(\theta)}{\partial \theta} \arrow[d, dashed, "\text{limite}"] \\
\bar{x}(\theta) \arrow[r, dashed, "\text{dérivée ``formelle''}"'] & ?
\end{tikzcd}$$

→ ? candidat naturel : $\dfrac{\partial \bar{x}}{\partial \theta}(\theta)$

Thm (Gilbert 1992) informel.

Si $\|A\| < 1$ alors $\bar{x}$ est C^1 et

$$\frac{\partial x_k(\theta)}{\partial \theta} \xrightarrow[k \to \infty]{\text{linéairement}} \frac{\partial \bar{x}}{\partial \theta}(\theta)$$

Note 1 $\|A\| < 1$ est suffisant mais pas nécessaire (Pauwels, V. SIOPT '23).

Note 2 $\bar{x}$ peut être également obtenu par <u>différentiation implicite</u>.

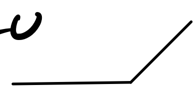
$$\frac{\partial \bar{x}}{\partial \theta}(\theta) = \left(\mathrm{Id} - \partial_1 F(\bar{x}(\theta), \theta)\right)^{-1} \partial_2 F(\bar{x}(\theta), \theta)$$

3) L'éléphant dans la salle : <u>dynamique non-lisse</u>.

<u>ex A</u> $\underbrace{\min_x \phi(x)}_{\text{pb d'optim}}$ $\qquad F(x,\theta) = x - \theta \nabla \phi(x)$

- Si $\phi \in C_p^2(\mathbb{R}^d)$.
  $\to$ F est C^1
  (Newton ?)

<u>ex B</u> F est un DEQ avec ReLU



- Si $\phi \in C_p^{1,1}(\mathbb{R}^d)$ cas typique
  $\to$ F pas C^1

Solution Jacobien conservatif.

(Bolte, Pauwels, Math Prog 21)

Def Si $f : \mathbb{R}^d \to \mathbb{R}$ continue loc. Lipschitz on dit que $\underbrace{J : \mathbb{R}^d \rightrightarrows \mathbb{R}^d}_{\text{app. multivoque}}$ est un gradient conservatif pour la fonction f (dite chemin dérivable) si

- J a un graphe fermé, loc. borné, non-vide pointwise.
- J satisfait la règle de la chaîne sur les chemins :

$\forall\, \gamma : [0,1] \to \mathbb{R}^d$ a.c wrt Lebesgue

$$\boxed{\frac{d}{dt} f(\gamma(t)) = J(\gamma(t)) \cdot \dot{\gamma}(t)}$$

Propriétés importantes :

- $\partial^c f(x) \subseteq J(x)$

Clarke.

Rademacher.

- $J(x) = \left\{ \frac{\partial f}{\partial x}(x) \right\}$ p.p

- compatible avec $+$ et $\circ$.

On peut définir de même pour les fonctions de plusieurs arguments $F(x, \theta)$.

4) Piggyback non-lisse

$$\frac{\partial x_{k+1}}{\partial \theta}(\theta) = A \frac{\partial x_k(\theta)}{\partial \theta} + B$$

$$\downarrow \qquad \downarrow$$

$$J_{x_{k+1}}(\theta) = \underset{?}{A} \; \underbrace{J_{x_k}(\theta)}_{\text{lequel ??}} + \underset{?}{B}$$

$\hookrightarrow$ $$J_{x_{k+1}}(\theta) = \left\{ AJ + B \;\middle|\; \begin{array}{l} [A, B] \in J_F(x_k(\theta), \theta) \\ J \in J_{x_k}(\theta) \end{array} \right\}$$

PyTorch

$\hookrightarrow$ sélection

$$\boxed{J_{k+1} = A_k J_k + B_k}.$$

avec $[A_k, B_k] \in J_F(x_k(\theta), \theta)$

5) Une règle de la chaîne infinie

Hypothèse J_F "contracte"

Assumption 1 (The conservative Jacobian of the iteration mapping is a contraction) F is locally Lipschitz, path differentiable, jointly in (x, θ), and J_F is a conservative Jacobian for F. There exists $0 \le \rho < 1$, such that for any $(x, \theta) \in \mathbb{R}^p \times \mathbb{R}^m$ and any pair $[A, B] \in J_F(x, \theta)$, with $A \in \mathbb{R}^{p \times p}$ and $B \in \mathbb{R}^{p \times m}$, the operator norm of A is at most ρ.

$$J_{\bar{x}}^{pb} : \theta \rightrightarrows \mathrm{Fix}\left[J_F(\bar{x}(\theta), \theta) \right]$$

$$= \mathrm{Fix}\left[J_F(\mathrm{Fix}[F_\theta], \theta) \right]$$

au sens des appli multivoque.

au sens des applications univoques

Thm (BPV 22) $J_{\bar{x}}^{pb}$ est bien définie.

$\simeq$ Banach-Picard point fixe pour les récursions sur les ensemble.

Cor $\lim_{k \to +\infty} \mathrm{gap}\left(J_{x_k}(\theta), J_{\bar{x}}^{pb}(\theta) \right) = 0$

$$\Leftrightarrow \lim_{k \to +\infty} \mathrm{gap}\left(\prod_{\ell=0}^{k} \mathcal{J}_F(x_\ell(\theta), \theta) \,,\, \mathcal{J}_{\bar{x}}^{pb}(\theta) \right) = 0 .$$

Version "unrolled"

2) Pour presque tout θ,

$$\frac{\partial x_k(\theta)}{\partial \theta} \xrightarrow{k \to +\infty} \frac{\partial \bar{x}(\theta)}{\partial \theta} \quad !$$

Note différentiation implicite pas compatible.

Note Si F a une structure "semi-alg" alors version quantitative (convergence linéaire).