Automatic differentiation of nonsmooth iterative algorithms

Jérôme Bolte
Toulouse School of Economics,
University of Toulouse Capitole.
Toulouse, France.

Edouard Pauwels
IRIT, CNRS, Université de Toulouse.
Institut Universitaire de France (IUF).
Toulouse, France.

Samuel Vaiter
CNRS & Université Côte d'Azur,
Laboratoire J. A. Dieudonné.
Nice, France.

Abstract

Differentiation along algorithms, i.e., piggyback propagation of derivatives, is now routinely used to differentiate iterative solvers in differentiable programming. Asymptotics is well understood for many smooth problems but the nondifferentiable case is hardly considered. Is there a limiting object for nonsmooth piggyback automatic differentiation (AD)? Does it have any variational meaning and can it be used effectively in machine learning? Is there a connection with classical derivative? All these questions are addressed under appropriate nonexpansivity conditions in the framework of conservative derivatives which has proved useful in understanding nonsmooth AD. We characterize the attractor set of nonsmooth piggyback iterations as a set-valued fixed point which remains in the conservative framework. Among various consequences we have almost everywhere convergence of classical derivatives. Our results are illustrated on parametric convex optimization with forward-backward, Douglas-Rachford and Alternating Direction of Multiplier algorithms as well as the Heavy-Ball method.

1 Introduction

Differentiable programming. We consider a Lipschitz function $F\colon \mathbb{R}^p \times \mathbb{R}^m \mapsto \mathbb{R}^p$, representing an iterative algorithm, parameterized by $\theta \in \mathbb{R}^m$, with Lipschitz initialization $x_0\colon \theta \mapsto x_0(\theta)$ and

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

where $F_\theta := F(\cdot, \theta)$, under the assumption that $x_k(\theta)$ converges to the unique fixed point of F_θ: $\bar{x}(\theta) = \mathrm{fix}(F_\theta)$. Such recursion represent for instance algorithms to solve an optimization problem $\min_x h(x)$ (e.g. empirical risk minimization), such as gradient descent: $F(x, \theta) = x - \theta \nabla h(x)$. But (1) could also be a fixed-point equation such as a deep equilibrium network [5].

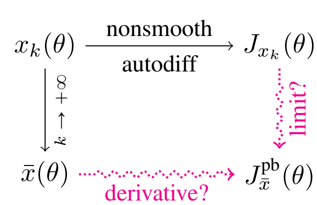

Figure 1: We study existence and meaning of $J^{\mathrm{pb}}_{\bar{x}}$ as a derivative of $\bar{x}$, compatible with automatic differentiation of the iterates $(x_k(\theta))_{k\in\mathbb{N}}$.

In the last years, a paradigm shift occurred: such algorithms are now implemented in algorithmic differentiation (AD)-friendly frameworks such as Tensorflow [1], PyTorch [42] or JAX [13]. For a differentiable F, it is possible to compute iteratively the derivatives of

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

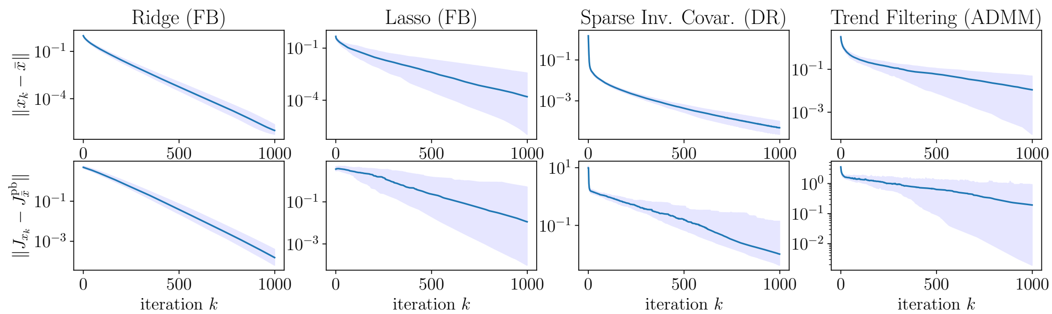

Figure 2: Illustration of the linear convergence of proximal splitting methods. *(First line)* Distance of the iterates to the fixed point. *(Second line)* Distance of the piggyback Jacobians to the Jacobian of the fixed point. The acronyms are FB for Forward-Backward, DR for Douglas-Rachford and ADMM for Alternating Direction Method of Multipliers. In all cases, despite nonsmoothness, piggyback Jacobians converge, illustrating Corollary 2. Blue lines represent the median of 100 repetitions with random data, and the blue shaded area represents the first and last deciles.

x_k with respect to θ using the differential calculus rules resulting in so called "piggyback" recursion:

$$\frac{\partial}{\partial \theta} x_{k+1}(\theta) = \partial_1 F(x_k(\theta), \theta) \cdot \frac{\partial}{\partial \theta} x_k(\theta) + \partial_2 F(x_k(\theta), \theta), \tag{2}$$

where $\frac{\partial}{\partial \theta} x_k$ is the Jacobian of x_k with respect to θ. In practice, automatic differentiation frameworks do not compute the full Jacobian, but compute either vector-Jacobian products in reverse-mode (or backpropagation) [48] or Jacobian-vector products in forward mode [53]. We rather consider the full Jacobian, and therefore, our findings *apply to both* modes. We focus on two issues arising with nonsmooth recursions, illustrated in Figure 1. *(i)* what can be said about the chain rule (2) and its asymptotics when the function F is not smooth (for example a projected gradient step)? *(ii)* how to interpret its asymptotics as a notion of derivative for $\bar{x}$, the fixed point of F_θ? We propose a *joint* answer to both questions, providing a solid theoretical ground to the idea of algorithmic differentiation of numerical solvers involving nonsmooth components in a differentiable programming context.

Related works. Algorithmic use of the chain rule (2) to differentiate programs takes its root in [53], with forward differentiation, and later in reverse mode [35]. Along with the development of AD, convergence of the iterative sequence (2) was investigated, notably in the optimization community as reviewed in [28]. This important survey paper gathers results in differentiable programming rediscovered/reused later: implicit differentiation [43, 45] and its stability [8], adjoint fixed point iteration [5] (a key aspect of the deep equilibrium network) and linear convergence of (2). Notably, linear convergence of Jacobians was investigated in [25, 27] for the forward mode and in [15, Theorem 2.3] for the reverse mode. This was more recently investigated – for C^2 *functions* – in imaging for primal-dual algorithms [14, 9] and in machine learning for gradient descent [39, 36] and the Heavy-ball [39] method. In the specific context where F solves a $\min$-$\min$ problem, the authors in [2] proved the linear convergence of the Jacobians. The use of AD for nonsmooth functions was justified with the notion of *conservative Jacobians* [12, 11] with a nonsmooth version of the chain rule for compositional models. Correctness of AD was also investigated in [34] for a large class of piecewise analytic functions, and in [33] where a qualification condition is used to compute a Clarke Jacobian. Along with AD, a natural way to differentiate a model (1) is by implicit differentiation, recently applied in several works [5, 3, 21]. In a nonsmooth context, an implicit function theorem [10] was proved for path-differentiable functions. In terms of applications, nonsmooth piggyback derivatives are applied to hyperparameter tuning for inverse problems in [8] while the case of Lasso was investigated in [7]. Other relevant applications include plug-and-play denoising [32], parameter selection [19], bilevel programming [41]

Contributions: Under suitable nonexpansivity assumptions, our contributions are as follows.

• We address both questions illustrated in Figure 1 for nonsmooth recursions. Set-valued extensions of the piggyback recursion (2) have a well defined limit: the fixed point of subset map (Theorem 1), it is conservative for the fixed point map $\bar{x}$. This is a nonsmooth "infinite" chain rule for AD (Theorem 2).

• For almost all θ, despite nonsmoothness, recursion (2) is well defined using the classical Jacobian and converges to the classical Jacobian of the fixed point $\bar{x}$ (Corollary 2). This has implications for both forward and reverse modes of AD.
• For a large class of functions (Lipschitz-gradient selection), it is possible to give a quantitative rate estimate (Corollary 3), namely to prove linear convergence of the derivatives.
• We show that these results can be applied to proximal splitting algorithms in nonsmooth convex optimization. We include forward–backward (Proposition 2), as well Douglas–Rachford (Proposition 3) and ADMM, a numerical illustration of the convergence of derivatives is given in Figure 2.
• We also illustrate that, contrarily to the smooth case, nonsmooth piggy back derivatives of momentum methods such as Heavy-ball, may diverge even if the iterates converge linearly (Proposition 4).

Notations. A function $f : \mathbb{R}^p \to \mathbb{R}^m$ is locally Lipschtiz if, for each $x \in \mathbb{R}^n$, there exists a neighborhood of x on which f is Lipschitz. Denoting by $R \subseteq \mathbb{R}^p$, the full measure set where f is differentiable, the Clarke Jacobian [16] at $x \in \mathbb{R}^p$ is defined as

$$\mathrm{Jac}^{\,c} f(x) = \mathrm{conv}\left\{ M \in \mathbb{R}^{p\times m}, \exists (y_k)_{k\geqslant 0} \text{ s.t. } \lim_{k\to\infty} y_k = x, y_k \in R, \lim_{k\to\infty} \frac{\partial f}{\partial y}(y_k) = M \right\}. \quad (3)$$

The Clarke subdifferential, $\partial^c f$ is defined similarly. Given two matrices A, B with compatible dimension, $[A, B]$ is their concatenation. For a set $\mathcal{X}$, $\mathrm{conv}\mathcal{X}$ is its convex hull. The symbol $\mathbb{B}$ denotes a unit ball, the corresponding norm should be clear from the context.

2 Nonsmooth piggyback differentiation

We first show how the use of the notion of *conservative Jacobians* allow us to justify rigorously the existence of a nonsmooth equivalent of piggyback iterations in (2) that is compatible with AD.

Conservative Jacobians. Conservative Jacobians were introduced in [12] as a generalization of derivatives to study automatic differentiation of nonsmooth functions. Given a locally Lipschitz continuous function $f : \mathbb{R}^p \to \mathbb{R}^m$, the set-valued $J : \mathbb{R}^p \rightrightarrows \mathbb{R}^{m\times p}$ is a *conservative Jacobian* for the *path differentiable* f if J has a closed graph, is locally bounded and nowhere empty with

$$\frac{d}{dt} f(\gamma(t)) = J(\gamma(t))\dot{\gamma}(t) \quad \text{a.e.} \quad (4)$$

for any $\gamma \colon [0, 1] \to \mathbb{R}^p$ absolutely continuous with respect to the Lebesgue measure. Conservative gradients are defined similarly. We refer to [12] for extensive examples and properties of this class of function, key ideas are recalled in Appendix A for completeness. Let us mention that the classes of convex functions, definable functions, or semialgebraic functions are contained in the set of path differentiable functions. Given $D_f \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, a conservative gradient for $f \colon \mathbb{R}^p \to \mathbb{R}$, we have:

• (**Clarke subgradient**), for all $x \in \mathbb{R}^p$, $\partial^c f(x) \subset \mathrm{conv}(D_f(x))$.
• (**Gradient almost everywhere**) $D_f(x) = \{\nabla f(x)\}$ for almost all $x \in \mathbb{R}^p$.
• (**Calculus**) differential calculus rules preserve conservativity, *e.g.* sum and compositions of conservative Jacobians are conservative Jacobians.

Finally, D_f can be used as a first order optimization oracle for methods of gradient type [11].

Piggyback differentiation of recursive algorithms. The following is standing throughout the text.

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 $0 \leqslant \rho < 1$, 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 1, F_θ is a strict contraction so that $(x_k(\theta))_{k\in\mathbb{N}}$ converges linearly to $\bar{x}(\theta) = \mathrm{fix}(F_\theta)$ the unique fixed point of the iteration mapping F_θ. More precisely, for all $k \in \mathbb{N}$, we have

$$\|x_k(\theta) - \bar{x}(\theta)\| \leqslant \rho^k \frac{\|x_0 - F_\theta(x_0)\|}{1-\rho}.$$

Furthermore, for every $k \in \mathbb{N}$, let us define the following set-valued piggyback recursion:

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

We will show in Section 3 that (PB) plays the same role as (2) in the nonsmooth setting. Note that one can recursively evaluates a sequence $J_k \in J_{x_k}$, $k \in \mathbb{N}$, through operations actually implemented in nonsmooth AD frameworks, as follows

$$J_{k+1} = A_k J_k + B_k \quad \text{where} \quad [A_k, B_k] \in J_F(x_k(\theta), \theta), \tag{5}$$

Remark 1 (Local contractions) Assumption 1 may be relaxed as follows: for all θ, the fixed point set $\mathrm{fix}(F_\theta)$ is a singleton $\bar{x}_\theta$ such that $x_k(\theta) \to \bar{x}(\theta)$ as $k \to \infty$, and the operator norm condition on J_F in Assumption 1 holds at the point $(\bar{x}(\theta), \theta)$. By graph closedness of J_F, in a neighborhood of $(\bar{x}(\theta), \theta)$, F_θ is a strict contraction and the operator norm condition on J_F holds, possibly with a larger contraction factor ρ. After finitely many steps, the iterates $(x_k)_{k\in\mathbb{N}}$ remain on some neighborhood and all our convergence results hold, due to their asymptotic nature.

Remark 2 (Relation to existing work) For a smooth F a natural conservative Jacobian is the classical one. The, hypotheses in [39, 36] for gradient descent (F is C^1), are exactly the classical counterpart of Assumption 1. On the other hand [25, 27, 15] use a more general assumption on spectral radius, which allow to treat the Heavy-Ball method, *e.g.* in [39]. However this does not generalize to sets of matrices, as shown in Section 5. Hence Assumption 1 is on operator norm and not on spectral radius, which is sharp, contrary to the smooth case.

3 Asymptotics of nonsmooth piggyback differentiation

3.1 Fixed point of affine iterations

Gap and Haussdorf distance. Being given two nonempty compact subsets $\mathcal{X}, \mathcal{Y}$ of $\mathbb{R}^p$, set

$$\mathrm{gap}(\mathcal{X}, \mathcal{Y}) = \max_{x\in\mathcal{X}} d(x, \mathcal{Y}) \quad \text{where} \quad d(x, \mathcal{Y}) = \min_{y\in\mathcal{Y}} \|x - y\|,$$

and define the Hausdorff distance between $\mathcal{X}$ and $\mathcal{Y}$ by $\mathrm{dist}(\mathcal{X}, \mathcal{Y}) = \max(\mathrm{gap}(\mathcal{X}, \mathcal{Y}), \mathrm{gap}(\mathcal{Y}, \mathcal{X}))$. Note that $\mathrm{gap}(\mathcal{X}, \mathcal{Y}) = 0$ if, and only if, $\mathcal{X} \subseteq \mathcal{Y}$, whereas $\mathrm{dist}(\mathcal{X}, \mathcal{Y}) = 0$ if, and only if, $\mathcal{X} = \mathcal{Y}$. Moreover, $\mathcal{X} \subseteq \mathcal{Y} + \mathrm{gap}(\mathcal{X}, \mathcal{Y})\mathbb{B}$ where $\mathbb{B}$ is the unit ball. It means that $\mathrm{gap}(\mathcal{X}, \mathcal{Y})$ "measures" the default of inclusion of $\mathcal{X}$ in $\mathcal{Y}$, see [46, Chapter 4] for more details.

Affine iterations by packets of matrices. Let $\mathcal{J} \subset \mathbb{R}^{p\times(p+m)}$ be a compact subset of matrices such that any matrix of the form $[A, B] \in \mathcal{J}$ with $A \in \mathbb{R}^{p\times p}$ is such that A has operator norm at most $\rho < 1$. We let $\mathcal{J}$ act naturally on the matrices of size $p \times m$ as follows $\mathcal{J}\colon \mathbb{R}^{p\times m} \rightrightarrows \mathbb{R}^{p\times m}$ the function from $\mathbb{R}^{p\times m}$ to subsets of $\mathbb{R}^{p\times m}$ which is defined for each $X \in \mathbb{R}^{p\times m}$ as follows: $\mathcal{J}(X) = \{AX + B,\ [A, B] \in \mathcal{J}\}$. This defines a set-valued map through, for any $\mathcal{X} \subset \mathbb{R}^{p\times m}$,

$$\mathcal{J}(\mathcal{X}) = \{AX + B,\ [A, B] \in \mathcal{J},\ X \in \mathcal{X}\}. \tag{6}$$

Recursions of the form (PB) generate sequences $(\mathcal{X}_k)_{k\in\mathbb{N}}$ of subsets of $\mathbb{R}^{p\times m}$ satisfying

$$\mathcal{X}_{k+1} = \mathcal{J}(\mathcal{X}_k) \quad \forall k \in \mathbb{N}. \tag{7}$$

The following is an instance of the Banach–Picard theorem (whose proof is recalled in Appendix B).

Theorem 1 (Set-valued affine contractions) *Let $\mathcal{J} \subset \mathbb{R}^{p\times(p+m)}$ be a compact subset of matrices as above with $\rho < 1$. Then there is a unique nonempty compact set $\mathrm{fix}(\mathcal{J}) \subset \mathbb{R}^{p\times m}$ satisfying $\mathrm{fix}(\mathcal{J}) = \mathcal{J}(\mathrm{fix}(\mathcal{J}))$, where the action of $\mathcal{J}$ is given in (6).*

Let $(\mathcal{X}_k)_{k\in\mathbb{N}}$ be a sequence of compact subsets of $\mathbb{R}^{p\times m}$, such that $\mathcal{X}_0 \neq \emptyset$, and satisfying the recursion (7). We have for all $k \in \mathbb{N}$

$$\mathrm{dist}(\mathcal{X}_k, \mathrm{fix}(\mathcal{J})) \leqslant \rho^k \frac{\mathrm{dist}(\mathcal{X}_0, \mathcal{J}(\mathcal{X}_0))}{1 - \rho},$$

where dist *is the Hausdorff distance related to the Euclidean norm on* $p \times m$ *matrices.*

3.2 An infinite chain rule and its consequences

Define the following set-valued map based on the fix operator from Theorem 1,

$$J_{\bar{x}}^{\text{pb}} \colon \theta \rightrightarrows \text{fix}\left[J_F(\bar{x}(\theta), \theta)\right].$$

where $\bar{x}(\theta)$ is the unique fixed point of the algorithmic recursion. Since $\bar{x}(\theta) = \text{fix}(F_\theta)$, we have equivalently that $J_{\bar{x}}^{\text{pb}}$ is the fixed-point of the Jacobian at the fixed-point: $J_{\bar{x}}^{\text{pb}} \colon \theta \rightrightarrows \text{fix}\left[J_F(\text{fix}(F_\theta), \theta)\right]$. We have the following (proved in Appendix C) and a consequence from Theorem 1.

Theorem 2 (A conservative mapping for the fixed point map) *Under Assumption 1, $J_{\bar{x}}^{\text{pb}}$ is well-defined, and is a conservative Jacobian for the fixed point map $\bar{x}$.*

Corollary 1 (Convergence of the piggyback derivatives) *Under Assumption 1, for all θ, the recursion (PB) satisfies*

$$\lim_{k \to \infty} \text{gap}(J_{x_k}(\theta), J_{\bar{x}}^{\text{pb}}(\theta)) = 0. \tag{8}$$

Unrolling the expression of J_{x_k}, using (6) and (7), we can rewrite (8) with a compositional product:

$$\lim_{K \to +\infty} \text{gap}\left(\left(\bigcirc_{k=0}^{K} J_F(x_k(\theta), \theta)\right)(J_{x_0}(\theta)), J_{\bar{x}}^{\text{pb}}(\theta)\right) = 0.$$

In plain words, this is a limit-derivative exchange result: *Asymptotically, the gap between the automatic differentiation of x_k and the derivative of the limit is zero.* In particular the recursion (5) produces bounded sequences whose accumulation points are in $J_{\bar{x}}^{\text{pb}}$. Since conservative Jacobians equal classical Jacobians almost everywhere [12], we have convergence of classical derivatives.

Corollary 2 (Convergence a.e. of the classical piggyback derivatives) *Under Assumption 1, for almost all θ, the classical Jacobian $\frac{\partial}{\partial \theta} x_k(\theta)$, is well defined for all k and converges towards the classical Jacobian of $\bar{x}$. That is*

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

Remark 3 (Connection to implicit differentiation) The authors in [10] proved a qualification-free version of the implicit function theorem. Assuming that for every $[A, B] \in J(\bar{x}(\theta), \theta)$, the matrix $I - A$ is invertible, we have that

$$J_{\bar{x}}^{\text{imp}} \colon \theta \rightrightarrows \left\{(I - A)^{-1}B, [A, B] \in J_F(\bar{x}(\theta), \theta)\right\} \tag{9}$$

is a conservative Jacobian for $\bar{x}$. Under Assumption 1, one has $J_{\bar{x}}^{\text{imp}}(\theta) \subset J_{\bar{x}}^{\text{pb}}(\theta)$ for all θ. Unfortunately, as soon as F is not differentiable, the inclusion may be strict, see details in Appendix D.

3.3 Consequence for algorithmic differentiation

Given $k \in \mathbb{N}$, $\dot{\theta} \in \mathbb{R}^m$, $\bar{w}_k \in \mathbb{R}^p$, the following algorithms allow us to compute $\dot{x}_k = J_k \dot{\theta}$ using the forward mode of automatic differentiation (Jacobian Vector Products, JVP), and $\bar{\theta}_k^T = \bar{w}_k^T J_k$ using the backward mode of automatic differentiation (Vector Jacobian Products, VJP). In a compositional model $\dot{\theta}$ is the derivative of an inner functions controlling algorithm parameters θ, with another variable real variable $\lambda \in \mathbb{R}$, for example an hyper parameter. The goal is to combine $\frac{\partial \theta(\lambda)}{\partial \lambda}$ and $\frac{\partial x_k(\theta)}{\partial \theta}$ with the chain rule in a forward pass to obtain the total derivative $\frac{\partial x_k(\theta(\lambda))}{\partial \lambda}$. On the other hand, in a compositional model, $\bar{w}_k$ is typically the gradient of an outer loss functions ℓ evaluated at $x_k(\theta)$. In this case the goal is to combine derivatives of iterates $\frac{\partial x_k(\theta)}{\partial \theta}$ with $\bar{w}_k = \frac{\partial \ell(x_k)}{\partial x_k}$ in a backward pass to obtain $\frac{\partial \ell(x_k(\theta))}{\partial \theta}$.

Algorithm 1: Algorithmic differentiation of recursion (1), forward and reverse modes

Input: $k \in \mathbb{N}$, $\theta \in \mathbb{R}^m$, $\dot{\theta} \in \mathbb{R}^m$, $\bar{w}_k \in \mathbb{R}^p$, initialization function $x_0(\theta)$, recursion function $F(x, \theta)$, conservative Jacobians $J_F(x, \theta)$ and $J_{x_0}(\theta)$. 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 $i = 1, \ldots, k$ **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 $i = 1, \ldots, k$ **do**
  $x_i = F(x_{i-1}, \theta)$
for $i = k, \ldots, 1$ **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: $\bar{\theta}_k$

Proposition 1 (Convergence of VJP and JVP) *Let $k \in \mathbb{N}$, $\dot{\theta} \in \mathbb{R}^m$, $\bar{w}_k \in \mathbb{R}^p$, $x_k \in \mathbb{R}^p$, $\dot{x}_k \in \mathbb{R}^p$, $\bar{\theta}_k^T \in \mathbb{R}^m$ be as in Agorithm 1 under Assumption 1. Then for almost all $\theta \in \mathbb{R}^m$, $\dot{x}_k \to \frac{\partial \bar{x}}{\partial \theta}\dot{\theta}$.*

Assume furthermore that, as $k \to \infty$, $\bar{w}_k \to \bar{w}$ (for example, $\bar{w}_k = \nabla \ell(x_k)$ for a C^1 loss ℓ), then for almost all $\theta \in \mathbb{R}^m$, $\bar{\theta}_k^T \to \bar{w}^T \frac{\partial \bar{x}}{\partial \theta}$.

Remark 4 In addition to Proposition 1, in both cases, for all θ, all accumulation points of both $\dot{x}_k$ and $\bar{\theta}_k^T$ are elements of $J_{\bar{x}}^{\mathrm{pb}}\dot{\theta}$ and $\bar{w}^T J_{\bar{x}}^{\mathrm{pb}}$ respectively. This is a consequence of Corollary 2 combined with algorithmic differentiation arguments which proof is given in Appendix D.

3.4 Linear convergence rate for semialgebraic piecewise smooth selection function

Semialgebraic functions are ubiquitous in machine learning (piecewise polynomials, ℓ_1, ℓ_2 norms, determinant matrix rank ...). We refer the reader to [11] for a thorough discussion of their extensions, and use in machine learning. For more technical details, see [17, 18] for introductory material on semialgebraic and o-minimal geometry.

Lipschitz gradient selection functions. Let $F\colon \mathbb{R}^p \mapsto \mathbb{R}^q$ be semialgebraic and continuous. We say that F has a *Lipschitz gradient selection* $(s, F_1, \ldots, F_m)$ if $s\colon \mathbb{R}^p \mapsto (1, \ldots, m)$ is semialgebraic and there exists $L \geqslant 0$ such that for $i = 1 \ldots, m$, $F_i\colon \mathbb{R}^p \mapsto \mathbb{R}^p$ is semialgebraic with L-Lipschitz Jacobian, and for all $x \in \mathbb{R}^p$, $F(x) = F_{s(x)}(x)$. For any $x \in \mathbb{R}^p$, set $I(x) = \{i \in \{1, \ldots, m\}, F(x) = F_i(x)\}$. The set-valued map $J_F^s\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^{p \times q}$ given by $J_F^s\colon x \rightrightarrows \mathrm{conv}\left(\left\{\frac{\partial F_i}{\partial x}(x),\, i \in I(x)\right\}\right)$, is a conservative Jacobian for F as shown in [11]. Here $\frac{\partial F_i}{\partial x}$ denotes the classical Jacobian of F_i. Let us stress that such a structure is ubiquitous in applications [11, 34].

Rate of convergence. We may now strengthen Corollary 1 by proving the linear convergence of piggyback derivatives towards the fixed point. The following is a consequence of the fact that the proposed selection conservative Jacobians of Lipschitz gradient selection functions are Lipschitz-like (Lemma 4 in Appendix E.1). Note that semialgebraicity is only used as a *sufficient* condition to ensure conservativity of the selection Jacobian together with this Lipschitz like property. It could be relaxed if it can be guaranteed by other means, in particular one could consider the broader class of definable functions in order to handle log-likelihood data fitting terms.

Corollary 3 (Linear convergence of piggyback derivatives) *In addition to Assumption 1, assume that F has a Lipschitz gradient selection structure as above. Then, for any θ and $\epsilon > 0$, there exists $C > 0$ such that the recursion (PB) with $J_F = J_F^s$ satisfies for all $k \in \mathbb{N}$, $\mathrm{gap}(J_{x_k}(\theta), J_{\bar{x}}^{\mathrm{pb}}(\theta)) \leqslant C(\sqrt{\rho} + \epsilon)^k$. Moreover, classical Jacobians in Corollary 2 converge at a linear rate for almost all θ.*

4 Application to proximal splitting methods in convex optimization

Consider the composite parametric convex optimization problem, where $\theta \in \mathbb{R}^m$ represents parameters and $x \in \mathbb{R}^p$ is the decision variable

$$\bar{x}(\theta) = \arg\min_x f(x, \theta) + g(x, \theta).$$

The purpose of this section is to construct examples of functions F used in recursion (1) based on known algorithms. The following assumption will be standing throughout the section.

Assumption 2 f is semialgebraic, convex, its gradient with respect to x for fixed θ, $\nabla_x f$, is locally Lipschitz jointly in (x, θ) and L-Lipschitz in x for fixed θ. Semialgebraicity implies that $\nabla_x f$ is path-differentiable jointly in (x, θ), we denote by J_f^2 its Clarke Jacobian. The function g is semialgebraic, convex in x for fixed θ, and lower semicontinuous. For all $\alpha > 0$, we assume that $G_\alpha(x, \theta) \mapsto \mathrm{prox}_{\alpha g(\cdot,\theta)}(x)$ is locally Lipschitz jointly in (x, θ). Semialgebraicity implies that it is also path differentiable jointly in (x, θ), we denote by J_{G_α} its Clarke Jacobian.

This assumption covers a very large diversity of problems in convex optimization as most gradient and prox operations used in practice are semialgebraic (or definable). Under Assumption 2, we will provide sufficient conditions on f and g for Assumption 1, for different algorithmic recursions. These will therefore imply convergence as stated in Corollary 1 and 2, Proposition 1, as well Corollary 3 in the piecewise selection case. The proofs are postponed to Appendix F.

4.1 Splitting algorithms

In this section we provide sufficient condition for Assumption 1 to hold. The underlying conservative Jacobian is obtained by combining Clarke Jacobians of elementary algorithmic operations (gradient, proximal operator in Assumption 2), using the compositional rules of differential calculus [11] and implicit differentiation [10]. Using [12], such Jacobians are conservative by semialgebraicity and their combination provide conservative Jacobians for the corresponding algorithmic recursion F. These objects are explicitly constructed in Appendix F.

Forward–backward algorithm. The forward–backward iterations are given for $\alpha > 0$ by

$$x_{k+1} = \mathrm{prox}_{\alpha g(\cdot,\theta)}\left(x_k - \alpha \nabla_x f(x_k, \theta)\right). \tag{10}$$

Proposition 2 *Under Assumption 2 with $0 < \alpha < \frac{2}{L}$, denote by $F_\alpha \colon \mathbb{R}^{p \times m} \to \mathbb{R}^p$ the forward-backward recursion in (10). For $\mu > 0$, if either f or g is μ-strongly convex in x for all θ, then F_α is a strict contraction and Assumption 1 holds.*

It is well known that if f is μ-strongly convex, choosing $\alpha = 2/(L + \mu)$ provides a contraction factor $\rho_{FB} = (\tau - 1)/(1 + \tau)$, where $\tau = L/\mu \geqslant 1$ is the condition number of the problem.

Douglas–Rachford algorithm. Given $\alpha > 0$, the algorithm goes as follows

$$y_{k+1} = \frac{1}{2}(I + R_{\alpha f(\cdot,\theta)} R_{\alpha g(\cdot,\theta)}) y_k, \tag{11}$$

where $R_{\alpha f(\cdot,\theta)} = 2\mathrm{prox}_{\alpha f(\cdot,\theta)} - I$ is the reflected proximal operator, which is 1-Lipschitz (and similarly for g). Following [6, Theorem 26.11], if the problem has a minimizer, then $(y_k)_{k \in \mathbb{N}}$ converges to a fixed point of (11), $\bar{y}$ such that $\bar{x} = \mathrm{prox}_{\alpha g}(\bar{y})$ is a solution to the optimization problem. Following [26, Theorem 1], if f is strongly convex, then $R_{\alpha f(\cdot,\theta)}$ is ρ-Lipschitz for some $\rho < 1$ and our differentiation result applies to Douglas-Rachford splitting in this setting.

Proposition 3 *Under Assumption 2 with $\alpha > 0$, denote by $F_\alpha \colon \mathbb{R}^{p \times m} \to \mathbb{R}^p$ the Douglas-Rachford recursion in (11). If f is μ-strongly convex in x for all θ, then F_α is a strict contraction and Assumption 1 holds.*

Following [26, Proposition 3], choosing $\alpha = 1/\sqrt{L\mu}$ provides a contraction factor of order $\rho_{DR}(\sqrt{\tau} - 1)/(\sqrt{\tau} + 1) < \rho_{FB}$, where again $\tau = L/\mu$ is the condition number of the problem. In this respect Douglas-Rachford's iterations provide a faster asymptotic rate than those of Forward-Backward, which may also impact the convergence of derivatives in the context of Corollary 3.

Alternating Direction Method of Multipliers. Consider the separable convex problem

$$\min_{u,v} \phi_\theta(u) + \psi_\theta(v) \quad \text{subject to} \quad A_\theta u + B_\theta v = c_\theta. \tag{12}$$

The alternating direction method of multipliers (ADMM) algorithm combines two partial minimization of an augmented Lagrangian, and a dual update:

$$\begin{aligned} u_{k+1} &= \arg\min_u \left\{ \phi_\theta(u) + x_k^\top A_\theta u + \frac{\alpha}{2}\|A_\theta u + B_\theta v_k - c_\theta\|_2^2 \right\} \\ v_{k+1} &= \arg\min_v \left\{ \psi_\theta(v) + x_k^\top B_\theta v + \frac{\alpha}{2}\|A_\theta u_{k+1} + B_\theta v_k - c_\theta\|_2^2 \right\} \\ x_{k+1} &= x_k + \alpha(A_\theta u_{k+1} + B_\theta v_{k+1} - c_\theta). \end{aligned} \tag{13}$$

As observed in [23], the ADMM algorithm can be seen as the Douglas-Rachford splitting method applied to the Fenchel dual of problem (12) (see Appendix F.3 for more details). More precisely, ADMM updates are equivalent to Douglas-Rachford iterations applied to the following problem

$$\min_x \underbrace{c_\theta^\top x + \phi_\theta^*(-A_\theta^\top x)}_{f(x,\theta)} + \underbrace{\psi_\theta^*(-B_\theta^\top x)}_{g(x,\theta)}. \tag{14}$$

Therefore, if ϕ_θ is strongly convex with Lipschitz gradient and A_θ is injective, then ADMM converges linearly and one is able to combine derivatives of proximal operators to differentiate ADMM.

4.2 Numerical illustrations

We now detail how Figure 2 discussed in the introduction is obtained, and how it illustrates our theoretical results. We consider four scenarios (Ridge, Lasso, Sparse inverse covariance selection and Trend filtering) corresponding to the four columns. For each of them, the first line shows the empirical linear rate of the iterates x_k and the second line shows the empirical linear rate of the derivative $\frac{\partial}{\partial\theta} x_k$. All experiments are repeated 100 times and we report the median along with the first and last deciles.

Forward–Backward for the Ridge. The Ridge estimator is defined for $\theta > 0$ as $\bar{x}(\theta) = \arg\min_{x\in\mathbb{R}^p} \frac{1}{2}\|Ax - b\|_2^2 + \theta\|x\|_2^2$ Among several possibilities to solve it, one can use the Forward–Backward algorithm applied to $f\colon (x,\theta) \mapsto \frac{1}{2}\|Ax - b\|_2^2$ and $g : \theta\|x\|_2^2$. Since g is strongly convex, the operator F_α is strongly convex, and thus Proposition 2 may be applied.

Forward–Backward algorithm for the Lasso. Consider the Forward–Backward algorithm applied to the Lasso problem [49], with parameter $\theta > 0$, $\bar{x}(\theta) \in \arg\min_{x\in\mathbb{R}^p} \frac{1}{2}\|Ax - b\|_2^2 + \theta\|x\|_1 = \arg\min_x \frac{1}{2L}\|Ax - b\|_2^2 + \frac{\theta}{L}\|x\|_1$, where L is any upper bound on the operator norm of $A^T A$. The gradient of the quadratic part is 1-Lipschitz, so we may consider the forward backward algorithm (10), with unit step size and $f\colon (x,\theta) \mapsto \frac{1}{2L}\|Ax - b\|_2^2$ and $g\colon (x,\theta) \mapsto \frac{\theta}{L}\|x\|_1$.

A well known qualification condition involving a generalized support at optimality ensures uniqueness of the Lasso solution [20, 37]. It holds for generic problem data [50]. Following [10, Proposition 5], under this qualification condition, the implicit conservative Jacobian J_F is such that, at the solution x^*, the matrix set $I - J_F$ only contains invertible matrices. This means that there exists $\rho < 1$, such that any $M \in J_F(x^*)$ has operator norm at most ρ. Following Remark 1, all our convergence results apply qualitatively. Note that we recover the results of [7, Proposition 2] for the Lasso.

Douglas–Rachford for the Sparse Inverse Covariance Selection. The Sparse Inverse Covariance Selection [52, 22] reads $\bar{x}(\theta) \in \arg\min_{x\in\mathbb{R}^{n\times n}} \operatorname{tr}(Cx) - \log\det x + \theta\sum_{i,j}|x_{i,j}|$, where C is a symmetric positive matrix and $\theta > 0$. It is possible to apply Douglas–Rachford method to $f : (x,\theta) \mapsto \operatorname{tr}(Cx) - \log\det x$ and $g : (x,\theta) \mapsto \theta\|x\|_{1,1}$. It is known that f is locally strongly convex, indeed $x \mapsto -\log\det x$ is the standard self-concordant barrier in semidefinite programming [40]. Following Remark 1, all our convergence results apply qualitatively.

ADMM for Trend Filtering. Introduced in [51] in statistics as a generalization of the Total Variation, the trend filtering estimator with observation $\theta \in \mathbb{R}^p$ reads $\bar{x}(\theta) = \arg\min_{x\in\mathbb{R}^p} \frac{1}{2}\|x - \theta\|_2^2 + \lambda\|D^{(k)}x\|_1$, where $D^{(k)}$ is a forward finite–difference approximation of a differential operator of order k (here $k = 2$). Using $\psi_\theta : u \mapsto \lambda\|u\|_1$, $\phi_\theta : v \mapsto \|v - \theta\|_2^2$ (strongly convex), $A_\theta = -I$ (injective), $B_\theta = D^{(k)}$, and $c_\theta = 0$, we can apply the ADMM to solve trend filtering.

5 Failure of automatic differentiation for inertial methods

This section focuses on the Heavy-Ball method for strongly convex objectives, in its global linear convergence regime. For C^2 objectives, piggyback derivatives converge to the derivative of the solution map [28, 39, 36]. However, we provide a $C^{1,1}$ strongly convex parametric objective with path differentiable derivative, such that piggyback derivatives of the Heavy Ball algorithm contain diverging vectors for a given parameter value. In this example, other conservative differentiation means (implicit differentiation, piggyback on gradient descent), avoid this divergent behaviors.

5.1 Heavy-ball algorithm and global convergence

Consider a function $f\colon \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$, and $\beta > 0$, for simplicity, when the second argument is fixed, we write $f_\theta\colon x \mapsto f(x,\theta)$. Set for all x, y, θ, $F(x,y,\theta) = (x - \nabla f_\theta(x) + \beta(x-y), x)$, consider the Heavy-Ball algorithm $(x_{k+1}, y_{k+1}) = F(x_k, y_k, \theta)$ for $k \in \mathbb{N}$.

If f_θ is μ-strongly convex with L-Lipschitz gradient, then, choosing $\alpha = 1/L$ and $\beta < \frac{1}{2}\left(\frac{\mu}{2L} + \sqrt{\frac{\mu^2}{4L^2} + 2}\right)$, the algorithm will converge globally at a linear rate to the unique solution, $\bar{x}(\theta)$ [24, Theorem 4], local convergence is due to Polyak [44]. Furthermore, if in addition f is C^2 forward propagation of derivatives converge to the derivative of the solution [28, 29, 39].

5.2 A diverging Jacobian accumulation

Details and proof of the following result are given in Section G.

Proposition 4 (Piggyback differentiation fails for the Heavy Ball method) *Consider $f\colon \mathbb{R}^2 \to \mathbb{R}$, such that for all $\theta \in \mathbb{R}$, $f(x,\theta) = x^2/2$ if $x \geqslant 0$ and $f(x,\theta) = x^2/8$ if $x < 0$. Assume that $\alpha = 1$ and $\beta = 3/4$. Then the heavy ball algorithm converges globally to 0 and ∇f is path differentiable. The Clarke Jacobian of F with respect to (x,y) at $(0,0,0)$ is $J_F(0,0,0) = \mathrm{conv}\,\{M_1, M_2\}$, where the product $M_1 M_1 M_2 M_2$ has eigenvalue $-9/8$.*

The presence of an eigenvalue with modulus greater than 1 may produce divergence in (PB). Set

$$f_1\colon (x,\theta) \mapsto \begin{cases} x^2/2 & \text{if } x \geqslant 0 \\ x^2/8 & \text{if } x < 0. \end{cases} \qquad f_2\colon (x,\theta) \mapsto \begin{cases} x^2/2 & \text{if } x > 0 \\ x^2/8 & \text{if } x \leqslant 0. \end{cases}$$

Note that f_1 and f_2 are both equivalent to f as they implement the same function. With initializations $x(\theta) = y(\theta) = \theta$, we run a few iterations of the Heavy Ball algorithm for $\theta = 0$, and implement (PB) alternating between two steps on f_1 and two steps on f_2 and differentiate the resulting sequence $(x_k)_{k\in\mathbb{N}}$ with respect to θ using algorithmic differentiation. The divergence phenomenon predicted by Proposition 4 is illustrated in Figure 3, while the true derivative is 0 (the sequence is constant).

6 Conclusion

We have developed a flexible theoretical framework to describe convergence of piggyback differentiation applied to nonsmooth recursions – providing, in particular, a rigorous meaning to differentiation of nonsmooth solvers. The relevance of our approach is illustrated on composite convex optimization

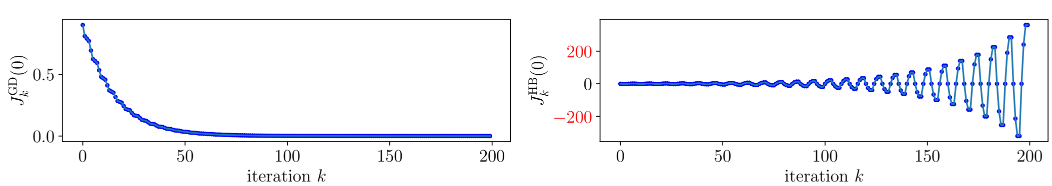

Figure 3: Behaviour of automatic differentiation for first-order methods on a quadratic function. (Left) Stability of the propagation of derivatives for the fixed step-size gradient descent. (Right) Instability of the propagation of Heavy-Ball initialized. Both methods are initialized at optimum.

through widely used methods as forward-backward, Douglas-Rachford or ADMM algorithms. Our framework allows however to consider many other abstract algorithmic recursions and provides thus theoretical ground for more general problems such as variational inequalities or saddle point problems as in [14, 9]. As a matter for future work, we shall consider relaxing Assumption 1 to study a wider class of methods, e.g., when F is not a strict contraction.

Acknowledgments

J. B. and E. P. acknowledge the financial support of the AI Interdisciplinary Institute ANITI funding under the grant agreement ANR-19-PI3A-0004, Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant numbers FA9550-19-1-7026, FA8655-22-1-7012 and ANR MaSDOL 19-CE23-0017-01. J. B. also acknowledges the support of ANR Chess, grant ANR-17-EURE-0010, TSE-P and the Centre Lagrange. S. V. acknowledges the support ANR GraVa, grant ANR-18-CE40-0005.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Pierre Ablin, Gabriel Peyré, and Thomas Moreau. Super-efficiency of automatic differentiation for functions defined as a minimum. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 32–41. PMLR, 13–18 Jul 2020.

[3] Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J. Zico Kolter. Differentiable convex optimization layers. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[4] Charalambos D Aliprantis and Kim C Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Science & Business Media, 2006.

[5] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[7] Quentin Bertrand, Quentin Klopfenstein, Mathieu Blondel, Samuel Vaiter, Alexandre Gramfort, and Joseph Salmon. Implicit differentiation of lasso-type models for hyperparameter optimization. In *International Conference on Machine Learning*, pages 810–821. PMLR, 2020.

[8] Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation, 2021.

[9] Lea Bogensperger, Antonin Chambolle, and Thomas Pock. Convergence of a Piggyback-style method for the differentiation of solutions of standard saddle-point problems. Technical report, 2022.

[10] Jérôme Bolte, Tam Le, Edouard Pauwels, and Antonio Silveti-Falls. Nonsmooth Implicit Differentiation for Machine Learning and Optimization. In *Advances in Neural Information Processing Systems*, Online, 2021.

[11] Jerome Bolte and Edouard Pauwels. A mathematical model for automatic differentiation in machine learning. In *Conference on Neural Information Processing Systems*, Vancouver, Canada, 2020.

[12] Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient method and deep learning. *Mathematical Programming*, 2020.

[13] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

[14] Antonin Chambolle and Thomas Pock. Learning consistent discretizations of the total variation. *SIAM Journal on Imaging Sciences*, 14(2):778–813, 2021.

[15] Bruce Christianson. Reverse accumulation and attractive fixed points. *Optimization Methods and Software*, 3(4):311–326, 1994.

[16] Frank H. Clarke. *Optimization and nonsmooth analysis*. Canadian Mathematical Society Series of Monographs and Advanced Texts. John Wiley & Sons, Inc., New York, 1983. A Wiley-Interscience Publication.

[17] Michel Coste. *An introduction to o-minimal geometry*. Istituti editoriali e poligrafici internazionali Pisa, 2000.

[18] Michel Coste. An introduction to semialgebraic geometry, 2000.

[19] Charles-Alban Deledalle, Samuel Vaiter, Jalal M. Fadili, and Gabriel Peyré. Stein Unbiased GrAdient estimator of the Risk (SUGAR) for multiple parameter selection. *SIAM Journal on Imaging Sciences*, 7(4):2448–2487, 2014.

[20] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *Annals of statistics*, 32(2):407–499, 2004.

[21] Laurent El Ghaoui, Fangda Gu, Bertrand Travacca, Armin Askari, and Alicia Tsai. Implicit deep learning. *SIAM Journal on Mathematics of Data Science*, 3(3):930–958, 2021.

[22] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 12 2007.

[23] Daniel Gabay. Applications of the method of multipliers to variational inequalities. In *Studies in mathematics and its applications*, volume 15, pages 299–331. Elsevier, 1983.

[24] Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE, 2015.

[25] Jean Charles Gilbert. Automatic differentiation and iterative processes. *Optimization Methods and Software*, 1(1):13–21, 1992.

[26] Pontus Giselsson and Stephen Boyd. Linear convergence and metric selection for douglas-rachford splitting and admm. *IEEE Transactions on Automatic Control*, 62(2):532–544, 2016.

[27] Andreas Griewank, Christian Bischof, George Corliss, Alan Carle, and Karen Williamson. Derivative convergence for iterative equation solvers. *Optimization Methods and Software*, 2(3-4):321–355, 1993.

[28] Andreas Griewank and Christèle Faure. Piggyback differentiation and optimization. In *Large-scale PDE-constrained optimization*, pages 148–164. Springer, 2003.

[29] Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008.

[30] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[31] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[32] Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Gradient step denoiser for convergent plug-and-play. In *International Conference on Learning Representations*, 2021.

[33] Sham M Kakade and Jason D Lee. Provably correct automatic sub-differentiation for qualified programs. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[34] Wonyeol Lee, Hangyeol Yu, Xavier Rival, and Hongseok Yang. On correctness of automatic differentiation for non-differentiable functions. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6719–6730. Curran Associates, Inc., 2020.

[35] Seppo Linnainmaa. The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. Master's thesis, Univ. Helsinki, 1970.

[36] Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1540–1552. PMLR, 26–28 Aug 2020.

[37] Julien Mairal and Bin Yu. Complexity analysis of the lasso regularization path. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1835–1842, 2012.

[38] Swann Marx and Edouard Pauwels. Path differentiability of ode flows. *arXiv preprint arXiv:2201.03819*, 2022.

[39] Sheheryar Mehmood and Peter Ochs. Automatic differentiation of some first-order methods in parametric optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1584–1594. PMLR, 2020.

[40] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

[41] Peter Ochs, René Ranftl, Thomas Brox, and Thomas Pock. Techniques for gradient-based bilevel optimization with non-smooth lower level problems. *Journal of Mathematical Imaging and Vision*, 56(2):175–194, 2016.

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[43] Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International conference on machine learning*, pages 737–746. PMLR, 2016.

[44] Boris Polyak. Introduction to optimization. In *Optimization Software, Publications Division*. Citeseer, 1987.

[45] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

[46] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[47] Halsey Lawrence Royden and Patrick Fitzpatrick. *Real analysis*, volume 32. Macmillan New York, 1988.

[48] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986.

[49] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[50] Ryan J Tibshirani. The lasso problem and uniqueness. *Electronic Journal of statistics*, 7:1456–1490, 2013.

[51] Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285 – 323, 2014.

[52] Martin J Wainwright, John Lafferty, and Pradeep Ravikumar. High-dimensional graphical model selection using \ell_1-regularized logistic regression. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

[53] R. E. Wengert. A simple automatic derivative evaluation program. *Communications of the ACM*, 7(8):463–464, August 1964.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Our submission follows the order mentionned in the abstract and introduction.
   (b) Did you describe the limitations of your work? [Yes] The two main assumptions are discussed in the main text.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We believe that our work, being fully theoretical, does not bring negative societal impacts, especially points 1 to 7 of the ethics guidelines of NeurIPS 2022.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] Points 1 to 5 are satisfied.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Our analysis requires a small number of assumptions: assumption 1 on contraction and assumption 2 on semialgebraic regularity.
   (b) Did you include complete proofs of all theoretical results? [Yes] Most of the appendices are dedicated to them.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The (self-contained) jupyter notebook to generate the figures is submitted in supllemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See appendix H.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See appendix H. Every curves report the median over 100 repetitions, along with first and last deciles.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See appendix H.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] See appendix I.
(b) Did you mention the license of the assets? [Yes] See appendix I.
(c) Did you include any new assets either in the supplemental material or as a URL? [N/A] No new assets.
(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] No nonpublic data used.
(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] No personally identifiable information.

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

This is the appendix for "Convergence of piggyback differentiation of nonsmooth iterative solvers".

Appendices

A Reminder on conservative calculus

For the sake of completeness, we recall important definitions and results from [12] on conservative calculus which are extensively used throughout the paper.

Definitions: We first collect the necessary definitions and details for Equation (4). We then collect important results from [12], which will be used throughout the paper. Recall from multivariable calculus that the *Jacobian* of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$ is given by

$$\frac{\partial f}{\partial x} := \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

Definition 1 (Absolutely continuous curves) *A continuous function $\gamma : \mathbb{R} \to \mathbb{R}^n$ is an absolutely continuous curve if it has a derivative $\dot{\gamma}(t)$, for almost all $t \in \mathbb{R}$, which furthermore satisfies*

$$\gamma(t) - \gamma(0) = \int_0^t \dot{\gamma}(\tau) d\tau$$

for all $t \in \mathbb{R}$.

The *graph* of a set-valued mapping $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is the set graph $D := \{(x, z) : x \in \mathbb{R}^n, z \in D(x)\}$.

Definition 2 (Closed graphs) *A set-valued mapping $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ has closed graph or is graph closed if* graph D *is a closed subset of $\mathbb{R}^{n+m}$ or, equivalently, if, for any convergent sequences $(x_k)_{k\in\mathbb{N}}$ and $(z_k)_{k\in\mathbb{N}}$ with $z_k \in D(x_k)$ for all $k \in \mathbb{N}$, it holds*

$$\lim_{k\to\infty} z_k \in D(\lim_{k\to\infty} x_k).$$

Definition 3 (Locally bounded set-valued mappings) *A set-valued mapping $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is locally bounded if for all $x \in \mathbb{R}^n$, there exists a neighborhood $\mathcal{U}$ of x and $M > 0$ such that, for all $u \in \mathcal{U}$, for all $y \in D(u)$, $\|y\| < M$.*

We provide an equivalent alternative to Definition Equation (4) see [12, Lemma 2].

Definition 4 (Conservative Jacobians) *The set-valued $J : \mathbb{R}^p \rightrightarrows \mathbb{R}^{m\times p}$ is a* conservative Jacobian *if J has a closed graph, is locally bounded and nowhere empty with*

$$\int_{t=0}^{t=1} J(\gamma(t))\dot{\gamma}(t)dt = 0 \tag{15}$$

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

Given such a J, the potential f as in Equation (4) can be reconstructed up to a constant using integration along absolutely continuous through.

$$f(\gamma(1)) - f(\gamma(0)) = \int_{t=0}^{t=1} J(\gamma(t))\dot{\gamma}(t)dt, \tag{16}$$

where the value of the integral does not depend on the choice of γ provided that the endpoints are fixed.

First results and examples : We have the following results, see [12, Theorem 1, Corollary 2].

Theorem 3 *Let $F\colon \mathbb{R}^p \to \mathbb{R}^m$ be locally Lipschitz. Then F is path differentiable if and only if* $\mathrm{Jac}^c F$ *in (3) is a conservative Jacobian. In this case, setting $\mathcal{J}\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^{m\times p}$ any conservative Jacobian for F, we have*

- $\mathcal{J}(x) = \{\mathrm{Jac}\, F(x)\}$ *for Lebesgue almost all x.*
- $\mathrm{Jac}^c(x) \subset \mathrm{conv}(\mathcal{J}(x))$ *for all x.*

Example of path differentiable functions include

- Convex or concave functions
- Clarke regular functions
- Prox regular functions

we refer to [46] for details on these classes of functions. Another relevant class is that of semi-algebraic or more generally definable functions, see [17, 18]. Beyond technical definition, this class is relevant because it contains the vast majority of operations used in applications, independently of smoothness. These include: the relu function, the absolute value function, the max-pooling operation, ℓ_1 and ℓ_∞ norms, any polynomial or piecewise polynomial function such sorting a vector by increasing coordinates order, the operator norm, the rank function .... Furthermore, the class of semi-algebraic functions is closed under many operations, as for instance:

- usual arithmetic operations $+, \times, -, /$
- functional composition
- differentiation
- partial minimization
- more broadly, any functional operation which can be described with a first order logical formula: a boolean formula with quantification on variables only (not sets), see [17].

Conservative Jacobians and calculus: The main reason for the introduction of conservative Jacobians in [12] is the lack of an efficient differential calculus for Clarke Jacobians (recall (3)). For example, if $f = |\cdot|$ and $g = -|\cdot|$, we have

$$\partial^c(f+g)(0) = \partial^c(t \mapsto 0)(0) = \{0\} \neq [-2,2] = \partial^c f(0) + \partial^c g(0).$$

On the contrary, conservative Jacobians have an appealing calculus.

Lemma 1 [12, Lemma 5] *Let $F_1 : \mathbb{R}^p \mapsto \mathbb{R}^m$ and $F_2 : \mathbb{R}^m \mapsto \mathbb{R}^l$ be locally Lipschitz continuous mappings. Let $J_1 : \mathbb{R}^p \rightrightarrows \mathbb{R}^{m\times p}$ be a conservative mapping for F_1 and $J_2 : \mathbb{R}^p \rightrightarrows \mathbb{R}^{l\times m}$ be a conservative mapping for F_2. Then the product mapping $J_2 \cdot J_1$ is a conservative mapping for $F_2 \circ F_1$.*

as a consequence, beyond composition, conservative gradients are compatible with basic arithmetic operations, such as addition. In general conservative gradients and Jacobian provide a variational meaning to the formal application of the rules of differential calculus to generalized derivatives arising in nonsmooth analysis, this goes beyond simple arithmetic operations and composition, for example with implicit differentiation [10].

Optimization: Let D be a conservative gradient, v is called a selection in D if for all x, $v(x) \in D(x)$. Selection conservative gradients can be used as surrogate gradients, or subgradients, while preserving convergence guaranties, examples are given in [12, 11, 10].

B Properties of affine iterations on compact subsets

B.1 Banach–Picard theorem: Proof of Theorem 1

For a compact set, $\mathcal{Z}$ we denote by $\|\mathcal{Z}\|_{\sup}$ the maximal norm of elements in $\mathcal{Z}$:

$$\|\mathcal{Z}\|_{\sup} = \sup_{z\in\mathcal{Z}} \|z\|.$$

In order to prove our fixed point result, we need first the following lemma.

Lemma 2 (Bounding Hausdorff distances) *Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z} \subset \mathbb{R}^p$ be nonempty compact sets, such that $\mathcal{X} \subset \mathcal{Y} + \mathcal{Z}$ and $\mathcal{Y} \subset \mathcal{X} + \mathcal{Z}$ then*

$$\mathrm{dist}(\mathcal{X}, \mathcal{Y}) \leqslant \|\mathcal{Z}\|_{\sup}.$$

Proof : The first inclusion says that for any $x \in \mathcal{X}$, there is $y(x) \in \mathcal{Y}$, $z(x) \in \mathcal{Z}$ such that $x = y(x) + z(x)$. We deduce that for any $x \in \mathcal{X}$

$$\min_{y\in\mathcal{Y}} \|x - y\| = \min_{y\in\mathcal{Y}} \|y(x) - z(x) - y\| \leqslant \|z(x)\| \leqslant \max_{z\in\mathcal{Z}} \|z\|$$

Therefore, $\max_{x\in\mathcal{X}} \min_{y\in\mathcal{Y}} \|x - y\| \leqslant \max_{z\in\mathcal{Z}} \|z\|$. Symmetrically, $\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} \|x - y\| \leqslant \max_{z\in\mathcal{Z}} \|z\|$ and the result follows. □

We now prove Theorem 1.

Proof of Theorem 1: Recall that the action of $\mathcal{J}$ on matrices is defined in (6) and by $\mathcal{A}$ and $\mathcal{B}$ the projections of $\mathcal{J}$ on the first p and last l columns respectively, that is $\mathcal{A} = \{A, \exists B, [A, B] \in \mathcal{J}\}$ and similarly for B. Note that $\mathcal{A}$ is a compact set and that all matrices in $\mathcal{A}$ have an operator norm of at most ρ. We claim that the restriction of $\mathcal{J}$ to compact subsets is a strict contraction in Hausdorff metric. Indeed, for any $\mathcal{X}, \mathcal{Y}$ compact subsets of $\mathbb{R}^{p\times m}$, we have by using Lemma 2 and noting that $\mathcal{J}$ preserves the inclusion,

$$\mathcal{J}(\mathcal{X}) \subset \mathcal{J}(\mathcal{Y} + \mathrm{dist}(\mathcal{X}, \mathcal{Y})\mathbb{B}) \subset \mathcal{J}(\mathcal{Y}) + \mathrm{dist}(\mathcal{X}, \mathcal{Y})\mathcal{A}\mathbb{B} \subset \mathcal{J}(\mathcal{Y}) + \rho\mathrm{dist}(\mathcal{X}, \mathcal{Y})\mathbb{B}$$
$$\mathcal{J}(\mathcal{Y}) \subset \mathcal{J}(\mathcal{X} + \mathrm{dist}(\mathcal{X}, \mathcal{Y})\mathbb{B}) \subset \mathcal{J}(\mathcal{X}) + \mathrm{dist}(\mathcal{X}, \mathcal{Y})\mathcal{A}\mathbb{B} \subset \mathcal{J}(\mathcal{X}) + \rho\mathrm{dist}(\mathcal{X}, \mathcal{Y})\mathbb{B}$$

where the last inclusion follows because $\mathcal{A}\mathbb{B} \subset \rho\mathbb{B}$, where $\mathbb{B}$ is the unit ball (for the Euclidean norm) of $p \times m$ matrices, since by assumption all square matrices in $\mathcal{A}$ have operator norm at most ρ. We deduce that $\mathrm{dist}(\mathcal{J}(\mathcal{X}), \mathcal{J}(\mathcal{Y})) \leqslant \rho\mathrm{dist}(\mathcal{X}, \mathcal{Y})$ using Lemma 2, that is the action of $\mathcal{J}$ on subsets of $p \times m$ matrices is ρ Lipschitz with respect to Hausdorff metric.

Let us show that $(\mathcal{X}_k)_{k\in\mathbb{N}}$ remains in a bounded set, we have for all k

$$\|\mathcal{X}_{k+1}\|_{\sup} \leqslant \|\mathcal{A}\mathcal{X}_k + \mathcal{B}\|_{\sup} \leqslant \|\mathcal{A}\mathcal{X}_k\|_{\sup} + \|\mathcal{B}\|_{\sup} \leqslant \rho\|\mathcal{X}_k\|_{\sup} + \|\mathcal{B}\|_{\sup},$$

which entails

$$\|\mathcal{X}_{k+1}\|_{\sup} - \frac{\|\mathcal{B}\|_{\sup}}{1-\rho} \leqslant \rho\left(\|\mathcal{X}_k\|_{\sup} - \frac{\|\mathcal{B}\|_{\sup}}{1-\rho}\right).$$

We distinguish two cases

- if $\|\mathcal{X}_k\|_{\sup} > \frac{\|\mathcal{B}\|_{\sup}}{1-\rho}$, then $\|\mathcal{X}_{k+1}\|_{\sup}$ gets either closer to $\frac{\|\mathcal{B}\|_{\sup}}{1-\rho}$ or below it, in particular it decreases.
- if $\|\mathcal{X}_k\|_{\sup} \leqslant \frac{\|\mathcal{B}\|_{\sup}}{1-\rho}$ then $\|\mathcal{X}_{k+1}\|_{\sup} \leqslant \frac{\|\mathcal{B}\|_{\sup}}{1-\rho}$ and we remain below this threshold for all k.

All in all, for all $k \in \mathbb{N}$,

$$\|\mathcal{X}_{k+1}\|_{\sup} \leqslant \max\left\{\|\mathcal{X}_k\|_{\sup}, \frac{\|\mathcal{B}\|_{\sup}}{1-\rho}\right\} \leqslant \ldots \leqslant \max\left\{\|\mathcal{X}_0\|_{\sup}, \frac{\|\mathcal{B}\|_{\sup}}{1-\rho}\right\},$$

and $\limsup_k \|\mathcal{X}_k\|_{\sup} \leqslant \frac{\|\mathcal{B}\|_{\sup}}{1-\rho}$.

We have shown that the sequence remains in a bounded set so that the recursion actually takes place in a compact set $\mathcal{C} \subset \mathbb{R}^{p\times m}$ which contains all the iterates in its interior, we consider the restriction of the topology to this subset. By [4, Theorem 3.85], the closed subsets form a complete metric space. The result is an application of Banach-Picard theorem (for example [47, Section 10.3]). In particular (see [4, Theorem 3.88]), $\mathcal{L}$ is the unique fixed point of $\mathcal{J}$ and it is closed and bounded, hence compact. Note that we can consider larger compact sets to take into account larger initializations, the fixed point remains the same. Indeed for a larger compact $\tilde{\mathcal{C}}$ containing $\mathcal{C}$, $\mathcal{L}$ is in the interior of $\mathcal{C}$ and is still a fixed point of $\mathcal{J}$ when the topology is restricted to $\tilde{\mathcal{C}}$ and this fixed point must be unique. □

B.2 Properties of the fixed-set mapping

We now equip the set of matrices $\mathbb{R}^{p\times(p\times m)}$ with the norm $\|[A,B]\|_{p,m} = \max\{\|A\|_{\mathrm{op}}, \|B\|\}$ where $A \in \mathbb{R}^{p\times p}$ and $B \in \mathbb{R}^{p\times m}$. The set of compact subsets of $\mathbb{R}^{p\times(p+m)}$ is endowed with the corresponding Hausdorff distance.

Definition 5 (Affine contraction sets) *For $\rho \in [0,1)$, we denote by $\mathcal{C}_\rho$, the set of compact subsets of matrices in $\mathbb{R}^{p\times(p+m)}$ such that for all $\mathcal{S} \subset \mathbb{R}^{p\times(p+m)}$, $\mathcal{S} \subset \mathcal{C}_\rho$ and all $M \in \mathcal{S}$, we have $\|A\|_{\mathrm{op}} \leqslant \rho$ where $A \in \mathbb{R}^{p\times p}$ is the matrix made of the first p columns of M.*

Given $\mathcal{J} \in \mathcal{C}_\rho$, we denote by $\mathrm{fix}(\mathcal{J})$ the unique fixed point of the corresponding iteration mapping as defined in Theorem 1. We have the following

Proposition 5 (Monotonicity of the fixed set) *Given $\mathcal{J} \in \mathcal{C}_\rho$ and $\tilde{\mathcal{J}} \in \mathcal{C}_\rho$ (as in Definition 5), such that $\mathcal{J} \subset \tilde{\mathcal{J}}$, we have*

$$\mathrm{fix}(\mathcal{J}) \subset \mathrm{fix}(\tilde{\mathcal{J}}).$$

Proof : Setting $\mathcal{X}_0 = \mathrm{fix}(\mathcal{J})$, we have

$$\mathcal{X}_0 = \mathcal{J}(\mathcal{X}_0) \subset \tilde{\mathcal{J}}(\mathcal{X}_0),$$

and the result follows by the same argument as in the last paragraph of the proof of Theorem 1. □

Proposition 6 (The fixed-set mapping is locally Lipschitz continuous) *The function* fix *is locally Lipschitz continuous on $\mathcal{C}_\rho$ (as in Definition 5). More precisely, for any $\mathcal{J}_0 \in \mathcal{C}_\rho$ and $\mathcal{J} \in \mathcal{C}_\rho$,*

$$\mathrm{dist}\left(\mathrm{fix}(\mathcal{J}_0), \mathrm{fix}(\mathcal{J})\right) \leqslant \left(\frac{1}{1-\rho} + \frac{\sup_{[A_0,B_0]\in\mathcal{J}_0}\|B_0\|}{(1-\rho)^2}\right)\mathrm{dist}(\mathcal{J}_0, \mathcal{J})$$

Proof : Given $\mathcal{J}_0 \in \mathcal{C}_\rho$ and $\mathcal{J} \in \mathcal{C}_\rho$, we remark that $\mathcal{J} \subset \mathcal{J}_0 + \mathrm{dist}(\mathcal{J}_0, \mathcal{J})\mathbb{B}_{pm}$, where dist and $\mathbb{B}_{pm}$ are considered with respect to the norm $\|\cdot\|_{pm}$. This means

$$\mathcal{J} \subset \{[A_0,B_0] + [C,D],\ [A_0,B_0] \in \mathcal{J}_0, \|[C,D]\|_{p,m} \leqslant \mathrm{dist}(\mathcal{J}_0,\mathcal{J})\}$$

We have

$$\begin{aligned}
\mathcal{J}(\mathrm{fix}(\mathcal{J}_0)) &= \{AX + B,\ [A,B] \in \mathcal{J},\ X \in \mathrm{fix}(\mathcal{J}_0)\} \\
&\subset \{A_0X + B_0,\ [A_0,B_0] \in \mathcal{J}_0,\ X \in \mathrm{fix}(\mathcal{J}_0)\} \\
&\quad + \{CX + D,\ \|[C,D]\|_{mp} \leqslant \mathrm{dist}(\mathcal{J}_0,\mathcal{J}),\ X \in \mathrm{fix}(\mathcal{J}_0)\} \\
&= \mathcal{J}_0(\mathrm{fix}(\mathcal{J}_0)) + \{CX + D,\ \|[C,D]\|_{mp} \leqslant \mathrm{dist}(\mathcal{J}_0,\mathcal{J}),\ X \in \mathrm{fix}(\mathcal{J}_0)\} \\
&= \mathrm{fix}(\mathcal{J}_0) + \{CX + D,\ \|[C,D]\|_{mp} \leqslant \mathrm{dist}(\mathcal{J}_0,\mathcal{J}),\ X \in \mathrm{fix}(\mathcal{J}_0)\}.
\end{aligned}$$

This sets one inclusion. Similarly, we have

$$\begin{aligned}\mathrm{fix}(\mathcal{J}_0) &= \mathcal{J}_0(\mathrm{fix}(\mathcal{J}_0)) \\ &\subset \mathcal{J}(\mathrm{fix}(\mathcal{J}_0)) + \{CX + D,\ \|[C, D]\|_{mp} \leqslant \mathrm{dist}(\mathcal{J}_0, \mathcal{J}),\ X \in \mathrm{fix}(\mathcal{J}_0)\}\,.\end{aligned}$$

Recall that $\|[C, D]\|_{mp} = \max\{\|C\|_{\mathrm{op}}, \|D\|\}$, we have for any $[C, D]$ with $\|[C, D]\|_{mp} \leqslant \mathrm{dist}(\mathcal{J}_0, \mathcal{J})$ and $X \in \mathrm{fix}(\mathcal{J}_0)$,

$$\|CX + D\| \leqslant \|C\|_{\mathrm{op}}\|\mathrm{fix}(\mathcal{J}_0)\|_{\sup} + \|D\| \leqslant \mathrm{dist}(\mathcal{J}_0, \mathcal{J})(1 + \|\mathrm{fix}(\mathcal{J}_0)\|_{\sup}).$$

We deduce using Lemma 2 that $\mathrm{dist}(\mathrm{fix}(\mathcal{J}_0), \mathcal{J}(\mathrm{fix}(\mathcal{J}_0))) \leqslant \mathrm{dist}(\mathcal{J}_0, \mathcal{J})(1+\|\mathrm{fix}(\mathcal{J}_0)\|_{\sup})$. Setting $\mathcal{X}_0 = \mathrm{fix}(\mathcal{J}_0)$, invoking Theorem 1 with $\mathcal{J}$ and $k = 0$, we have

$$\begin{aligned}\mathrm{dist}(\mathrm{fix}(\mathcal{J}_0), \mathrm{fix}(\mathcal{J})) &\leqslant \frac{\mathrm{dist}(\mathcal{J}_0, \mathcal{J})(1 + \|\mathrm{fix}(\mathcal{J}_0)\|_{\sup})}{1-\rho} \\ &\leqslant \mathrm{dist}(\mathcal{J}_0, \mathcal{J})\frac{(1-\rho+\sup_{[A_0,B_0]\in\mathcal{J}_0}\|B_0\|)}{(1-\rho)^2}.\end{aligned}$$

□

B.3 Perturbed iterations

The following proposition shows that the linear convergence property is actually stable to perturbations. It will be useful to show that all potential limits of unrolling algorithmic differentiation recursions are contained in the corresponding fixed point set.

Proposition 7 (Perturbed set sequences) *Let $\rho < 1$ and $\epsilon > 0$ such that $\rho + \epsilon < 1$. Let $(\mathcal{J}_k)_{k\in\mathbb{N}}$ be a sequence in $\mathcal{C}_{\rho+\epsilon}$ and $\bar{\mathcal{J}} \in \mathcal{C}_\rho$ (as in Definition 5). Assume that for all $k \in \mathbb{N}$*

$$\mathrm{gap}_{pm}(\mathcal{J}_k, \bar{\mathcal{J}}) \leqslant \epsilon$$

or in other words $\mathcal{J}_k \subset \bar{\mathcal{J}} + \epsilon\mathbb{B}_{pm}$ where $\mathbb{B}_{pm}$ is the unit ball of the norm $\|\cdot\|_{pm}$. Then the recursion on compact sets

$$\mathcal{X}_{k+1} = \mathcal{J}_k(\mathcal{X}_k)$$

satisfies for all $k \in \mathbb{N}$

$$\begin{aligned}&\mathrm{gap}(\mathcal{X}_k, \mathrm{fix}(\mathcal{J})) \\ \leqslant\ & (\rho+\epsilon)^k\frac{(1+\rho+\epsilon)\|\mathcal{X}_0\|_{\sup} + \sup_{[A,B]\in\bar{\mathcal{J}}}\|B\| + \epsilon}{1-\rho-\epsilon} + \epsilon\frac{(1-\rho+\sup_{[A,B]\in\bar{\mathcal{J}}}\|B\|)}{(1-\rho)^2}.\end{aligned}$$

In other words, $\mathcal{X}_k \subset \mathrm{fix}(\bar{\mathcal{J}}) + C(\rho, \epsilon, k)\mathbb{B}$ where $C(\rho, \epsilon, k)$ is the constant above.

Proof : Set $\mathcal{J}_\epsilon := \{J + [C, D],\ J \in \bar{\mathcal{J}},\ \|[C, D]\|_{mp} \leqslant \epsilon\}$. Denote by $(\tilde{\mathcal{X}}_k)_{k\in\mathbb{N}}$ the sequence satisfying the recursion, $\tilde{\mathcal{X}}_{k+1} = \mathcal{J}_\epsilon(\tilde{\mathcal{X}}_k)$ with $\mathcal{X}_0 = \tilde{\mathcal{X}}_0$. We have

$$\mathcal{X}_1 = \bar{\mathcal{J}}(\mathcal{X}_0) \subset \mathcal{J}_\epsilon(\mathcal{X}_0) = \tilde{\mathcal{X}}_1$$

and by recursion $\mathcal{X}_k \subset \tilde{\mathcal{X}}_k$ for all $k \in \mathbb{N}$. By Theorem 1, we have

$$\mathrm{dist}(\tilde{\mathcal{X}}_k, \mathrm{fix}(\mathcal{J}_\epsilon)) \leqslant (\rho+\epsilon)^k\frac{\mathrm{dist}(\mathcal{X}_0, \mathcal{J}_\epsilon(\mathcal{X}_0))}{1-\rho-\epsilon}.$$

We deduce from Proposition 6 that for all $k \in \mathbb{N}$,

$$\begin{aligned}&\mathrm{dist}(\tilde{\mathcal{X}}_k, \mathrm{fix}(\bar{\mathcal{J}})) \\ \leqslant\ & \mathrm{dist}(\tilde{\mathcal{X}}_k, \mathrm{fix}(\mathcal{J}_\epsilon)) + \mathrm{dist}(\mathrm{fix}(\mathcal{J}_\epsilon), \mathrm{fix}(\bar{\mathcal{J}})) \\ \leqslant\ & (\rho+\epsilon)^k\frac{\mathrm{dist}(\mathcal{X}_0, \mathcal{J}_\epsilon(\mathcal{X}_0))}{1-\rho-\epsilon} + \frac{(1-\rho+\sup_{[A,B]\in\bar{\mathcal{J}}}\|B\|)}{(1-\rho)^2}\mathrm{dist}(\mathcal{J}_\epsilon, \bar{\mathcal{J}}) \\ \leqslant\ & (\rho+\epsilon)^k\frac{(1+\rho+\epsilon)\|\mathcal{X}_0\|_{\sup} + \sup_{[A,B]\in\bar{\mathcal{J}}}\|B\| + \epsilon}{1-\rho-\epsilon} + \frac{(1-\rho+\sup_{[A,B]\in\bar{\mathcal{J}}}\|B\|)}{(1-\rho)^2}\epsilon.\end{aligned}$$

And the result follows because

$$\max_{X \in \mathcal{X}_k} \min_{L \in \mathrm{fix}(\bar{\mathcal{J}})} \|X - L\| \leqslant \max_{X \in \tilde{\mathcal{X}}_k} \min_{L \in \mathrm{fix}(\bar{\mathcal{J}})} \|X - L\| \leqslant \mathrm{dist}(\tilde{\mathcal{X}}_k, \mathrm{fix}(\bar{\mathcal{J}})).$$

□

This allows to obtain explicit convergence results as follows

Corollary 4 (Limit of iterations with vanishing perturbations) *Let $\rho < 1$ and $\bar{\mathcal{J}} \in \mathcal{C}_\rho$ (as in Definition 5). Let $(\mathcal{J}_k)_{k \in \mathbb{N}}$ be a sequence of matrices such that for all $k \in \mathbb{N}$*

$$\mathrm{gap}_{pm}(\mathcal{J}_k, \bar{\mathcal{J}}) \leqslant \epsilon_k$$

where $(\epsilon_k)_{k \in \mathbb{N}}$ is a positive sequence such that there exists a constant $a > 0$ such that $\epsilon_k \leqslant a\rho^k$ for all $k \in \mathbb{N}$. Then for the recursion on compact sets of $p \times m$ matrices

$$\mathcal{X}_{k+1} = \mathcal{J}_k(\mathcal{X}_k)$$

There are constants $C, c > 0$ such that for all $k \in \mathbb{N}$

$$\mathrm{gap}(\mathcal{X}_k, \mathrm{fix}(\mathcal{J})) \leqslant Ce^{-ck}.$$

Furthermore, one can take $c = \log\left(\frac{1}{\sqrt{\rho+\epsilon}}\right)$ for arbitrary $\epsilon > 0$.

Proof : We consider $K \in \mathbb{N}$ such that $\epsilon_k \leqslant \epsilon$ for all $k \in \mathbb{N}$ where $\epsilon + \rho < 1$. Without loss of generality, we may assume that $K = 0$. Using the same notations as in the proof of Proposition 7, we have $\mathcal{X}_k \subset \tilde{\mathcal{X}}_k$ for all $k \in \mathbb{N}$. Furthermore, it follows from the same arguments as in the proof of Theorem 1 that

$$\|\mathcal{X}_k\|_{\sup} \leqslant \|\tilde{\mathcal{X}}_k\|_{\sup} \leqslant M, \tag{17}$$

for a constant $M > 0$. Now choose $k \in \mathbb{N}$, applying Proposition 7 shifting the initialization 0 to k, we have for all $m \in \mathbb{N}$

$$\begin{aligned}
&\max_{X \in \mathcal{X}_{k+m}} \min_{L \in \mathrm{fix}(\mathcal{J})} \|X - L\| \\
\leqslant \quad & (\rho + \epsilon_k)^m \frac{(1 + \rho + \epsilon_k)\|\mathcal{X}_k\|_{\sup} + \sup_{[A,B] \in \bar{\mathcal{J}}} \|B\| + \epsilon_k}{1 - \rho - \epsilon_k} + \epsilon_k \frac{(1 - \rho + \sup_{[A,B] \in \bar{\mathcal{J}}} \|B\|)}{(1-\rho)^2} \\
\leqslant \quad & (\rho + \epsilon)^m \frac{(1 + \rho + \epsilon)M + \sup_{[A,B] \in \bar{\mathcal{J}}} \|B\| + \epsilon}{1 - \rho - \epsilon} + a\rho^k \frac{(1 - \rho + \sup_{[A,B] \in \bar{\mathcal{J}}} \|B\|)}{(1-\rho)^2},
\end{aligned}$$

where we have used the bound (17) and the fact that $\epsilon_k \leqslant \epsilon$ and $\epsilon_k \leqslant a\rho^k$. Setting $u = \frac{(1+\rho+\epsilon)M + \sup_{[A,B] \in \bar{\mathcal{J}}} \|B\| + \epsilon}{1-\rho-\epsilon}$ and $v = a\frac{(1-\rho+\sup_{[A,B] \in \bar{\mathcal{J}}} \|B\|)}{(1-\rho)^2}$ we have

$$\begin{aligned}
\max_{X \in \mathcal{X}_{2k}} \min_{L \in \mathrm{fix}(\mathcal{J})} \|X - L\| \leqslant \quad & u(\rho + \epsilon)^k + v\rho^k \leqslant (u+v)(\rho+\epsilon)^{2k/2} \leqslant \frac{u+v}{(\rho+\epsilon)^{1/2}}(\rho+\epsilon)^{2k/2}, \\
\max_{X \in \mathcal{X}_{2k+1}} \min_{L \in \mathrm{fix}(\mathcal{J})} \|X - L\| \leqslant \quad & u(\rho + \epsilon)^{k+1} + v\rho^k \leqslant \frac{u+v}{(\rho+\epsilon)^{1/2}}(\rho+\epsilon)^{(2k+1)/2}.
\end{aligned}$$

Since k was arbitrary, this proves the desired result. □

C Existence of a conservative Jacobian for autodiff

C.1 Regularity of $J_{\bar{x}}^{\mathrm{pb}}$

We recall the main notations and elements of Assumption 1. We assume that F is locally Lipschitz, path differentiable, and denote by $J_F \colon \mathbb{R}^{p+m} \rightrightarrows \mathbb{R}^{p \times (p+m)}$ a conservative Jacobian for F. Now assume that any pair $[A, B] \in J_F(x, \theta)$ is such that the operator norm of A is at most $\rho < 1$, that is for all x and θ, $J_F(x, \theta) \in \mathcal{C}_\rho$ (as in Definition 5). Define the following set-valued map

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

Here, $\bar{x}(\theta) = \mathrm{fix}(F_\theta)$ is the unique fixed point of the algorithmic recursion, so that we actually have

$$J_{\bar{x}}^{\mathrm{pb}} \colon \theta \rightrightarrows \mathrm{fix}\left[J_F(\mathrm{fix}(F_\theta), \theta)\right].$$

We have the following

Lemma 3 (Regularity of $J_{\bar{x}}^{\text{pb}}$) *The mapping $J_{\bar{x}}^{\text{pb}}$ is nonempty valued, locally bounded and has a closed graph.*

Proof : The fact that $J_{\bar{x}}^{\text{pb}}$ is locally bounded and nonempty valued comes from the fact that J_F is locally bounded with nonempty values and $\bar{x}$ is locally Lipschitz combined with Theorem 1.

By local Lipschitz continuity of $\bar{x}$ and the fact that J_F has a closed graph, the set-valued map $\theta \rightrightarrows J_F(\bar{x}(\theta), \theta)$ also has a closed graph. By continuity of $\text{fix}(\mathcal{J})$ with respect to the Hausdorff distance, see Proposition 6, $J_{\bar{x}}^{\text{pb}}$ has a closed graph. □

C.2 Proof of Theorem 2

Proof : Following Remark 3, we consider

$$J_{\bar{x}}^{\text{imp}} \colon \theta \rightrightarrows \left\{ (I - A)^{-1} B, [A, B] \in J_F(\bar{x}(\theta), \theta) \right\},$$

a conservative Jacobian for $\bar{x}$ and $L_0 = J_{\bar{x}}^{\text{imp}}$. Now, define by recursion for all $k \in \mathbb{N}$

$$L_{k+1} \colon \theta \rightrightarrows J_F(\bar{x}(\theta), \theta)(L_k(\theta)).$$

Recall that this means that for all $\theta \in \mathbb{R}^m$ and $k \in \mathbb{N}$

$$L_{k+1}(\theta) = \{AL + B,\ [A, B] \in J_F(\bar{x}(\theta), \theta),\ L \in L_k(\theta)\}.$$

Since $F(\bar{x}(\theta), \theta) = \bar{x}(\theta)$ for all θ, J_F is conservative for F and L_0 is conservative for $\bar{x}$, we have by induction that for all $k \in \mathbb{N}$, L_k is conservative for $\bar{x}$.

Fix $l \colon \mathbb{R}^m \to \mathbb{R}^m$ an arbitrary Borel measurable selection in $J_{\bar{x}}^{\text{pb}}$, that is $l(\theta) \in J_{\bar{x}}^{\text{pb}}(\theta)$ for all $\theta \in \mathbb{R}^m$. Such a selection exist by [4, Theorem 18.20] because $J_{\bar{x}}^{\text{pb}}$ has a closed graph by Lemma 3. Consider for all $k \in \mathbb{N}$, a measurable selection

$$l_k \colon \theta \to \arg \min_{z \in L_k(\theta)} \|z - l(\theta)\|.$$

The function $(z, \theta) \to \|z - l(\theta)\|$ is Caratheodory (continuous in z, measurable in θ), so such a selection exists (Aliprantis Theorem 18.19). By Theorem 1, we have that $\text{dist}(L_k(\theta), J_{\bar{x}}^{\text{pb}}(\theta))$ tends to 0 as k grows, for all $\theta \in \mathbb{R}^m$, where the convergence is in Hausdorff distance. Actually since all set-valued objects are locally bounded, the convergence occurs uniformly on every compact. This implies in particular that l_k converges pointwise to l.

Fix an absolutely continuous path $\gamma \colon [0, 1] \to \mathbb{R}^m$. We have for all $k \in \mathbb{N}$, by conservativity,

$$\bar{x}(\gamma(1)) - \bar{x}(\gamma(0)) = \int_0^1 l_k(\gamma(t))\dot{\gamma}(t)dt.$$

Furthermore, $l_k \circ \gamma$ is measurable, converges pointwise to $l \circ \gamma$ and $l_k \circ \gamma$ can be uniformly bounded, let K be such a bound. The integrable function $g \colon t \mapsto K\|\dot{\gamma}(t)\|$ dominates the integrand and $l_k \circ \gamma \times \dot{\gamma}$ converges pointwise to $l \circ \gamma \times \dot{\gamma}$. By the dominated convergence theorem (see [47, Section 4.4]), we have

$$\bar{x}(\gamma(1)) - \bar{x}(\gamma(0)) = \int_0^1 l(\gamma(t))\dot{\gamma}(t)dt.$$

L has a Castaing representation with a dense sequence of measurable selection [4, Theorem 18.14]. Since l was an arbitrary measurable selection in L, conservativity of L follows by [38, Lemma 8]. □

C.3 Proof of Corollary 1

Proof : Fix θ. We have $x_k(\theta) \to \bar{x}(\theta)$, so that for any $\epsilon > 0$, there exists $K \in \mathbb{N}$ such that $J_F(x_k(\theta), \theta) \subset J_F(\bar{x}(\theta), \theta) + \epsilon\mathbb{B}$ for all $k \geqslant K$. The result is then a consequence of Proposition 7, letting $\epsilon \to 0$. The last part is due to the conservativity of $J_{\bar{x}}^{\text{pb}}$ which must be a singleton almost everywhere, equal to the classical Jacobian. □

C.4 Proof of Corollary 3

Proof : Define $(L_k)_{k\in\mathbb{N}}$, a sequence of conservative Jacobians for $\bar{x}$ as in the begining of the proof of Theorem 2 in Appendix C.2. By [12, Theorem 1], for each $k \in \mathbb{N}$, there is a full measure set $S_k \subset \mathbb{R}^m$ such that $L_k(\theta) = \left\{\frac{\partial \bar{x}}{\partial \theta}(\theta)\right\}$ for all $\theta \in S_k$. Similarly, there exists a full measure set $S_{-1} \subset \mathbb{R}^m$ such that $J_{\bar{x}}^{\mathrm{pb}}(\theta) = \left\{\frac{\partial \bar{x}}{\partial \theta}(\theta)\right\}$ for all $\theta \in S_{-1}$. Setting $S = \cap_{i=-1}^{+\infty} S_i$, S has full measure and for all $\theta \in S$ and for all $k \in \mathbb{N}$,

$$J_{\bar{x}}^{\mathrm{pb}}(\theta) = \left\{\frac{\partial \bar{x}}{\partial \theta}(\theta)\right\} \qquad L_k(\theta) = \left\{\frac{\partial \bar{x}}{\partial \theta}(\theta)\right\}.$$

Following the proof of Theorem 2 in Appendix C.2, L_k converges to $J_{\bar{x}}^{\mathrm{pb}}$ in Hausdorff distance, which means that convergence occurs in the classical sense since all sets in the sequence are singletons. □

C.5 Proof of Proposition 1

Proof : Under the setting of Corollary 2, for almost all $\theta \in \mathbb{R}^m$, recursion (PB) or (5) reduce to the following, and all $k \in \mathbb{N}$

$$J_{k+1} = A_k J_k + B_k \tag{18}$$

where $J_k = \frac{\partial x_k}{\partial \theta}$, $A_k = \frac{\partial F}{\partial x}(x_k, \theta)$ and $B_k = \frac{\partial F}{\partial \theta}(x_k, \theta)$ are classical Jacobians and J_k converges to the classical Jacobian of $\frac{\partial \bar{x}}{\partial \theta}(\theta)$. Fix such a $\theta \in \mathbb{R}^m$ and $k \in \mathbb{N}$, $k \geqslant 1$. With the notation of Algorithm 1, for the forward mode, multiplying (18) on the right by $\dot{\theta}$, we have for all $i \in 1, \ldots k$

$$J_i \dot{\theta} = A_{i-1} J_{i-1} \dot{\theta} + B_{i-1} \dot{\theta}.$$

Setting $\dot{x}_i = J_i \dot{\theta}$, this is exactly the recursion implemented by Algorithm 1 in forward mode. Corollary 2 and the result follows from convergence of J_k.

As for the backward mode a simple recursion shows that

$$\begin{aligned} J_k = \;& A_{k-1} A_{k-2} \ldots A_0 J_0 \\ &+ A_{k-1} A_{k-2} \ldots A_1 B_0 \\ &+ \ldots \\ &+ A_{k-1} A_{k-2} \ldots A_i B_{i-1} \\ &+ \ldots \\ &+ A_{k-1} B_{k-2} \\ &+ B_{k-1}. \end{aligned} \tag{19}$$

Setting $B_{-1} = J_0$, we may rewrite equivalently,

$$J_k = B_{k-1} + \sum_{i=0}^{k-1} \left(\prod_{j=k-1}^{i} A_j \right) B_{i-1}. \tag{20}$$

Transposing and multiplying on the right by $\bar{w}_k$, we have

$$J_k^T \bar{w}_k = B_{k-1}^T \bar{w}_k + \sum_{i=0}^{k-1} B_{i-1}^T \left(\prod_{j=i}^{k-1} A_j^T \right) \bar{w}_k. \tag{21}$$

We set for all $i = 0, \ldots, k-1$,

$$\bar{w}_i = \prod_{j=i}^{k-1} A_j^T \bar{w}_k. \tag{22}$$

We have the backward recursion relation, for $i = k, \ldots, 1$

$$\bar{w}_{i-1} = A_{i-1}^T \bar{w}_i,$$

which is the recursion implemented by Algorithm 1 in reverse mode. Combining (21) and (22), we obtain

$$J_k^T \bar{w}_k = B_{k-1}^T \bar{w}_k + \sum_{i=0}^{k-1} B_{i-1} \bar{w}_i = \sum_{i=1}^{k} B_{i-1}^T \bar{w}_i + J_0^T \bar{w}_0,$$

which is the quantity accumulated in $\bar{\theta}_k$ in Algorithm 1. This proves that $\bar{\theta}_k^T$ returned by the backward mode is indeed equal to $\bar{w}_k^T J_k$ and the convergence follows from convergence of both $\bar{w}_k$ and J_k as $k \to \infty$. □

D Connection with implicit differentiation

Recall that for all θ

$$\begin{aligned} J_{\bar{x}}^{\text{imp}}(\theta) &= \left\{(I - A)^{-1}B, [A, B] \in J_F(\bar{x}(\theta), \theta)\right\} \\ &= \left\{M, \exists [A, B] \in J_F(\bar{x}(\theta), \theta)\; M = AM + B\right\}. \end{aligned}$$

Setting $\mathcal{J} = J_F(\bar{x}(\theta), \theta)$, we have therefore that $J_{\bar{x}}^{\text{imp}}(\theta) \subset \mathcal{J}(J_{\bar{x}}^{\text{imp}}(\theta))$. By recursion, for all $k \in \mathbb{N}$, $J_{\bar{x}}^{\text{imp}}(\theta) \subset \mathcal{J}^k(J_{\bar{x}}^{\text{imp}}(\theta))$ and passing to the limit using Theorem 1, $J_{\bar{x}}^{\text{imp}}(\theta) \subset \text{fix}(\mathcal{J}) = J_{\bar{x}}^{\text{pb}}(\theta)$. In particular, if F is continuously differentiable, then (PB) with classical Jacobians converges towards a classical implicit derivative.

However, the inclusion $J_{\bar{x}}^{\text{imp}}(\theta) \subset J_{\bar{x}}^{\text{pb}}(\theta)$ may be strict as the following example shows.

Example 1 Set $\mathcal{J} = \{[A, B],\ A \in \mathcal{A}, B \in \mathcal{B}\}$, where

$$\mathcal{A} = \left\{ \begin{pmatrix} \frac{\lambda+1}{4} & 0 \\ 0 & \frac{2-\lambda}{4} \end{pmatrix}, \lambda \in [0, 1] \right\} \qquad \mathcal{B} = \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}.$$

We set

$$\mathcal{T} = (I - \mathcal{A})^{-1}\mathcal{B} = \left\{ \begin{pmatrix} \frac{4}{3-\lambda} \\ \frac{4}{2+\lambda} \end{pmatrix}, \lambda \in [0, 1] \right\}.$$

As already observed, we have $\mathcal{T} \subset \mathcal{A}\mathcal{T} + \mathcal{B}$, but the inclusion is strict. Therefore $\mathcal{T}$ is not a fixed point of the affine iteration and it is only contained in it.

Indeed, we have

$$\begin{pmatrix} \frac{1+1}{4} & 0 \\ 0 & \frac{2-1}{4} \end{pmatrix} \begin{pmatrix} \frac{4}{3-0} \\ \frac{4}{2+0} \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{5}{3} \\ \frac{3}{2} \end{pmatrix} \in \mathcal{A}\mathcal{T} + \mathcal{B}.$$

However solving for λ

$$\begin{pmatrix} \frac{5}{3} \\ \frac{3}{2} \end{pmatrix} = \begin{pmatrix} \frac{4}{3-\lambda} \\ \frac{4}{2+\lambda} \end{pmatrix},$$

the first equation requires $\lambda = \frac{3}{5}$ while the second requires $\lambda = \frac{2}{3}$ which shows that the given vector does not belong to $\mathcal{T}$.

E Semialgebraic Lipschitz gradient selection functions

E.1 Lipschitz property of conservative Jacobians of selections

Lemma 4 (Conservative Jacobians of selections are Lipschitz-like) *Let F be continuous, semi-algebraic with Lipschitz gradient selection. Then for each $x_0 \in \mathbb{R}^p$, there exists $R > 0$ such that*

$$\text{gap}(J_F^s(x), J_F^s(x_0)) \leqslant L\|x - x_0\|, \qquad \forall x,\ \|x - x_0\| \leqslant R,$$

where L is the Lipschitz constant given by the selection structure of F.

Proof : Fix $x_0 \in \mathbb{R}^p$ and consider the function g which associates to $r > 0$ a subset of $\{1, \dots, m\}$ defined as

$$g(r) = \cup_{\|x - x_0\| \leqslant r} I(x).$$

The function g is semialgebraic and therefore it admits a limit as $r \to 0$. The function g is actually piecewise constant so that the limit is reached for some $R > 0$ by semialgebraicity. This means that there is $R > 0$ and an index set $I \subset \{1, \dots, m\}$ such that $I(x) \subset I$ for all x such that $\|x - x_0\| \leqslant R$. Furthermore, for each $i \in I$ and all $0 < r \leqslant R$, there exists x such that $\|x - x_0\| \leqslant r$ and $F_i(x) = F(x)$. By continuity of each component F_i, we have for each $i \in I$, $F_i(x_0) = F(x_0)$, that is $I \subset I(x_0)$.

We deduce that for each x such that $\|x - x_0\| \leqslant R$ and $i \in I(x)$, we have

$$\min_{V \in J_F^s(x_0)} \left\| V - \frac{\partial F_i}{\partial x}(x) \right\| \leqslant \left\| \frac{\partial F_i}{\partial x}(x_0) - \frac{\partial F_i}{\partial x}(x) \right\| \leqslant L\|x - x_0\|.$$

Fix any $Z \in J_F^s(x)$, it is a convex combination of $\frac{\partial F_i}{\partial x}(x)$ for $i \in I(x)$ so by convexity of the distance, we have

$$\min_{V \in J_F^s(x_0)} \|V - Z\| \leqslant L\|x - x_0\|,$$

which proves the result since this allows to bound the supremum over $Z \in J_F^s(x)$ by the desired quantity. □

E.2 Proof of Corollary 3

Proof : This is a consequence of linear convergence of the recursion $x_{k+1} = F(x_k, \theta)$ combined with Lemma 4 and Corollary 4. □

F Proximal splitting algortihms in convex optimization

F.1 Proof of Proposition 2

Proof : We consider the gradient step operation $H_\alpha \colon (x, \theta) \mapsto x - \alpha \nabla_x f(x, \theta)$. We have for all (x, θ),

$$F_\alpha(x, \theta) = G_\alpha(H_\alpha(x, \theta), \theta).$$

By Assumption 2, both G_α and H_α are 1-Lipschitz in x for fixed θ and we are going to show that if either f or g satisfy the strong convexity condition, the corresponding map is a strict contraction in x for fixed θ. Furthermore, the mapping $\mathrm{Jac}^{c}_{H_\alpha} \colon (x, \theta) \rightrightarrows \left\{[I - \alpha A, -\alpha B],\ [A, B] \in J_f^2(x, \theta)\right\}$ is the Clarke Jacobian of H_α. By Assumption 2, all the functions are path-differentiable [12] and one may obtain a conservative jacobian for F by applying differential calculus rules [12]. We set for all (x, θ) a conservative Jacobian for F_α,

$$J_{F_\alpha}(x, \theta) = \left\{[C(I - \alpha A), -\alpha CB + D], \quad [A, B] \in J_f^2(x, \theta),\ [C, D] \in J_{G_\alpha}(x - \alpha \nabla_x f(x, \theta), \theta)\right\} \tag{23}$$

Whenever $\nabla_x f$ is differentiable at (x, θ), the first p columns of its Jacobian form a symmetric positive definite square matrix with eigenvalues at most L. This implies that the matrix $(I - \alpha A)$ in (23) is symmetirc with eignevalues in $[-1, 1]$ and strictly greater than -1. Similarly, whenever G_α is differentiable, since it is 1-Lipschitz in x for fixed θ and the gradient of a C^1 function, the first p columns of its Jacobian form a symmetric positive definite square matrix with eigenvalues at most 1. This implies that the matrix C in (23) is symmetric with eignevalues in $[0, 1]$. In addition, we have the following;

- Assume that for all θ, f is μ-strongly convex. In this case, similarly as above the matrix $(I - \alpha A)$ in (23) has eigenvalue in $(-1, 1)$ for all (x, θ).
- Assume that for all θ, g is μ-strongly convex. In this case, similarly as above the matrix C in (23) has eigenvalue in $[0, 1/(1 + \alpha\mu)]$ for all (x, θ) [6, Proposition 23.13].

In both cases, the product $C(I - \alpha A)$ in (23) has operator norm strictly smaller than 1 and Assumption 1 holds. □

F.2 Proof of Proposition 3

Proof : From [6, Proposition 23.11], both $R_{\alpha f}$ and $R_{\alpha g}$ are 1-Lipschitz. We are going to show that $R_{\alpha f}$ is a strict contraction and the result will follow. Since f is $C^{1,1}$ in x, we have for all $\theta \in \mathbb{R}^m$,

$$z = \mathrm{prox}_{\alpha f(\cdot,\theta)}(x) \Leftrightarrow z + \alpha \nabla_x f(z,\theta) - x = 0$$

Set $H_\alpha(z,x,\theta) = z + \alpha \nabla_x f(z,\theta) - x$, we have that

$$\mathrm{Jac}^{\,c}_{H_\alpha}(z,x,\theta) \rightrightarrows \{[I + \alpha A,\, -I,\, \alpha B]\} \tag{24}$$

is the Clarke Jacobian of H_α. Similarly as in Appendix F.1, by strong convexity of f, the matrix $I + \alpha A$ in (24) is symmetric with eigenvalues strictly greater than 0 and smaller than 1. By implicit differential calculus rule in [10, Theorem 2], the mapping

$$J_{\mathrm{prox}_{\alpha f(\cdot,\theta)}}(x,\theta) \rightrightarrows \left\{[(I+\alpha A)^{-1},\, -\alpha(I+\alpha A)^{-1}B],\, [A,B] \in J_f^2(\mathrm{prox}_{\alpha f(\cdot,\theta)},\theta)\right\} \tag{25}$$

is conservative for $(x,\theta) \mapsto \mathrm{prox}_{\alpha f(\cdot,\theta)}$. Furthermore, the matrix $(I+\alpha A)^{-1}$ in (25) is symmetric eigenvalues in $(0,1)$. This entails that the mapping

$$J_{R_{\alpha f(\cdot,\theta)}}(x,\theta) \rightrightarrows \left\{[2(I+\alpha A)^{-1} - I,\, -2\alpha(I+\alpha A)^{-1}B - I],\, [A,B] \in J_f^2(\mathrm{prox}_{\alpha f(\cdot,\theta)},\theta)\right\} \tag{26}$$

is conservative for $R_{\alpha f(\cdot,\theta)}$ and the matrix $2(I+\alpha A)^{-1} - I$ is symmetric with eigenvalues in $(-1,1)$.

Similarly, the mapping

$$J_{R_{\alpha g(\cdot,\theta)}}(x,\theta) \rightrightarrows \left\{[2C - I,\, 2D - I],\, [C,D] \in J_{\mathrm{prox}_{\alpha g(x,\theta)}}\right\} \tag{27}$$

is the Clarke Jacobian of $R_{\alpha g(\cdot,\theta)}$ and the matrix $2C - I$ in (27) is symmetric with eigenvalues in $[-1,1]$. One may combine $J_{R_{\alpha f(\cdot,\theta)}}$ and $J_{R_{\alpha g\cdot,\theta)}}$, using differential calculus rule to obtain a conservative Jacobian J_{F_α} for F_α, such that for all (x,θ) and $[E,F] \in J_{F_\alpha}(x,\theta)$, the square matrix E is of the form $\frac{I}{2} + ((I+\alpha A)^{-1} - I)(2C - I)$ where A is from (26) and C is from (27). Such a matrix E has operator norm strictly smaller than 1 which is Assumption 1. □

F.3 Equivalence between ADMM and dual Douglas–Rachford

We need the following lemma.

Lemma 5 *Let F, G two convex, lower semicontinuous and closed functions and h defined by*

$$h(x) = F^*(-A^\top x) + G^*(x).$$

Then, h is convex, lower semicontinuous, closed, and

$$\mathrm{prox}_{\alpha h}(x) = x + \alpha(A\hat{u} - \hat{v}) \tag{28}$$

where

$$(\hat{u},\hat{v}) \in \arg\min_{u,v} \left\{F(u) + G(v) + x^\top(Au - v) + \frac{\alpha}{2}\|Au - v\|_2^2\right\}.$$

The material contained in this section is already known in the litterature accross several papers and lecture notes, but for the sake of completeness, we include a full derivation of the equivalence.

In this appendix, we drop the dependency to the variable θ since we are only concerned on the behaviour with respect to x. We recall that the iteration of Douglas–Rachford are defined by an initialization y_0 and the recursion

$$\begin{aligned} x_{k+1} &= \mathrm{prox}_f(y_k) \\ y_{k+1} &= y_k + \mathrm{prox}_g(2x_{k+1} - y_k) - x_{k+1}. \end{aligned} \tag{29}$$

By denoting $\tilde{x}_k = x_{k+1}$ and $\tilde{y}_k = y_k$, we can rewrite the updates of Douglas–Rachford (given $\tilde{x}_0$ and $\tilde{y}_0$) as

$$\begin{aligned} \tilde{y}_{k+1} &= \tilde{y}_k + \mathrm{prox}_g(2\tilde{x}_k - \tilde{y}_k) - \tilde{x}_k. \\ \tilde{x}_{k+1} &= \mathrm{prox}_f(\tilde{y}_{k+1}) \end{aligned} \tag{30}$$

Introducing the variable $\hat{r} = \text{prox}_g(2\hat{x} - \hat{y})$, this is also equivalent to

$$\begin{aligned} \hat{r}_{k+1} &= \text{prox}_g(2\hat{x}_k - \hat{y}_k) \\ \hat{x}_{k+1} &= \text{prox}_f(\hat{y}_k + \hat{r}_{k+1} - \hat{x}_k) \\ \hat{y}_{k+1} &= \hat{y}_k + \hat{r}_{k+1} - \hat{x}_k \end{aligned} \tag{31}$$

Using the change of variable $\hat{w}_k = \hat{x}_k - \hat{y}_k$, we have

$$\begin{aligned} \hat{r}_{k+1} &= \text{prox}_g(\hat{x}_k + \hat{w}_k) \\ \hat{x}_{k+1} &= \text{prox}_f(\hat{r}_{k+1} - \hat{w}_k) \\ \hat{w}_{k+1} &= \hat{w}_k + \hat{x}_{k+1} - \hat{r}_{k+1}. \end{aligned} \tag{32}$$

This formulation will be convenient to show how to retrieve the equations of ADMM (13).

The dual problem of (12) is given by (14)

$$\max_x -f(x) - g(x). \tag{33}$$

where $f(x) = \phi^\star(-Ax) + c^\top x$ and $g(x) = \psi(-Bx)$

We consider the update rules given by (32), i.e.,

$$\hat{r} = \text{prox}_{\alpha g}(x + w) \tag{34}$$

$$\hat{x} = \text{prox}_{\alpha f}(\hat{r} - w) \tag{35}$$

$$\hat{w} = w + \hat{x} - \hat{r}. \tag{36}$$

Applying Lemma 5 to $F = \phi$ and $G = \iota_c$, we rewrite (34) by

$$\hat{r} = x + w + \alpha(A\hat{u} - c)$$

where

$$\hat{u} = \arg\min_u \left\{ \phi(x) + x^\top(Au - v) + \frac{\alpha}{2}\|Au - c + w/\alpha\|_2^2 \right\}.$$

Using the same lemma to $F = \psi$ and $G = 0$, we rewrite (35) by

$$\hat{x} = x + \alpha(A\hat{u} + B\hat{v} - c)$$

where

$$\hat{v} = \arg\min_v \left\{ \psi(v) + x^\top Bv + \frac{\alpha}{2}\|A\hat{u} + Bv - c \right\}.$$

Finaly, combining the expression of $\hat{r}$ and $\hat{x}$, we obtain

$$\hat{w} = \alpha B\hat{v}.$$

G Inertial methods

Let us first recall notations from Section 5. Consider a function $f : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$, and $\beta > 0$, for simplicity, when the second argument is fixed we write $f_\theta : x \mapsto f(x, \theta)$. Set for all x, y, θ, $F(x, y, \theta) = (x - \nabla f_\theta(x) + \beta(x - y), x)$, consider the Heavy-Ball algorithm $(x_{k+1}, y_{k+1}) = F(x_k, y_k, \theta)$ for $k \in \mathbb{N}$. If f_θ is μ-strongly convex with L-Lipschitz gradient, then, choosing $\alpha = 1/L$ and $\beta < \frac{1}{2}\left(\frac{\mu}{2L} + \sqrt{\frac{\mu^2}{4L^2} + 2}\right)$, the algorithm will converge globally at a linear rate to the unique solution,

G.1 Failure of Forward differentiation for $C^{1,1}$ objectives

The Jacobian of F for the Heavy-Ball agorithm (in x, y) is of the form

$$\text{Jac}_F(x, y, \theta) = \begin{pmatrix} (I - \alpha\nabla^2 f_\theta(x)) + \beta I & -\beta I \\ I & 0 \end{pmatrix}, \tag{37}$$

when f is C^2. If f is $C^{1,1}$, then the Hessian can be replaced by a set-valued conservative Jacobian of the gradient: $J_{\nabla f_\theta}$.

Proof of Proposition 4:

Recall that the function $f\colon \mathbb{R}^2 \to \mathbb{R}$ is given by

$$f\colon (x,\theta) \mapsto \begin{cases} \frac{x^2}{2} & \text{if } x \geqslant 0 \\ \frac{x^2}{8} & \text{if } x < 0. \end{cases}$$

We have $f'(x) = x$ for $t \geqslant 0$ and $f'(x) = \frac{x}{4}$ for $t < 0$, therefore, f' is 1-Lipschitz. The Clarke subdifferential of f' is $\{\frac{1}{4}\}$ for $t < 0$, $\{1\}$ for $t > 0$ and the segment $\left[\frac{1}{4}, 1\right]$ at $t = 0$. Finally, f is $\mu = \frac{1}{4}$ strongly convex and has $L = 1$ Lipschitz gradient and the unique fixed point of the Heavy-Ball algorithm applied to $f(\cdot,\theta)$ is $x = y = \theta$. Choosing $\alpha = 1$, $\beta = 0.75$, we have

$$\beta < \frac{1}{2}\left(\frac{\mu}{2L} + \sqrt{\frac{\mu^2}{4L^2} + 2}\right) = \frac{1}{2}\left(\frac{1}{8} + \sqrt{\frac{1}{64} + 2}\right) \simeq 0.77.$$

Therefore, the heavy ball algorithm with this choice of parameter converges linearly to the unique solution which is 0, a fixed point of the iteration mapping.

Set

$$F(x,y,\theta) = (x - \nabla_x f(x,\theta) + \beta(x-y), x).$$

At $(0,0,0)$, the last column of the Jacobian of F is $(0,0)$ and the first two columns are given by

$$J = \operatorname{conv}\{M_1, M_2\},$$

where

$$M_1 = \begin{pmatrix} \frac{3}{2} & -\frac{3}{4} \\ 1 & 0 \end{pmatrix} \qquad M_2 = \begin{pmatrix} \frac{3}{4} & -\frac{3}{4} \\ 1 & 0 \end{pmatrix}.$$

Therefore, the Clarke Jacobian of F (with respect to x, y) at $(0,0,0)$ is given by

$$J_F(0,0,0) = \operatorname{conv}\{M_1, M_2\}, \qquad M_1 = \begin{pmatrix} \frac{3}{2} & -\frac{3}{4} \\ 1 & 0 \end{pmatrix}, \qquad M_2 = \begin{pmatrix} \frac{3}{4} & -\frac{3}{4} \\ 1 & 0 \end{pmatrix}.$$

We have

$$M_1 M_1 M_2 M_2 = \frac{-1}{32}\begin{pmatrix} 36 & 0 \\ 27 & 9 \end{pmatrix},$$

which has two eigenvalues $\frac{-9}{8} < -1$ and $\frac{-9}{32}$. Setting for any $\theta \in \mathbb{R}$ $x_0(\theta) = \theta$, $y_0(\theta) = \theta$, we have for all $k \in \mathbb{N}$ $x_k(\theta) = y_k(\theta) = \theta$, in other words, this is the unique fixed point of the Heavy-Ball algorithm.

□

Given $l \in \mathbb{N}$, the forward propagation recursion in (PB) presented in Figure 3 satisfies for $k = 8l$

$$(M_1 M_1 M_2 M_2)^{2l} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

This products will diverge diverge due to the eigenvalue of $(M_1 M_1 M_2 M_2)^2$ strictly above 1. In other words, for all k, $J_{x_{8k}}$ given by (PB) contains elements which magnitude diverges at a geometric rate. We conclude that, for all $k \in \mathbb{N}$, J_{x_k} contains elements which magnitude diverge at a geometric rate.

This illustrates the failure of forward derivative propagation on $f(\cdot,\theta)$: the Heavy Ball algorithm is stable and globally linearly convergent, its fixed point is differentiable (it is actually constant in θ), yet there is a parametric initialization $x(\theta), y(\theta)$ such that forward propagation of derivatives produces diverging elements for $\theta = 0$. Note that implicit differentiation provides the correct derivative, which is 0, since $x(\theta) = 0$ is the unique fixed point of the gradient iterations. Forward derivative propagation on the gradient descent algorithms also results in the limit in 0 derivative since it only contains element which converge to 0 at a geometric rate.

Le us emphasize again that such pathology would not happen if f was C^2. Indeed, in this case, J_f^2 would be single valued and the divergence phenomenon would not appear. This illustrate a fundamental difference between $C^{1,1}$ and C^2 objectives in terms of forward derivative propagation for second order inertial methods.

H Experiments details

All the experiments where run on a MacBook M1 Pro (arm64), on Python 3.9 and `numpy` 1.21 for a compute time inferior to one hour. They are repeated 100 times, and we report the median as a blue line and the first and last deciles as a blue shaded area. The solutions are computed with 2000 iterations, and the curves are reported for the 1000 first iterations. The differentiation of all methods is performed in forward-mode with `jacfwd` of the module `jax`.

Forward–Backward for the Ridge. The dimensions of the problem are $n = 500$, $p = 300$. The design matrix is Gaussian, i.e., $X_{i,j} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$ and the observations $y_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$. The regularization parameter is set to $\theta = 0.05$.

Forward–Backward algorithm for the Lasso. The dimensions of the problem are $n = 50$, $p = 500$. The design matrix is Gaussian, i.e., $X_{i,j} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$ and the observations $y_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$. The regularization parameter is set to $\theta = 0.2 \times \theta_{\text{max}}$ where $\theta_{\text{max}} = \|X^\top y\|_\infty$.

Douglas–Rachford for the Sparse Inverse Covariance Selection. We consider covariance matrices of size $n \times n$ where $n = 50$ and $\theta = 0.1$. The matrix C is generated as $C = V^\top V$ where $V_{i,j} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$.

ADMM for Trend Filtering. We consider the cyclic 1D Total Variation $n = p = 75$ and $\lambda = 3.0$. Here $\theta \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$.

I Assets used

Our numerical experiments rely on:

- `numpy` [30], released under BSD-3 license.
- `matplotlib` [31], released under PSF license.
- `jax` [13], released under Apache-2.0 license.