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 $\operatorname{graph} D := \{(x, z) : x \in \mathbb{R}^n, z \in D(x)\}$.

**Definition 2 (Closed graphs)** *A set-valued mapping $D : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ has closed graph or is graph closed if* $\operatorname{graph} D$ *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 $\gamma$ 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, $\ell_1$ and $\ell_\infty$ 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 $F_1 \colon \mathbb{R}^p \mapsto \mathbb{R}^m$ and $F_2 \colon \mathbb{R}^m \mapsto \mathbb{R}^l$ be locally Lipschitz continuous mappings. Let $J_1 \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^{m\times p}$ be a conservative mapping for $F_1$ and $J_2 \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^{l\times m}$ 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*

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

**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 $\rho$. 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,

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

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 $\rho$. We deduce that $\operatorname{dist}(\mathcal{J}(\mathcal{X}), \mathcal{J}(\mathcal{Y})) \leqslant \rho\operatorname{dist}(\mathcal{X}, \mathcal{Y})$ using Lemma 2, that is the action of $\mathcal{J}$ on subsets of $p \times m$ matrices is $\rho$ 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 $m\in\mathbb{N}$

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

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

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

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 $\theta$, $J_F(x,\theta)\in\mathcal{C}_\rho$ (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 $J^{\text{pb}}_{\bar{x}}$)** *The mapping $J^{\text{pb}}_{\bar{x}}$ is nonempty valued, locally bounded and has a closed graph.*

**Proof :** The fact that $J^{\text{pb}}_{\bar{x}}$ 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, $J^{\text{pb}}_{\bar{x}}$ has a closed graph. □

### C.2 Proof of Theorem 2

**Proof :** Following Remark 3, we consider

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

a conservative Jacobian for $\bar{x}$ and $L_0 = J^{\text{imp}}_{\bar{x}}$. 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 $\theta$, $J_F$ 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 $J^{\text{pb}}_{\bar{x}}$, that is $l(\theta) \in J^{\text{pb}}_{\bar{x}}(\theta)$ for all $\theta \in \mathbb{R}^m$. Such a selection exist by [4, Theorem 18.20] because $J^{\text{pb}}_{\bar{x}}$ 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 $\theta$), so such a selection exists (Aliprantis Theorem 18.19). By Theorem 1, we have that $\text{dist}(L_k(\theta), J^{\text{pb}}_{\bar{x}}(\theta))$ 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 $\theta$. 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 $J^{\text{pb}}_{\bar{x}}$ 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 $J^{\text{pb}}_{\bar{x}}(\theta) = \left\{\frac{\partial \bar{x}}{\partial \theta}(\theta)\right\}$ 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 $k \in \mathbb{N}$,

$$J^{\text{pb}}_{\bar{x}}(\theta) = \left\{\frac{\partial \bar{x}}{\partial \theta}(\theta)\right\} \qquad\qquad L_k(\theta) = \left\{\frac{\partial \bar{x}}{\partial \theta}(\theta)\right\}.$$

Following the proof of Theorem 2 in Appendix C.2, $L_k$ converges to $J^{\text{pb}}_{\bar{x}}$ 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 $i \in 1, \dots k$

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

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

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

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 $i = 0, \dots, k-1$,

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

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

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

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 $\theta$

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

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 $\lambda$

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

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 $(x, \theta)$,

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

By Assumption 2, both $G_\alpha$ and $H_\alpha$ are 1-Lipschitz in $x$ for fixed $\theta$ 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 $\theta$. 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 $H_\alpha$. 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 $(x, \theta)$ a conservative Jacobian for $F_\alpha$,

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

Whenever $\nabla_x f$ is differentiable at $(x, \theta)$, 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 $G_\alpha$ is differentiable, since it is 1-Lipschitz in $x$ for fixed $\theta$ 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 $\theta$, $f$ is $\mu$-strongly convex. In this case, similarly as above the matrix $(I - \alpha A)$ in (23) has eigenvalue in $(-1, 1)$ for all $(x, \theta)$.
- Assume that for all $\theta$, $g$ is $\mu$-strongly convex. In this case, similarly as above the matrix $C$ in (23) has eigenvalue in $[0, 1/(1 + \alpha\mu)]$ for all $(x, \theta)$ [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 $H_\alpha$. 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

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

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

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

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 $J_{F_\alpha}$ for $F_\alpha$, such that for all $(x,\theta)$ 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 $\theta$ 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 $\hat{r} = \operatorname{prox}_g(2\hat{x} - \hat{y})$, this is also equivalent to

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

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

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

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.,

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

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

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

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

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

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 $\theta$), 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.