Automatic differentiation of nonsmooth iterative algorithms

Jérôme Bolte[1], Edouard Pauwels[2], Samuel Vaiter[3] (NeurIPS 2022)
[1]TSE, Univ. Toulouse [2]IRIT, Univ. Toulouse [3]CNRS & Univ. Côte d'Azur

Summary

We characterize the attractor set of **nonsmooth piggyback iterations** as a **set-valued fixed point** which remains in the **conservative framework**.

$$\begin{tikzcd} x_k(\theta) \arrow[r, "\text{nonsmooth}", "\text{autodiff}"'] \arrow[d, "k\to\infty"'] & J_{x_k}(\theta) \arrow[d, squiggly, "\text{limit?}"] \\ \bar{x}(\theta) \arrow[r, squiggly, "\text{derivative?}"'] & J^{\text{pb}}_{\bar{x}}(\theta) \end{tikzcd}$$

- Piggyback propagation, *i.e.*, **differentiation along algorithms** is well understood [1] in the smooth case . We extend such results to nonsmooth problems.
- Our main assumption is **nonexpansivity conditions** on the algorithm studied.

Conservative Jacobian

Definition [2]. $f : \mathbb{R}^p \to \mathbb{R}^m$ locally Lipschitz. The set-valued $J : \mathbb{R}^p \rightrightarrows \mathbb{R}^{m\times p}$ is a *conservative Jacobian* for the **path differentiable** f if J is closed, locally bounded and nowhere empty with

$$\frac{d}{dt} f(\gamma(t)) = J(\gamma(t))\dot{\gamma}(t) \quad \text{a.e.}$$

for any $\gamma \colon [0,1] \to \mathbb{R}^p$ absolutely continuous with respect to the Lebesgue measure.

Fixed point of affine iterations

- $\mathcal{J} \subset \mathbb{R}^{p\times(p+m)}$: compact set of matrices such that
$$\forall [A,B] \in \mathcal{J}, \|A\|_{op} \leqslant \rho.$$
- **Action of $\mathcal{J}$ on matrices** of size $p \times m$
$$\mathcal{J} : X \rightrightarrows \{AX + B,\ [A,B] \in \mathcal{J}\}$$
- (Extended) **action of $\mathcal{J}$ on set of matrices**
$$\mathcal{J} : \mathcal{X} \rightrightarrows \{AX + B,\ [A,B] \in \mathcal{J},\ X \in \mathcal{X}\}.$$
- **Recursive action** of $\mathcal{J}$ on $(\mathcal{X}_k)_{k\in\mathbb{N}}$
$$\mathcal{X}_{k+1} = \mathcal{J}(\mathcal{X}_k) \quad \forall k \in \mathbb{N}.$$

Theorem 1 (**Set-valued affine contractions**). There is a unique nonempty compact set $\operatorname{fix}(\mathcal{J})$ satisfying $\operatorname{fix}(\mathcal{J}) = \mathcal{J}(\operatorname{fix}(\mathcal{J}))$,

$$\forall k \in \mathbb{N}, \quad \operatorname{dist}(\mathcal{X}_k, \operatorname{fix}(\mathcal{J})) \leqslant \rho^k \frac{\operatorname{dist}(\mathcal{X}_0, \mathcal{J}(\mathcal{X}_0))}{1-\rho}.$$

Iterative algorithm

Iterative algorithm. Pair of a **Lipschitz function** $F \colon \mathbb{R}^p \times \mathbb{R}^m \mapsto \mathbb{R}^p$ parameterized by $\theta \in \mathbb{R}^m$, with Lipschitz initialization $x_0 \colon \theta \mapsto x_0(\theta)$ and

$$x_{k+1}(\theta) = F(x_k(\theta), \theta) = F_\theta(x_k(\theta)),$$

where $F_\theta := F(\cdot, \theta)$, under the assumption that $x_k(\theta)$ converges to the unique fixed point of F_θ: $\bar{x}(\theta) = \operatorname{fix}(F_\theta)$.
Examples. **gradient descent** $F(x,\theta) = x - \theta\nabla h(x)$, deep equilibrium network.

Piggyback differentiation of iterative algorithms

Chain rule applied to smooth iterative algorithms ("Piggyback" recursion).

$$\frac{\partial}{\partial\theta} x_{k+1}(\theta) = \partial_1 F(x_k(\theta), \theta) \cdot \frac{\partial}{\partial\theta} x_k(\theta) + \partial_2 F(x_k(\theta), \theta), \qquad \text{(PB-S)}$$

where $\frac{\partial}{\partial\theta} x_k$ is the Jacobian of x_k with respect to θ.

Assumption A (**The conservative Jacobian of the iterations is a contraction**). F is locally Lipschitz, path differentiable, jointly in (x,θ), and J_F is a conservative Jacobian for F. There exists $0 \leqslant \rho < 1$, 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_{x_0} is a conservative Jacobian for the initialization function $\theta \mapsto x_0(\theta)$.

Under Assumption A, F_θ is a strict contraction: $(x_k(\theta))_k$ converges linearly to $\bar{x}(\theta) = \operatorname{fix}(F_\theta)$.
Chain rule applied to *non*smooth iterative algorithms ("Piggyback" recursion).

$$J_{x_{k+1}}(\theta) = \{AJ + B,\ [A,B] \in J_F(x_k(\theta), \theta),\ J \in J_{x_k}(\theta)\}. \qquad \text{(PB-NS)}$$

Main result: infinite chain rule

Set-valued (piggyback) map based on the fix operator from Theorem 1,

$$J^{\text{pb}}_{\bar{x}} \colon \theta \rightrightarrows \operatorname{fix}[J_F(\bar{x}(\theta), \theta)] = \operatorname{fix}[J_F(\operatorname{fix}(F_\theta), \theta)].$$

Theorem 2 (**Conservative mapping for the fixed point map**) Under Assumption A, $J^{\text{pb}}_{\bar{x}}$ is a conservative Jacobian for the fixed point map $\bar{x}$, and:

$$\text{for all } \theta, \quad \lim_{k\to\infty} \operatorname{gap}(J_{x_k}(\theta), J^{\text{pb}}_{\bar{x}}(\theta)) = 0;$$

$$\text{for **almost** all } \theta, \quad \lim_{k\to\infty} \frac{\partial}{\partial\theta} x_k(\theta) = \frac{\partial}{\partial\theta}\bar{x}(\theta),$$

where $\operatorname{gap}(\mathcal{X}, \mathcal{Y}) = \max_{x\in\mathcal{X}} d(x, \mathcal{Y})$, and $d(x, \mathcal{Y}) = \min_{y\in\mathcal{Y}} \|x - y\|$.

- **Limit-derivative exchange**: *Asymptotically, the gap between the differentiation of x_k and the derivative of the limit is zero.* (can be shown to be linear under additional hypotheses,)
- Assuming that for every $[A,B] \in J(\bar{x}(\theta), \theta)$, the matrix $I - A$ is invertible, we have [3]
$$J^{\text{imp}}_{\bar{x}} \colon \theta \rightrightarrows \left\{(I-A)^{-1}B, [A,B] \in J_F(\bar{x}(\theta), \theta)\right\}$$
is a conservative Jacobian for $\bar{x}$ (**implicit differentiation**). Under Assumption A, one has $J^{\text{imp}}_{\bar{x}}(\theta) \subset J^{\text{pb}}_{\bar{x}}(\theta)$. If F is not differentiable, the inclusion may be strict.

Consequence for automatic differentiation

Input: $k \in \mathbb{N}$, $\theta \in \mathbb{R}^m$, $\dot{\theta} \in \mathbb{R}^m$, $\bar{w}_k \in \mathbb{R}^p$. Initialize: $x_0 = x_0(\theta) \in \mathbb{R}^p$.

Forward mode (JVP):
$\dot{x}_0 = J\dot{\theta}$, $J \in J_{x_0}(\theta)$.
for $i = 1, \dots, k$ **do**
  $x_i = F(x_{i-1}, \theta)$
  $\dot{x}_i = A_{i-1}\dot{x}_{i-1} + B_{i-1}\dot{\theta}$
  $[A_{i-1}, B_{i-1}] \in J_F(x_{i-1}, \theta)$
Return: $\dot{x}_k$

Reverse mode (VJP): $\bar{\theta}_k = 0$.
for $i = 1, \dots, k$ **do**
  $x_i = F(x_{i-1}, \theta)$
for $i = k, \dots, 1$ **do**
  $\bar{\theta}_k = \bar{\theta}_k + B_{i-1}^T \bar{w}_i \quad \bar{w}_{i-1} = A_{i-1}^T \bar{w}_i$
  $[A_{i-1}, B_{i-1}] \in J_F(x_{i-1}, \theta)$
$\bar{\theta}_k = \bar{\theta}_k + J^T \bar{w}_0$, $J \in J_{x_0}(\theta)$
Return: $\bar{\theta}_k$

Theorem 3 (**Convergence of JVP and VJP**).
- (JVP). For almost all $\theta \in \mathbb{R}^m$, $\dot{x}_k \to \frac{\partial\bar{x}}{\partial\theta}\dot{\theta}$.
- (VJP). Assume that $\lim_{k\to\infty} \bar{w}_k = \bar{w}$ (for example, $\bar{w}_k = \nabla\ell(x_k)$ for a C^1 loss ℓ), then for almost all $\theta \in \mathbb{R}^m$, $\bar{\theta}_k^T \to \bar{w}^T \frac{\partial\bar{x}}{\partial\theta}$.

[1] Gilbert. **Automatic differentiation and iterative processes**. *Opt. Met. Soft.*, 1992.
[2] Bolte, Pauwels. **Conservative set valued fields, automatic differentiation, stochastic gradient method and deep learning**. *Math. Prog.* 2020.
[3] Bolte, Le, Pauwels, Silveti-Falls. **Nonsmooth Implicit Differentiation for Machine-Learning and Optimization**. *NeurIPS*. 2021.

Applications to proximal methods

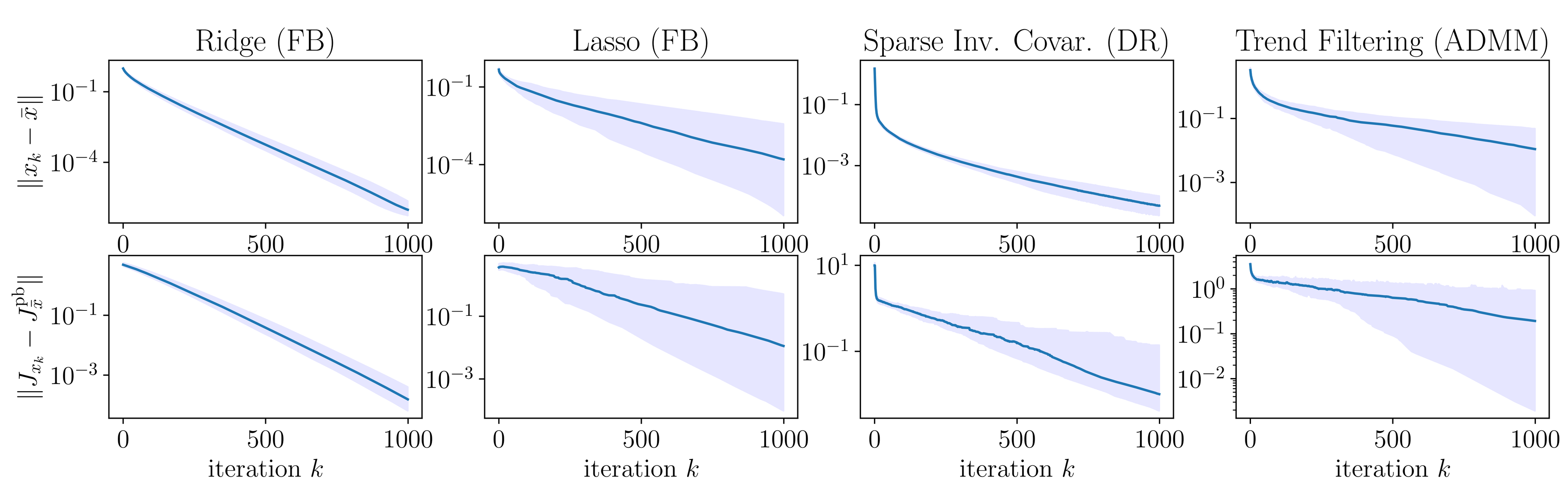
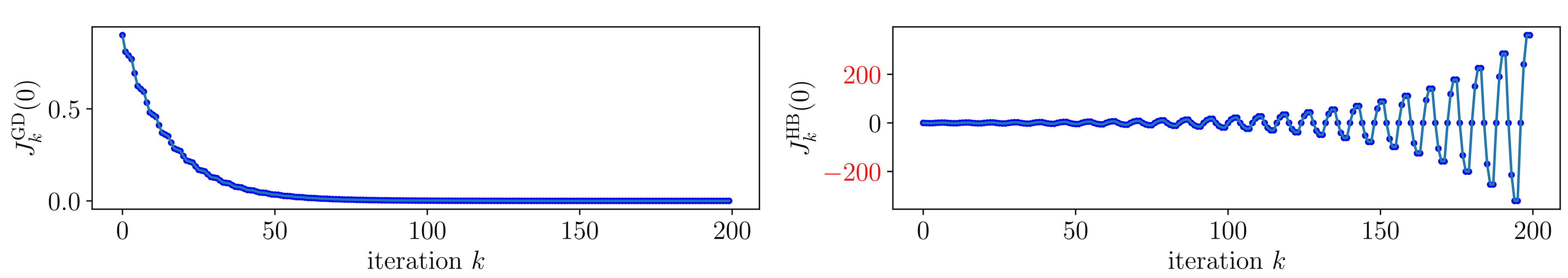


Illustration of the linear convergence. (First line) Distance of the iterates to the fixed point. (Second line) Distance of the piggyback Jacobians to the Jacobian of the fixed point.

Failure of inertial methods



Behavior of automatic differentiation for first-order methods on a piecewise quadratic function. (Left) Stability of the propagation of derivatives for the fixed step-size gradient descent. (Right) Instability of the propagation of Heavy-Ball initialized.