Differentiating Nonsmooth Solutions to Parametric Monotone Inclusion Problems

Jérôme Bolte, Edouard Pauwels, Antonio José Silveti-Falls

Differentiating Nonsmooth Solutions to Parametric Monotone Inclusion Problems

Jérôme Bolte* Edouard Pauwels† Antonio Silveti-Falls‡

Abstract. We leverage path differentiability and a recent result on nonsmooth implicit differentiation calculus to give sufficient conditions ensuring that the solution to a monotone inclusion problem will be path differentiable, with formulas for computing its generalized gradient. A direct consequence of our result is that these solutions happen to be differentiable almost everywhere. Our approach is fully compatible with automatic differentiation and comes with assumptions which are easy to check, roughly speaking: semialgebraicity and strong monotonicity. We illustrate the scope of our results by considering three fundamental composite problem settings: strongly convex problems, dual solutions to convex minimization problems and primal-dual solutions to min-max problems.

Key words. Maximal monotone operator, monotone inclusion, generalized equation, implicit differentiation, differentiating solutions, Clarke subdifferential, generalized gradient, conservative field.

AMS subject classifications. 49J40, 49J52, 49J53, 49K40, 65K15, 68T07

1 Introduction

Consider the following parametric maximal monotone inclusion problem

$$0 \in \mathcal{A}_\theta(x) + \mathcal{B}_\theta(x) \tag{1.1}$$

where θ is some parameter and, for each θ, $\mathcal{A}_\theta \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is maximal monotone (possibly set-valued) and $\mathcal{B}_\theta \colon \mathbb{R}^n \to \mathbb{R}^n$ is maximal monotone and Lipschitz continuous. For a fixed θ, a problem of this form is called a *generalized equation* [65] or *variational inequality*, and it models a wide range of optimization problems [8]. In fact, designing algorithms to find solution to maximal monotone inclusions is at the heart of convex optimization [7, 29, 30, 22, 21, 57, 53] and consequently has received a lot of attention in the last decades [5, 6, 20, 4].

Assuming that the solution $x^\star(\theta)$ is unique for each θ, our main objective in this paper is to investigate the regularity and differentiability of $x^\star$ with respect to θ as well as to develop calculus rules for computing a generalized derivative associated to it. In general, the solution $x^\star$ is not a function of θ because there can be several solutions for a given θ and, even under the assumption that $x^\star$ is unique, it is not guaranteed that $x^\star$ will be differentiable with respect to θ, therefore motivating further study. Understanding the regularity of

*Toulouse School of Economics, University of Toulouse
†IRIT, CNRS, Université Toulouse III Paul Sabatier. Institut Universitaire de France (IUF)
‡CVN, CentraleSupélec, Université Paris-Saclay

$x^\star$ as a function of θ has important applications in several areas: in deep learning for neural networks with implicit layers defined through monotone inclusions (e.g., monotone operator deep equilibrium networks [78], OptNet [3]), in machine learning (hyperparameter tuning [16], meta-learning [36], dataset distillation [13], adversarial examples [50]), signal processing ([29, 30, 35]), and general nonsmooth bilevel optimization [44, 76], without being exhaustive. For this reason, many other works have studied regularity properties of solutions to various forms of (1.1), using a myriad of different techniques that we will now discuss.

Variational analytic methods The study of solution mappings to parametric generalized equations can be traced back to variational analysis in the 1980s [65, 67]; [66] examined the Lipschitz continuity of the solution mapping when the single-valued monotone operator in the generalized equation is parametrized. This continued with results in [46, 47, 71] showing further the stability of the solution mapping but again avoiding parametrizing the nonsmooth/set-valued monotone operator. Another variational analytic approach is to use the notion of protodifferentiability developed in [68], which is used in [2] to analyze the stability of solutions to parametric monotone inclusions. This approach was extended in [75] to show that the directional differentiability of the solution map under the assumption that the Lipschitz continuous operator is strongly monotone. Both of these works consider generalized equations and allow for both the Lipschitz and set-valued monotone operators to be parametrized, the same as the current work. In the language of sensitivity analysis, all data in the problem (1.1) can be perturbed. A similar approach to [2] is used in [10] to analyze the Lipschitz constant of the solution mapping to the lasso problem as a function of the penalization parameter. Finally, we mention that similar methods have also been applied to study the differentiability of the prox operator in [62], which is a special case of a generalized equation in which the Lipschitz operator is identically 0.

Nonsmooth implicit differentiation As was already discussed, implicit differentiation is an important tool for characterizing properties of the solution to monotone inclusion problems, yet its application to nonsmooth problems remains a challenge. Specific cases involving the lasso and partial smoothness have been analyzed in [11] using the weak derivative and in [64, 74] using the Riemannian gradient. Other specific cases have been worked out for projections onto the cone of positive semidefinite matrices [52], or solutions of semidefinite programming problems [73], both of which use the Clarke implicit function theorem [28] to deduce Lipschitz continuity of the solution. From a computational perspective, a software library has been developed in [13] that can be used for automatic differentiation of implicitly defined functions. As we shall see, our approach is strongly based on the path differentiable implicit function theorem of [16] which comes with a calculus compatible with the Python library presented in [13].

Iterative differentiation/Unrolling Another growing field in which differentiation of solutions is fundamental is unrolling. In that case one wishes to find a solution of an optimization problem together with its derivative by differentiating some optimization algorithm with respect to external parameters. Pioneering works are [39, 9] and also [40]. In a machine learning context, research has been done for smooth algorithms setting in [60, 51, 54] and in the nonsmooth setting in [19] for path differentiability, [55] for partial smoothness, and [56] using a specific Bregman divergence. This is generally treated through ad hoc techniques, using for instance the smooth implicit function theorem approach. In [19], the approach is different and closer to ours as it uses the theory of conservative gradients. To understand the deep link this unrolling topic has with our present concerns, one needs to remind that iterates of algorithms are generally defined as solutions to a parametrized optimization problem. So unrolling offers a wide field of applications for solution's

differentiation techniques. Although we illustrate our results on general parametrized problems, let us emphasize that our results could also provide results for "unrolling", in the spirit of the iterative differentiation analysis of [19].

Each approach has its benefits, e.g., the variational analytic methods exploiting protodifferentiability are more adapted to giving information about the Lipschitz constant of the solution mapping than what we will propose. The most salient point of our contribution is that we are able to guarantee the path differentiability of the solution, which legitimates, in turn, the use of formal derivatives of the solution. In contrast to nearly all of these works and many others on this subject, this means our approach is compatible with modern differential tools like automatic differentiation and the backpropagation algorithm. This compatibility is achieved by way of a flexible calculus that allows all the usual operations of smooth calculus, in particular, the chain rule for differentiation. In the language of the optimization community, our differential results are qualification-free, and, in terms of differential regularity, everything will boil down to checking that the problem is semialgebraic (or definable).

Our approach and its advantages The general method we propose is to study the solution mapping $x^\star$ by first rewriting (1.1) as a locally Lipschitz fixed-point equation, using ideas from operator splitting methods for nonsmooth optimization [30]. For generalized equations of the form given in (1.1), we can use the resolvent $\mathcal{R}_{\mathcal{A}_\theta}$ to write the *forward-backward* map H, which a solution must be a fixed point of:

$$x^\star = H(x^\star, \theta) := \mathcal{R}_{\mathcal{A}_\theta}(x^\star - \mathcal{B}(x^\star)).$$

With this equation, we can continue by applying the nonsmooth implicit function theorem of [16] to deduce regularity properties and an expression for the generalized derivative (i.e., the conservative mapping) of the solution mapping $x^\star(\theta)$.

As has been discussed, the rise in popularity of modern automatic differentiation libraries [1, 23, 59] and their widespread use in machine learning calls for a flexible calculus at the crossroads of mathematics and computer science. For instance, almost the entire field of deep learning crucially relies on using the renowned backpropagation algorithm to do training. In spite of this, most prior work on this subject has either only considered the smooth case or has ignored non-differentiability issues. Thus a major advantage of the present approach, in contrast to other works, is to provide results that are broadly applicable (e.g., for nonsmooth solutions) but which are also compatible with automatic differentiation.

Besides compatibility with automatic differentiation, another advantage of our work is that the formula given by [16, Corollary 1] to compute an implicit conservative gradient is formally the same as the formula used in the smooth case. More specifically, a key feature of [16, Corollary 1] that our results will inherit is its coherence with the smooth implicit function theorem - that is to say, elements of the conservative Jacobian associated to the solution mapping $x^\star$ can be rigorously computed by formal differentiation just as in the smooth implicit function theorem.

Main results Our key technical result is an implicit function theorem for families of contractive Lipschitz continuous equations under path-differentiability assumptions. The simplicity of the contractivity assumption allows us to derive a wealth of regularity results for parametric strongly monotone problems. Our core result (Theorem 3.5) asserts that if $\mathcal{A}_\theta$ or $\mathcal{B}_\theta$ is strongly monotone, then $x^\star$ is path differentiable and there is a formula to compute a conservative Jacobian for it, based on the Clarke Jacobians of $\mathcal{R}_{\gamma\mathcal{A}_\theta}$ and $\mathcal{B}_\theta$. Let us insist on the fact that path differentiability easily follows from semi-algebraic or definable assumptions.

As a consequence, many fundamental parametrized optimization problems can be studied, we provide three important classes of examples. Theorem 4.2 deals with sum composite strongly convex optimization problems, which are ubiquitous in many fields from signal processing [30, 33, 77] to machine learning [12, 45]. In the framework of generalized convex duality, in the spirit of the Fenchel-Rockafellar theorem, we provide in Theorem 5.1 a regularity result for dual solutions to primal together with a calculus. In the min-max setting, under classical assumptions, we study the regularity of parametrize saddle points, this is Theorem 5.4. To be clear, the reach of our results extends beyond just these selected problems; our results represent a way to analyze solutions to any problem, which can be represented as a parametric monotone inclusion having this additive composite structure, encompassing a significant portion of the convex optimization problems in the literature.

Let us conclude by mentioning the fact the contractivity assumption which is behind our analysis in Lemma 3.4 is sharp in the sense that we are able to provide several counterexamples having apparently similar properties –as semialgebraic problems enjoying quadratic Łojasiewicz inequalities– which are not amenable to our optimization framework.

1.1 Organization of the Paper

In Section 2, we review the necessary background material and notation, mostly regarding convex analysis, conservative calculus, and the nonsmooth implicit function theorem for path differentiable functions. In Section 3, we develop results for path differentiability of the solution to parametric monotone inclusion problems, formally stating the monotone inclusion problem we consider and the assumptions needed to ensure path differentiability of the solution mapping. In Section 4, we turn to convex optimization and explore sufficient conditions in terms of properties of the objective function to ensure the solution to a convex optimization problem is path differentiable. In Section 5, we consider some general convex optimization problems and show how to apply the results of Section 3 and Section 4 to find expressions for implicit conservative Jacobians associated to the solution mappings. Finally, in Section 6, we conclude by noting some alternative formulations of the problem, its fixed point equations, and some other details that could have been chosen differently.

2 Background

Notation The set of real numbers will be written as $\mathbb{R}$ and the extended real numbers as $\bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$. We denote the set of p-times continuously differentiable functions on a given connected open subset $X \subset \mathbb{R}^n$ by $C^p(X)$ and denote the set of $C^1(X)$ functions whose gradient is Lipschitz continuous by $C^{1,1}(X)$. We will use Id_n to denote the identity matrix in $\mathbb{R}^{n\times n}$ and Id to denote the identity mapping. A set-valued map $\mathcal{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, is a function from $\mathbb{R}^n$ to subsets of $\mathbb{R}^m$ (including the empty set). The *graph* of a set-valued mapping $\mathcal{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ will be denoted $\operatorname{gph} \mathcal{A} := \{(x,u) \in \mathbb{R}^n \times \mathbb{R}^m \colon u \in \mathcal{A}(x)\}$. We denote the operator norm of a linear operator $K : \mathbb{R}^n \to \mathbb{R}^m$ as

$$\|K\|_{\mathrm{op}} := \sup_{v\in\mathbb{R}^n} \frac{\|Kv\|}{\|v\|}$$

and extend this to sets of linear operators $\mathcal{K} = \{K_\omega\}_{\omega\in\Omega}$ in the following way

$$\|\mathcal{K}\|_{\mathrm{op}} := \sup_{K\in\mathcal{K}} \sup_{v\in\mathbb{R}^n} \frac{\|Kv\|}{\|v\|}.$$

2.1 Convex Analysis

The following definitions and notations coming from convex analysis are well-known and can be found, for instance, in [7].

Definition 2.1 (Monotone Operator). A set-valued mapping $\mathcal{A}\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is called *monotone* if there exists $\alpha \geq 0$ such that for all $(x,u) \in \operatorname{gph}(\mathcal{A})$ and $(y,v) \in \operatorname{gph}(\mathcal{A})$,

$$\langle u - v, x - y \rangle \geq \alpha \|x - y\|^2 .$$

If $\alpha > 0$ then $\mathcal{A}$ is called α-*strongly monotone*.

A monotone operator $\mathcal{A}$ is said to be *maximal* if its graph is not contained in any other monotone operator. Recall that the *resolvent* of a maximal monotone operator $\mathcal{A}_\theta\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is the function $\mathcal{R}_{\mathcal{A}_\theta}\colon \mathbb{R}^n \to \mathbb{R}^n$ defined to be $\mathcal{R}_{\mathcal{A}_\theta} := (\mathrm{Id} + \mathcal{A}_\theta)^{-1}$. There is a special relationship between closed convex proper functions and maximal monotone operators, which we summarize in the next example.

Example 2.2. Let $f\colon \mathbb{R}^n \to \bar{\mathbb{R}}$ be a closed convex proper function, then the *subdifferential* of f, $\partial f(x) := \{u\colon \forall y \in \mathbb{R}^n, f(y) \geq f(x) + \langle u, y - x\rangle\}$, is a maximal monotone operator [70] and the resolvent $\mathcal{R}_{\partial f}$ is the *prox operator* [7, Example 23.3] given by $\operatorname{prox}_f(x) := \operatorname*{argmin}_{u \in \mathbb{R}^n} f(u) + \frac{1}{2}\|x - u\|^2$. Additionally, if f is α-strongly convex, then ∂f is α-strongly monotone [7, Example 22.3(iv)].

2.2 Conservative calculus and path differentiability

The following notions generalize the concept of differentiability to locally Lipschitz functions, from Clarke Jacobians to conservative Jacobians. In contrast with Clarke Jacobians, conservative Jacobians [18] offer a way to extend differentiation to locally Lipschitz functions in a way that is compatible with differential calculus.

Definition 2.3 (Clarke Jacobian). The *Clarke Jacobian* of a locally Lipschitz function $f\colon \mathbb{R}^n \to \mathbb{R}^m$ at a point x is defined to be

$$\operatorname{Jac}_f^c(x) := \operatorname{conv}\left\{ \lim_{k \to +\infty} \operatorname{Jac}_f(x_k) : x_k \in \operatorname{diff} f, \lim_{k \to +\infty} x_k = x \right\},$$

where $\operatorname{diff} f \subset \mathbb{R}^n$ is the set of full measure where f is differentiable in the classical sense.

The following lemma comes from [28, Proposition 2.1.2(a)], it is a nonsmooth generalization of the classical result that a β-Lipschitz continuous differentiable function has gradient bounded by β in norm.

Lemma 2.4 ([28]). *Let $U \subset \mathbb{R}^n$ be an open set and consider a Lipschitz continuous function $f\colon U \to \mathbb{R}^m$. Then f is Lipschitz continuous with constant β, if and only if, for all $x \in U$, $\left\|\operatorname{Jac}_f^c(x)\right\|_{\mathrm{op}} \leq \beta$.*

Definition 2.5 (Conservative Jacobian [18]). A *conservative Jacobian* for a locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}^m$ is a set-valued mapping $\mathcal{J}_f\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{m \times n}$ which is nonempty, locally bounded, graph closed, and satisfies, for any absolutely continuous curve $\gamma\colon [0,1] \to \mathbb{R}^n$,

$$\forall u \in \mathcal{J}_f(\gamma(t)), \quad \frac{d}{dt} f(\gamma(t)) = \langle u, \dot{\gamma}(t)\rangle \text{ for almost all } t \in [0,1]$$

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called *path differentiable* if it is locally Lipschitz and admits a conservative Jacobian $\mathcal{J}_f$. This is equivalent to the Clarke Jacobian Jac^c_f being conservative for f.

Path differentiable functions are ubiquitous among locally Lipschitz functions. The most prominent class of examples is that of semialgebraic functions. For an introduction to the subject of semialgebraic functions, we refer the interested reader to [32]. We simply recall here that a function $f\colon \mathbb{R}^n \to \mathbb{R}^n$ is said to be *semialgebraic* if $\mathrm{gph}\, f$ is a semialgebraic set, i.e., it can be written as the finite union and-or intersection of polynomial equations and inequalities. Let us also mention that all locally Lipschitz definable functions are path differentiable, see [15] and [18].

Given a path differentiable function $f : \mathbb{R}^{p+n} \to \mathbb{R}^n$, we will write $\mathrm{Jac}^c_{x,f}(\theta, x) := \{V\colon [U\ V] \in \mathrm{Jac}^c_f(\theta, x)\} \subset \mathbb{R}^{n\times n}$ and refer to this object as the Clarke Jacobian of f with respect to x (and the analog for when a conservative Jacobian has been specified). Similarly, we will write $\mathrm{Jac}^c_{\theta,f}(\theta, x) := \{U\colon [U\ V] \in \mathrm{Jac}^c_f(\theta, x)\} \subset \mathbb{R}^{n\times p}$ for the Clarke Jacobian of f with respect to θ. A very important but subtle point is that these sets are the projections of the joint Clarke Jacobians, which are possibly distinct from the sets given by fixing θ and computing the conservative Jacobian or the Clarke Jacobian with respect to x alone.

The following nonsmooth implicit differentiation theorem from [16, Corollary 1] is the main tool with which we can prove path differentiability and calculate elements of conservative Jacobians of solutions to monotone inclusions. Its main requirements are a path differentiable defining equation f and an invertibility condition on the elements of a conservative Jacobian associated with f.

Theorem 2.6 (Conservative implicit function theorem [16]). *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be path differentiable with conservative Jacobian $\mathcal{J}_f$. Let $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ be such that $f(\hat{x}, \hat{y}) = 0$. Assume that $\mathcal{J}_f(\hat{x}, \hat{y})$ is convex and that, for each $[U\ V] \in \mathcal{J}_f(\hat{x}, \hat{y})$, the matrix V is invertible. Then, there exists an open neighborhood $C \times D \subset \mathbb{R}^n \times \mathbb{R}^m$ of $(\hat{x}, \hat{y})$ and a path differentiable function $g : C \to D$ such that, for each $x \in C$,*

$$f(x, g(x)) = 0$$

and g admits a conservative Jacobian given, for each $x \in C$, by

$$\mathcal{J}_g\colon x \rightrightarrows \{-V^{-1}U : [U\ V] \in \mathcal{J}_f(x, g(x))\}.$$

In contrast to the accessibility of the formulas involved, the invertibility condition needed to apply Theorem 2.6 is more difficult to verify than its smooth counterpart due to the fact that the ordinary gradient is a singleton while the conservative gradient is set-valued. Indeed, for smooth implicit differentiation, it suffices to check the invertibility of a single matrix, while in the nonsmooth setting, one is tasked with showing the invertibility of *every* element of a set of matrices. Additionally, the invertibility of the matrix computed in the smooth setting can be checked during runtime while, in the nonsmooth setting, checking the matrix computed at runtime will not be sufficient to ensure that the invertibility condition is holding (see [16, Section 5]). For these reasons, it is imperative to have general sufficient conditions outlined which guarantee that the invertibility condition is satisfied, which is what we develop in the following sections.

Analogy with the Smooth Case One way to frame our work is as a study of how the calculus for solutions to smooth convex parametric optimization persists in nonsmooth convex parametric optimization thanks to conservative Jacobians. To explain this further, let us describe first a typical application of the smooth implicit function theorem to study convex parametric problems. Consider

$$\min_{x\in\mathbb{R}^n} f(\theta, x)$$

where $f\colon \mathbb{R}^p \times \mathbb{R}^n$ is continuously differentiable jointly in (θ, x), twice-differentiable in x, and convex in x for all $\theta \in \mathbb{R}^p$. To examine the existence and regularity of a solution mapping $x^\star$ as a function of θ, one can use the smooth implicit function theorem on the optimality condition $\nabla_x f(\theta, x^\star) = 0$ to get $\frac{\partial x^\star}{\partial \theta}(\theta) = -(\nabla_x^2 f(\theta, x^\star(\theta)))^{-1} \frac{\partial}{\partial \theta} \nabla_x f(\theta, x^\star(\theta))$. In addition to the differentiability assumptions we've made, the application of the smooth implicit function here requires the Hessian $\nabla_x^2 f(\theta, x^\star)$ to be invertible.

For a twice-differentiable convex function, the invertibility of the Hessian $\nabla_x^2 f(\theta, x)$ locally around $x^\star$ is equivalent to the strong convexity of the function f locally around $x^\star$, which is itself equivalent to a local quadratic growth condition $f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq \frac{\rho}{2} \|x - y\|^2$ for some $\rho > 0$, for all x, y in some neighborhood of $x^\star$. A byproduct of this paper is to investigate to what extent this trifold equivalency fails to hold when f is no longer assumed to be twice-differentiable. To this end, we are able to show positive and negative results in Section 4: strong convexity is sufficient to ensure the invertibility condition required by the nonsmooth implicit function theorem of [16, Corollary 1] holds, meanwhile a local quadratic growth condition is insufficient even when f is a semialgebraic function.

3 Solutions to monotone inclusions

3.1 Regularity assumptions and Lipschitz reformulations

A Lipschitz reformulation We begin this section by formally defining the parametric monotone inclusion problem we are considering (whose solution we seek to differentiate) and the assumptions we impose on it. When dealing with parametrized mappings like $\mathcal{A}\colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$, it will be convenient to use subscript notation $\mathcal{A}_\theta$ to denote the mapping corresponding to $\mathcal{A}(\theta, \cdot)\colon \mathbb{R}^n \to \mathbb{R}^n$ – this notation will be used frequently throughout the rest of the paper.

Assumption 3.1 (Path differentiability). Let $\Theta \subset \mathbb{R}^p$ be a nonempty connected open set, $\mathcal{A}\colon \Theta \times \mathbb{R}^n \to \mathbb{R}^n$, $\mathcal{B}\colon \Theta \times \mathbb{R}^n \to \mathbb{R}^n$, and $\gamma > 0$ be a stepsize. For all $\theta \in \Theta$, assume the following two conditions hold,

1. $\mathcal{A}(\theta, \cdot) : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is maximal monotone and $\mathcal{B}(\theta, \cdot) : \mathbb{R}^n \to \mathbb{R}^n$ is maximal monotone and locally Lipschitz. Furthermore the solution set to the following inclusion is nonempty

$$0 \in \mathcal{A}(\theta, \cdot) + \mathcal{B}(\theta, \cdot). \tag{$\mathcal{P}_{\text{mono}}$}$$

2. The resolvent $\mathcal{R}_{\gamma \mathcal{A}_\theta}$ and the map $\mathcal{B}$ are both locally Lipschitz and path differentiable, jointly in (θ, x), so that the function

$$H(\theta, x) := \mathcal{R}_{\gamma \mathcal{A}_\theta}(x - \gamma \mathcal{B}_\theta(x)) \tag{3.1}$$

is path differentiable jointly in (θ, x).

There are many different fixed-point reformulations of ($\mathcal{P}_{\text{mono}}$) one can choose from, each one inducing a function H_θ, which we discuss in Section 6. The Lipschitz mapping H_θ we have opted for is reminiscent of the forward-backward splitting algorithm [49]. It is general enough to cover a variety of monotone inclusions coming from both smooth and nonsmooth convex optimization problems. Similar to the choice of H_θ, the choice of the constant γ in Assumption 3.1 is also arbitrary provided γ is positive. Note also that the solution $x^\star$ does not depend on γ, although it can be defined as a fixed point of H_θ, which depends on γ. In each theorem, this constant will be finely tuned so as to obtain the most general regularity results possible. It is important to understand here that the regularity properties we will obtain in the following sections may depend on the reformulation in Assumption 3.1 that we have chosen. Whether it is the case or not is a matter for future research.

We define the *residual function* $\mathrm{res}\,:\mathbb{R}^{p+n}\to\mathbb{R}^n$ to be

$$\mathrm{res}\,(\theta,x):=x-H_\theta(x) \tag{$\mathcal{P}_{\mathrm{mono-bis}}$}$$

so that

$$\mathrm{res}\,(\theta,x^*(\theta))=0, \tag{3.2}$$

whenever the expression is well-defined.

An essential fact is that this equation will be automatically path differentiable if its constituents are semi-algebraic, or more generally definable [32, 15, 18]. Classically, this will also imply that the solution mapping is semialgebraic (as a set-valued mapping).

First-order properties of the residual equation Let us describe the key first-order objects that will ensure the existence of solution maps and allow us to establish their path-differentiability. We work under Assumption 3.1.

Set $T\colon \Theta\times\mathbb{R}^n\to\mathbb{R}^n$ to be $T(\theta,x):=\mathcal{R}_{\gamma\mathcal{A}_\theta}(x)$ and consider the two following conservative Jacobians relative to the path differentiable mapping H:

$$\begin{aligned}\mathcal{J}_H(\theta,x) &= \{\begin{bmatrix}U-\gamma VW & V(\mathrm{Id}_n-\gamma Z)\end{bmatrix} : [U\ V]\in \mathrm{Jac}^c_T(\theta,x-\gamma\mathcal{B}_\theta(x)),[W\ Z]\in\mathrm{Jac}^c_{\mathcal{B}}(\theta,x)\}.\\ \mathcal{J}_{x,H}(\theta,x) &= \mathrm{Jac}^c_{x,T}(\theta,x-\gamma\mathcal{B}_\theta(x))\times(\mathrm{Id}_n-\gamma\mathrm{Jac}^c_{x,\mathcal{B}}(\theta,x)).\end{aligned} \tag{3.3}$$

The first set valued map $\mathcal{J}_H$ given in (3.3) is simply obtained by a formal application of the rules of differential calculus to the composite structure of H in Assumption 3.1 using Clarke Jacobians instead of classical Jacobians, while the second is a partial derivative version obtained by mere projection.

Indeed, define $S\colon\Theta\times\mathbb{R}^n\to\Theta\times\mathbb{R}^n$ to be $S(\theta,x):=(\theta,x-\gamma\mathcal{B}(\theta,x))$ so that $T(S(\theta,x))=H(\theta,x)$. One can check that $\mathcal{J}_H$ is the following product of Clarke Jacobians for $(\theta,x)\in\Theta\times\mathbb{R}^n$

$$\begin{aligned}\mathcal{J}_H(\theta,x) &= \mathrm{Jac}^c_T(S(\theta,x))\times\mathrm{Jac}^c_S(\theta,x)\\ &= \{\begin{bmatrix}U & V\end{bmatrix}\times\begin{bmatrix}\mathrm{Id}_p & 0\\ -\gamma W & \mathrm{Id}_n-\gamma Z\end{bmatrix} : [U\ V]\in\mathrm{Jac}^c_T(\theta,x-\gamma\mathcal{B}_\theta(x)),[W\ Z]\in\mathrm{Jac}^c_{\mathcal{B}}(\theta,x)\}\\ &= \{\begin{bmatrix}U-\gamma VW & V(\mathrm{Id}_n-\gamma Z)\end{bmatrix} : [U\ V]\in\mathrm{Jac}^c_T(\theta,x-\gamma\mathcal{B}_\theta(x)),[W\ Z]\in\mathrm{Jac}^c_{\mathcal{B}}(\theta,x)\}\end{aligned}$$

which is a conservative Jacobian for H. Consequently we have the following conservative Jacobian with respect to x for H, for each fixed $(\theta,x)\in\Theta\times\mathbb{R}^n$,

$$\begin{aligned}\mathcal{J}_{x,H}(\theta,x) &= \{V(\mathrm{Id}_n-\gamma Z)\colon [U\ V]\in\mathrm{Jac}^c_T(\theta,x-\gamma\mathcal{B}_\theta(x)),[W\ Z]\in\mathrm{Jac}^c_{\mathcal{B}}(\theta,x)\}\\ &= \mathrm{Jac}^c_{x,T}(\theta,x-\gamma\mathcal{B}_\theta(x))\times(\mathrm{Id}_n-\gamma\mathrm{Jac}^c_{x,\mathcal{B}}(\theta,x)),\end{aligned}$$

which concludes our explanation.

Let us now define a conservative Jacobian for the residual function res from ($\mathcal{P}_{\mathrm{mono-bis}}$) through:

$$\mathcal{J}_{\mathrm{res}}\,(\theta,x)=\{\begin{bmatrix}\gamma VW-U & \mathrm{Id}_n-V(\mathrm{Id}_n-\gamma Z)\end{bmatrix} : [U\ V]\in\mathrm{Jac}^c_T(\theta,x-\gamma\mathcal{B}_\theta(x)),[W\ Z]\in\mathrm{Jac}^c_{\mathcal{B}}(\theta,x)\},$$

for each $(\theta,x)\in\Theta\times\mathbb{R}^n$.

Remark 3.2. The choice of $\mathcal{J}_{x,H}(\theta, x)$ in (3.3) corresponds to the set-valued object computed by applying formal differentiation, i.e., the chain rule, to H as a composition of functions. Additionally, the particular form of $\mathcal{J}_{x,H}(\theta, x)$ we consider above allows one to control $\|\mathcal{J}_{x,H}(\theta, x)\|_{\mathrm{op}}$ when one of the monotone operators $\mathcal{A}_\theta$ or $\mathcal{B}_\theta$ is strongly monotone, as we will expose later. In terms of the residual function res from ($\mathcal{P}_{\mathrm{mono-bis}}$), the choice made in (3.3) induces a particular conservative Jacobian $\mathcal{J}_{\mathrm{res}}$ and thus also a particular conservative Jacobian with respect to x,

$$\begin{aligned}\mathcal{J}_{x,\mathrm{res}}(\theta, x) &= \{\mathrm{Id}_n - V(\mathrm{Id}_n - \gamma Z) \colon [U\ V] \in \mathrm{Jac}^c_T(\theta, x - \gamma\mathcal{B}_\theta(x)), [W\ Z] \in \mathrm{Jac}^c_{\mathcal{B}}(\theta, x)\} \\ &= \mathrm{Id}_n - \mathrm{Jac}^c_{x,T}(\theta, x - \gamma\mathcal{B}_\theta(x)) \times (\mathrm{Id}_n - \gamma \mathrm{Jac}^c_{x,\mathcal{B}}(\theta, x)).\end{aligned}$$

3.2 Contractivity implies the path differentiatiability of solutions

With this particular choice of conservative Jacobians for H and res given by (3.3), and under the additional assumption of strong monotonicity, we show below that a simple contractivity condition holds.

Definition 3.3 (Contractivity of residual equations). Under Assumption 3.1, we shall say that the residual equation res $= 0$ is *contractive*, or has the *contractivity property*, if for each $\theta \in \Theta$, there is a solution $x^\star(\theta)$ to ($\mathcal{P}_{\mathrm{mono}}$) with

$$\|\mathcal{J}_{x,H}(\theta, x^\star(\theta))\|_{\mathrm{op}} < 1$$

where we recall that $H = \mathrm{Id} - \mathrm{res}$.

Contractivity in Definition 3.3 is key to applying nonsmooth implicit path differentiation, first because it actually warrants the well-posedness and the single-valuedness of the solution mapping ([7, Proposition 26.1], see as well [43, Theorem 11]), and also because it is closely related to the invertiblity condition required in the conservative implicit function theorem (Theorem 2.6).

The next lemma shows that contractivity is key to our approach to solutions of monotone inclusions.

Lemma 3.4 (Path differentiability of the solution map). *Under* Assumption 3.1 *and assuming* res *has the contractivity property, then $x^\star$ is unique, path differentiable on Θ, and admits a conservative Jacobian of the form*

$$\mathcal{J}_{x^\star} : \theta \rightrightarrows$$
$$\{(\mathrm{Id}_n - V(\mathrm{Id}_n - \gamma Z))^{-1}(U - \gamma VW) : [U\ V] \in \mathrm{Jac}^c_T(\theta, x^\star(\theta) - \gamma\mathcal{B}_\theta(x^\star(\theta))), [W\ Z] \in \mathrm{Jac}^c_{\mathcal{B}}(\theta, x^\star(\theta))\}.$$

Proof. We begin with the uniqueness of $x^\star$ for each $\theta \in \Theta$. By convexity of the operator norm, it holds

$$\|\mathrm{conv}\,(\mathcal{J}_{x,H}(\theta, x^\star(\theta)))\|_{\mathrm{op}} \leq \|\mathcal{J}_{x,H}(\theta, x^\star(\theta))\|_{\mathrm{op}} < 1.$$

However, $\mathrm{Jac}^c_{x,H}(\theta, x^\star(\theta)) \subset \mathrm{conv}\,(\mathcal{J}_{x,H}(\theta, x^\star(\theta)))$ and so $\left\|\mathrm{Jac}^c_{x,H}(\theta, x^\star(\theta))\right\|_{\mathrm{op}} < 1$. From this we conclude that H is locally a strict contraction around $(\theta, x^\star(\theta))$ and thus the solution $x^\star(\theta)$ is unique.

We will apply Theorem 2.6 to res using the pointwise convex hull $(\theta, x) \rightrightarrows \mathrm{conv}(\mathcal{J}_{\mathrm{res}}(\theta, x))$ as a conservative Jacobian for res. Note that it is indeed conservative since $\mathcal{J}_{\mathrm{res}}$ is and pointwise convex hulls preserves Definition 2.5. We will use the shorthand notation $\mathrm{conv}(\mathcal{J}_{x,\mathrm{res}})$ to denote for each fixed $\theta \in \Theta$ the set valued map $(\theta, x) \rightrightarrows \mathrm{conv}(\mathcal{J}_{\mathrm{res}}(\theta, x))$, and similarly for H. Note that $\mathrm{conv}(\mathcal{J}_{x,\mathrm{res}}) = \mathrm{Id}_n - \mathrm{conv}(\mathcal{J}_{x,H})$.

Fix $\theta \in \Theta$. Let us show that the contractivity condition entails that every element of $\mathrm{conv}(\mathcal{J}_{x,\mathrm{res}}(\theta, x^\star(\theta)))$ is invertible. Indeed, set $\rho = \|\mathrm{conv}(\mathcal{J}_{x,H}(\theta, x^\star(\theta)))\|_{\mathrm{op}} = \|\mathcal{J}_{x,H}(\theta, x^\star(\theta))\|_{\mathrm{op}} < 1$ (use triangle inequality), so that for any $M_{\mathrm{res}} \in \mathrm{conv}(\mathcal{J}_{x,\mathrm{res}}(\theta, x^\star(\theta)))$, there is $M_H \in \mathrm{conv}(\mathcal{J}_{x,H}(\theta, x^\star(\theta)))$ such that $M_{\mathrm{res}} = \mathrm{Id}_n - M_H$, and for all $v \in \mathbb{R}^n$, we have

$$\begin{aligned}\|M_{\mathrm{res}}\, v\| &= \|(\mathrm{Id}_n - M_H)v\| \\ &\geq \|v\| - \|M_H v\| \\ &\geq \|v\|\,(1-\rho)\end{aligned}$$

which shows that M_{res} is invertible because $\rho < 1$. Since res is path differentiable and all of the elements of $\mathcal{J}_{x,\mathrm{res}}(\theta, x^\star(\theta))$ are invertible for each $\theta \in \Theta$, the conditions to apply Theorem 2.6 to the equation $\mathrm{res}\,(\theta, x^\star(\theta)) = 0$ hold, and thus $x^\star$ is path differentiable on Θ. The formula for the conservative Jacobian follows from Theorem 2.6 because it defines a graph closed and locally bounded set valued map which is a subset of the set valued map obtained by applying Theorem 2.6 to $\mathrm{conv}(\mathcal{J}_{\mathrm{res}})$ which satisfies the chain rule of Definition 2.5. □

The expression for the conservative Jacobian of $x^\star$ given in Lemma 3.4 can be more compactly expressed in terms of the conservative Jacobian of H defined in (3.3), for each $\theta \in \Theta$,

$$\mathcal{J}_{x^\star} : \theta \rightrightarrows \{(\mathrm{Id}_n - V)^{-1}\, U : [U\ V] \in \mathcal{J}_H(\theta, x^\star(\theta))\}.$$

3.3 Strongly monotone inclusions have path differentiable solutions

The following theorem is related to [7, Prop. 26.16], which provides sufficient conditions for linear convergence of the forward-backward algorithm applied to finding a zero of the sum of two maximally monotone operators $\mathcal{A}$ and $\mathcal{B}$. It is however important to observe that linear convergence is not enough to reach the same conclusions (see Section 4.2).

Theorem 3.5 (Path differentiability: strongly monotone case). *Under* Assumption 3.1, *consider* ($\mathcal{P}_{\mathrm{mono}}$) *and, for each $\theta \in \Theta$, assume that $\mathcal{B}_\theta$ is β-Lipschitz continuous and that either $\mathcal{A}_\theta$ or $\mathcal{B}_\theta$ is α-strongly monotone, for some $\alpha, \beta > 0$, uniformly in θ.*

Then, for $\gamma \in \left(0, \frac{2\alpha}{(\alpha+\beta)^2}\right)$, the residual map res *of* ($\mathcal{P}_{\mathrm{mono}}$) *is contractive, i.e., the inequality in* Definition 3.3 *holds. Furthermore, $x^\star$ is unique and path differentiable on Θ with a conservative Jacobian given for each $\theta \in \Theta$ by*

$$\begin{aligned}&\mathcal{J}_{x^\star} : \theta \rightrightarrows \\ &\{(\mathrm{Id}_n - V(\mathrm{Id}_n - \gamma Z))^{-1}\,(U - \gamma VW) : [U\ V] \in \mathrm{Jac}^c_T(\theta, x^\star(\theta) - \gamma\mathcal{B}_\theta(x^\star(\theta))), [W\ Z] \in \mathrm{Jac}^c_{\mathcal{B}}(\theta, x^\star(\theta))\}.\end{aligned}$$

Proof. Under Assumption 3.1 with the conservative Jacobians given in (3.3), the forward-backward mapping H is path differentiable on $\Theta \times \mathbb{R}^n$ with a conservative Jacobian with respect to x given by

$$\mathcal{J}_{x,H}(\theta, x) = \mathrm{Jac}^c_{x,T}(\theta, x - \gamma\mathcal{B}_\theta(x)) \times (\mathrm{Id}_n - \gamma\mathrm{Jac}^c_{x,\mathcal{B}}(\theta, x)).$$

We take an arbitrary $\theta \in \Theta$ and divide the proof into cases depending on whether $\mathcal{A}_\theta$ or $\mathcal{B}_\theta$ is α-strongly monotone.

Assume that $\mathcal{B}_\theta$ is α-strongly monotone with $\alpha \leq \beta$. Since $\alpha > 0$ and $0 < \gamma < \frac{2\alpha}{(\alpha+\beta)^2} < \frac{2\alpha}{\beta^2}$, it holds that $\gamma(2\alpha - \gamma\beta^2) > 0$, and furthermore $\gamma(2\alpha - \gamma\beta) \leq \gamma 2\alpha < \frac{4\alpha^2}{(\alpha+\beta)^2} \leq 1$. Setting $\tau = \sqrt{1 - \gamma(2\alpha - \gamma\beta^2)}$, we

have that $\mathcal{R}_{\gamma\mathcal{A}_\theta}$ is nonexpansive [7, Proposition 23.8] and $\mathrm{Id}-\gamma\mathcal{B}_\theta$ is τ-Lipschitz continuous with $0 \leq \tau < 1$. Thus, by applying Lemma 2.4 for each $(\theta, x) \in \Theta \times \mathbb{R}^n$,

$$\|\mathcal{J}_{x,H}(x)\|_{\mathrm{op}} \leq \left\|\mathrm{Jac}^c_{x,T}(\theta, x - \gamma\mathcal{B}_\theta(x))\right\|_{\mathrm{op}} \left\|(\mathrm{Id}_n - \gamma\mathrm{Jac}^c_{x,\mathcal{B}}(\theta, x))\right\|_{\mathrm{op}} \leq \tau < 1.$$

Now we consider case where $\mathcal{A}_\theta$ is α-strongly monotone. We have by [7, Proposition 23.13] that the resolvent $\mathcal{R}_{\gamma\mathcal{A}_\theta}$ is Lipschitz continuous with constant $\frac{1}{1+\gamma\alpha}$ and the mapping $\mathrm{Id} - \gamma\mathcal{B}_\theta$ is $\sqrt{1+\gamma^2\beta^2}$-Lipschitz continuous by Lemma A.1, giving for each $(\theta, x) \in \Theta \times \mathbb{R}^n$,

$$\|\mathcal{J}_{x,H}(x)\|_{\mathrm{op}} = \left\|\mathrm{Jac}^c_{x,T}(\theta, x - \gamma\mathcal{B}_\theta(x))\right\|_{\mathrm{op}} \left\|(\mathrm{Id}_n - \gamma\mathrm{Jac}^c_{x,\mathcal{B}}(\theta, x))\right\|_{\mathrm{op}} \leq \frac{\sqrt{1+\gamma^2\beta^2}}{1+\gamma\alpha}.$$

Since $\gamma \in \left(0, \frac{2\alpha}{(\alpha+\beta)^2}\right)$, it holds

$$1 + \gamma^2\beta^2 < 1 + \gamma(\gamma(\alpha+\beta)^2) < 1 + 2\alpha\gamma < 1 + \gamma 2\alpha + \gamma^2\alpha^2 = (1+\gamma\alpha)^2,$$

so that $\frac{\sqrt{1+\gamma^2\beta^2}}{1+\gamma\alpha} < 1$. Putting everything together we find, for each $(\theta, x) \in \Theta \times \mathbb{R}^n$,

$$\|\mathcal{J}_{x,H_\theta}(x)\|_{\mathrm{op}} = \left\|\mathrm{Jac}^c_{x,T}(\theta, x - \gamma\mathcal{B}_\theta(x))\right\|_{\mathrm{op}} \|(\mathrm{Id}_n - \gamma\mathrm{Jac}^c_{\mathcal{B}}(\theta, x))\|_{\mathrm{op}} \leq \frac{\sqrt{1+\gamma^2\beta^2}}{1+\gamma\alpha} < 1.$$

We have established that $\|\mathcal{J}_{H_\theta}(x)\|_{\mathrm{op}} < 1$ for all $(\theta, x) \in \Theta \times \mathbb{R}^n$ in both cases of the theorem, and we have contractivity. By Lemma 3.4, $x^\star$ is, therefore, path differentiable on Θ and the desired formula for the conservative Jacobian follows. □

Remark 3.6 (On the constant γ). The restriction on the values that γ can take in the different cases of Theorem 3.5 can be relaxed if more information about the operators $\mathcal{A}_\theta$ and $\mathcal{B}_\theta$ is specified. For instance, if $\mathcal{B}_\theta$ is β-cocoercive rather than β-Lipschitz and $\mathcal{A}_\theta$ is α-strongly monotone, then H_θ is a contraction for any $\gamma \in (0, 2\beta)$. It is important to notice that the choice of γ for implicit differentiation need not match the γ chosen for solving the problem (indeed, the algorithm to solve the problem and the fixed point equation for optimality need not match to begin with).

4 Path differentiation of solutions to convex optimization problems

Let $\Theta \subset \mathbb{R}^p$ be a connected open set and consider, for each $\theta \in \Theta$, the parametric optimization problem of finding a minimizer

$$x^\star := \operatorname*{argmin}_{x \in \mathbb{R}^n} f(\theta, x) + g(\theta, x) \tag{$\mathcal{P}_{\mathrm{opt}}$}$$

where $f_\theta := f(\theta, \cdot) \in C^{1,1}(\mathbb{R}^n)$ is convex and $g_\theta := g(\theta, \cdot)$ is a closed convex proper function from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$. It is well known [7, Theorem 26.2] that this problem is equivalent to finding a zero $x^\star$ of the sum of two monotone operators given by the subdifferentials of the functions,

$$0 \in \nabla f_\theta(x^\star) + \partial g_\theta(x^\star).$$

In this way, the problem of differentiating a solution of ($\mathcal{P}_{\mathrm{opt}}$) is equivalent to the problem of the previous section - differentiating a solution to a monotone inclusion ($\mathcal{P}_{\mathrm{mono}}$). This equivalence motivates the following assumptions on ($\mathcal{P}_{\mathrm{opt}}$), which are analogous to Assumption 3.1 with the conservative Jacobians defined in (3.3) for the case where the monotone operators $\mathcal{A}_\theta$ and $\mathcal{B}_\theta$ are subdifferentials of closed convex proper functions.

Assumption 4.1. Let Θ be a connected open set and let $\gamma > 0$. For all $\theta \in \Theta$, let $f_\theta := f(\theta, \cdot) \in C^{1,1}(\mathbb{R}^n)$ and $g_\theta := g(\theta, \cdot)$ be closed convex proper functions from $\mathbb{R}^n$ to $\bar{\mathbb{R}}$ and assume that the prox operator $\operatorname{prox}_{\gamma g_\theta} \colon \Theta \times \mathbb{R}^n \to \mathbb{R}^n$ and the gradient $\nabla f_\theta \colon \Theta \times \mathbb{R}^n \to \mathbb{R}^n$ are both path differentiable, jointly in (θ, x).

A sufficient condition guaranteeing the path differentiability in Assumption 4.1 holds is to assume that f and g are semialgebraic functions. Under Assumption 4.1, one can consider $\mathcal{A}_\theta = \partial g_\theta$ and $\mathcal{B}_\theta = \nabla f_\theta$ so that Assumption 3.1 is met and H is the forward-backward algorithm applied to ($\mathcal{P}_{\text{opt}}$). Using the conservative Jacobians defined in (3.3), we have in this case for all $(\theta, x) \in \Theta \times \mathbb{R}^n$

$$\begin{aligned}\mathcal{J}_H(\theta, x) &= \{\begin{bmatrix} U - \gamma VW & V(\mathrm{Id}_n - \gamma Z)\end{bmatrix} : [U\ V] \in \operatorname{Jac}^c_{\operatorname{prox}_{\gamma g_\theta}}(\theta, x - \gamma \nabla f_\theta(x)), [W\ Z] \in \operatorname{Jac}^c_{\nabla f_\theta}(\theta, x)\} \\ \mathcal{J}_{x,H}(\theta, x) &= \operatorname{Jac}^c_{x, \operatorname{prox}_{\gamma g_\theta}}(\theta, x - \gamma \nabla f_\theta(x)) \times (\mathrm{Id}_n - \gamma \operatorname{Jac}^c_{x, \nabla f_\theta}(\theta, x)). \end{aligned} \tag{4.1}$$

For the moment, we do not explicitly assume that the solution $x^\star$ to ($\mathcal{P}_{\text{opt}}$) is unique for each $\theta \in \Theta$; the results in later sections will make stronger assumptions that imply the uniqueness of $x^\star$ as a byproduct. We shall also provide assumptions on g_θ and f_θ that will ensure the invertibility condition of Theorem 2.6 holds at the solution $x^\star(\theta)$.

4.1 Solutions of strongly convex problems

Recall that the subdifferential of a strongly convex function is strongly monotone [7, Example 22.4]. As a consequence of Theorem 3.5 for strong monotonicity, we can then formulate the following analgous result for ($\mathcal{P}_{\text{opt}}$) with strong convexity of f_θ or g_θ.

Theorem 4.2 (Path differentiability: strongly convex case). *Let* Assumption 4.1 *hold with the conservative Jacobians given in* (4.1) *and consider* ($\mathcal{P}_{\text{opt}}$) *for $\theta \in \Theta$. Denote $\beta > 0$ a Lipschitz constant of ∇f_θ which is assumed to be uniform in θ. Assume that either f_θ or g_θ is α-strongly convex with $\alpha > 0$ which is also assumed to be uniform in θ. Then, for $\gamma = \frac{\alpha}{(\alpha+\beta)^2}$, the solution $x^\star(\theta)$ is unique for each $\theta \in \Theta$ and path differentiable on Θ with a conservative Jacobian given by*

$$\mathcal{J}_{x^\star} \colon \theta \rightrightarrows \{(\mathrm{Id}_n - V(\mathrm{Id}_n - \gamma Z))^{-1}(U - \gamma VW)\}$$

where $[U\ V]$ range in $\operatorname{Jac}^c_T(\theta, x^\star(\theta) - \gamma \nabla f_\theta(x^\star(\theta)))$ and $[W\ Z]$ in $\operatorname{Jac}^c_{\nabla f_\theta}(\theta, x^\star(\theta))$.

Proof. Due to Assumption 4.1, the function res (θ, x) is path differentiable on $\Theta \times \mathbb{R}^n$; fix an arbitrary $\theta \in \Theta$. Since $f_\theta \colon \mathbb{R}^n \to \mathbb{R}$ and $g_\theta \colon \mathbb{R}^n \to \bar{\mathbb{R}}$ are closed convex proper functions, ∇f_θ and ∂g_θ are maximal monotone operators [7, Example 22.4]. Furthermore, since one of f_θ or g_θ is α-strongly convex, one of the operators ∇f_θ or ∂g_θ is α-strongly monotone and Theorem 3.5 can be applied with $\mathcal{A}_\theta = \partial g_\theta$ and $\mathcal{B}_\theta = \nabla f_\theta$, yielding the path differentiability of $x^\star$ and the formula for its conservative Jacobian on Θ. □

The choice of γ can be relaxed to be any value in $\left(0, \frac{2\alpha}{(\alpha+\beta)^2}\right)$ without issue, we have simply taken $\frac{\alpha}{(\alpha+\beta)^2}$ for convenience. In contrast to some prior work, e.g., [34], we allow for both f_θ and g_θ to be parametrized functions, rather than just one or the other.

4.2 Beyond strong convexity?

Strong convexity may be generalized by means of quadratic growth conditions, such as global error bounds [58] or equivalently KL inequality [17]. On the other hand, quadratic Łojasiewicz inequality, quadratic error bound, turn out to be equivalent to linear convergence of forward-backward iterations under mild conditions; one may consult, for instance, [37]. Since contractivity obviously implies linear convergence, it is tempting to think that it could be somehow inserted into the equivalence chain. This would be a natural path beyond strong convexity assumptions.

This calls, for instance, for the following question: Does the contractivity of res always hold if f and g are such that H is linearly convergent to a unique fixed point? An element of motivation is that in the smooth case, when f is C^2 and $g = 0$, contractivity is indeed equivalent to (4.2) as discussed in Section 2, so that the questions relates to extension of this equivalence to the nonsmooth setting.

Linear convergence of the forward-backward mapping can simply be formulated as follows: for a fixed $\theta \in \Theta$ there exists $\rho \in (0, 1)$, such that, for all x,

$$\|H(\theta, x) - x^\star(\theta)\| \leq \rho \|x - x^\star(\theta)\| . \tag{4.2}$$

We provide below two examples having this property while being non strongly convex contradicting contractivity of res . The first one has $C^{1,1}$ objective ($g = 0$), while the second is nonsmooth Proposition 4.4 ($f = 0$). This answers negatively to the above question.

Let us start with the differentiable case for which H reduces to a gradient step.

Proposition 4.3 (Linear convergence does not imply contractivity I). *There exists a convex semialgebraic function $h \in C^{1,1}(\mathbb{R}^2)$ with ∇h 1-Lipschitz and $\|x - \nabla h(x)\| \leq \rho\|x\|$ for all x for some $\rho \in (0, 1)$, such that h is not strongly convex locally around 0 and $\mathrm{Jac}^c_{\nabla h}(0)$ contains singular matrices.*

Proof. Let $Q \subset \mathbb{R}^2$ be a closed convex set with $0 \in \mathrm{int}(Q)$ and a smooth boundary so that there is a differentiable outer pointing unit normal vector $\hat{n} \in C^1(\mathrm{bd}(Q))$. We consider the gauge function Ψ_Q associated to Q

$$\Psi_Q(x) = \inf\{\lambda > 0 : x \in \lambda Q\}.$$

The gauge function Ψ_Q [7, Example 8.36] is the unique positively homogeneous function such that the sublevel set of level 1 is Q [7, Corollary 14.13]. Furthermore, the sublevel sets of Ψ_Q of level $\lambda \geq 0$ are equal to λQ.

For $x \in \mathbb{R}^2$, we extend $\hat{n}(x)$ to be the outer pointing normal vector to the set $\Psi_Q(x)Q$ at x, it defines a C^1 function on $\mathbb{R}^2$. The gradient of Ψ_Q for $x \neq 0$ has to be of the form $\alpha(x)\hat{n}(x)$ for a positive function α. By homogeneity, we have for small t

$$\frac{\Psi_Q\left(x\left(1 + \frac{t}{\Psi_Q(x)}\right)\right) - \Psi_Q(x)}{t} = 1 = \left\langle \frac{x}{\Psi_Q(x)}, \nabla\Psi_Q(x) \right\rangle = \alpha(x) \left\langle \frac{x}{\Psi_Q(x)}, \hat{n}(x) \right\rangle$$

from which we obtain

$$\nabla\Psi_Q(x) = \frac{\Psi_Q(x)}{\langle x, \hat{n}(x)\rangle}\hat{n}(x).$$

Since $\hat{n}$ is homogeneous of order zero, so is $\nabla\Psi_Q$. We have for $x \neq 0$

$$\mathrm{Jac}_{\nabla\Psi_Q(x)} = \mathrm{Jac}\left(x \mapsto \nabla\Psi_Q\left(\frac{x}{\Psi_Q(x)}\right)\right) = \mathrm{Jac}_{\nabla\Psi_Q(x/\Psi_Q(x))}\left(\frac{\mathrm{Id}_n}{\Psi_Q(x)} - \frac{x\nabla\Psi_Q(x)^T}{\Psi_Q(x)^2}\right).$$

Now set $h(x) = \Psi_Q(x)^2/2$, which is convex and C^1 since

$$\nabla h(x) = \Psi_Q(x)\nabla\Psi_Q(x)$$

is a continuous function. We have for $x \neq 0$

$$\begin{aligned}\mathrm{Jac}_{\nabla h(x)} &= \nabla\Psi_Q(x)\nabla\Psi_Q(x)^T + \Psi_Q(x)\mathrm{Jac}_{\nabla\Psi_Q(x/\Psi_Q(x))}\left(\frac{\mathrm{Id}_n}{\Psi_Q(x)} - \frac{x\nabla\Psi_Q(x)^T}{\Psi_Q(x)^2}\right)\\ &= \nabla\Psi_Q(x)\nabla\Psi_Q(x)^T + \mathrm{Jac}_{\nabla\Psi_Q(x/\Psi_Q(x))}\left(\mathrm{Id}_n - \frac{x\nabla\Psi_Q(x)^T}{\Psi_Q(x)}\right).\end{aligned}$$

This expression remains bounded which shows that ∇h is Lipschitz and we may assume by rescaling that its Lipschitz constant is 1. Set for all x, $x^+ = x - \nabla h(x)$, we have using standard arguments in the analysis of gradient descent on h, whose global minimum is the origin, that

$$2h(x^+) + \left\|x^+\right\|^2 \leq \|x\|^2.$$

We have that h is positively homogenoeous of degree 2 so that

$$h(x^+) = \left\|x^+\right\|^2 \frac{h(x^+)}{\left\|x^+\right\|^2} = \left\|x^+\right\|^2 h\left(\frac{x^+}{\left\|x^+\right\|}\right) \geq \left\|x^+\right\|^2 \min_{\|y\|=1} h(y),$$

where the minimum is attained and is positive, call it $c > 0$. All in all, we have

$$\|x - \nabla h(x)\| \leq \frac{1}{\sqrt{1+2c}}\|x\|,$$

so that the constructed function complies with hypotheses of the Lemma, independently of Q.

By definition of the gauge function, the sublevel sets of Ψ_Q are of the form λQ for $\lambda \in \mathbb{R}$. If h was strongly convex locally around 0, one would have that its sublevel sets are also strongly convex (positively curved). This is not the case, for example if Q is a square with smoothed corners. This shows that h is not necessarily locally strongly convex. To ensure h is semialgebraic, it suffices to take Q a semialgebraic square with smoothed corners.

We conclude with the following implication: if $\mathrm{Jac}^c_{\nabla h}(0)$ contains only nonsingular elements then h is strongly convex locally around 0. Indeed in this case $\mathrm{Jac}^c_{\nabla h}(x)$ contains only positive definite elements for all x in a convex compact neighborhood of 0, set $\lambda > 0$ a lower bound on the minimum eigenvalue in this neighborhood (which exists by graph closedness and continuity of the smallest eigenvalue), we have by (Aumann) integration in Definition 2.5 using conservativity of $\mathrm{Jac}^c_{\nabla h}$, for all x, y in this neighborhood,

$$\begin{aligned}\langle \nabla h(x) - \nabla h(y), x - y\rangle &= \left\langle \int_0^1 \mathrm{Jac}^c_{\nabla h}((1-t)x + ty)(x-y)dt, x - y\right\rangle\\ &= \left\langle \int_0^1 \mathrm{Jac}^c_{\nabla h}((1-t)x + ty)dt(x-y), x - y\right\rangle\\ &\geq \lambda\|x-y\|^2,\end{aligned}$$

which means strong monotonicity of ∇h, equivalent to strong convexity of h. By contraposition, if h is not locally strongly convex around 0, then, $\mathrm{Jac}^c_{\nabla h}(0)$ contains singular matrices. □

Proposition 4.3 shows that the equivalence between contractivity and (4.2) does not hold in the $C^{1,1}$ case, higlighting a gap between $C^{1,1}$ and C^2 functions. This actually extends to the nonsmooth setting with a proximal point step using convex analysis and Moreau enveloppes.

Proposition 4.4 (Linear convergence does not imply contractivity II). *There exists a convex semialgebraic function $g\colon \mathbb{R}^2 \to \mathbb{R}$ such that $g(0) = 0$, prox_g is path differentiable, and, for some $0 < \rho < 1$, $\left\|\operatorname{prox}_g(x)\right\| \leq \rho \|x\|$ for all x, such that, for any convex conservative Jacobian $\mathcal{J}_{\operatorname{prox}_g}$, not every element of $\operatorname{Id}_2 - \mathcal{J}_{\operatorname{prox}_g}(0)$ is invertible.*

Proof. Let $h \in C^{1,1}(\mathbb{R}^2)$ be a function given by Proposition 4.3, it is convex and semialgebraic with ∇h 1-Lipschitz, path differentiable, and $\|x - \nabla h(x)\| \leq \rho\|x\|$ for some $\rho \in (0,1)$ such that h is not strongly convex locally around 0 and $\operatorname{Jac}^c_{\nabla h}(0)$ contains singular elements. The function $\tilde{h}\colon x \mapsto \frac{\|x\|^2}{2} - h(x)$ is convex [7, Theorem 18.15 (vi)] with 1-Lipschitz gradient. Recall that the *conjugate* of $\tilde{h}$ is $\tilde{h}^*(x)\colon x \mapsto \sup_u \langle x, u\rangle - \tilde{h}(u)$. The function $g\colon x \mapsto \tilde{h}^*(x) - \frac{\|x\|^2}{2}$ is convex and semialgebraic, because $\tilde{h}^*$ is 1-strongly convex and semialgebraic [7, Proposition 10.8]. It satisfies $\operatorname{prox}_g(x) = \nabla\tilde{h}(x) = x - \nabla h(x)$ [7, Corollary 24.5], so that the gradient descent mapping for h with unit step size is equivalent to the prox operator for g.

Thus for all x, $\|\operatorname{prox}_g(x)\| = \|x - \nabla h(x)\| \leq \rho\|x\|$ for some $\rho \in (0,1)$, prox_g is path differentiable, and $\mathcal{J}_{\nabla h} := \operatorname{Id}_2 - \mathcal{J}_{\operatorname{prox}_g}$ is a convex conservative Jacobian for ∇h. Finally, by contraposition, $\operatorname{Id}_2 - \mathcal{J}_{\operatorname{prox}_g}(0) = \mathcal{J}_{\nabla h}(0)$ which contains at least one singular elemen by Proposition 4.3 using the fact that $\operatorname{Jac}^c_{\nabla h}(0) \subset \mathcal{J}_{\nabla h}(0)$ by convexity of $\mathcal{J}_{\nabla h}(0)$. □

Remark 4.5 (On local growth conditions). The result of this section may be refined by considering local growth conditions, which are sufficient to ensure linear convergence of the forward-backward algorithm as in (4.2). Most important examples include global error bounds

$$g(x) - \min_z g(z) \geq \lambda \operatorname{dist}(x, \operatorname{argmin} g)^2$$

for some $\lambda > 0$ and for all x (see [58] and references therein) as well as global Kurdyka-Łojasiewicz inequality

$$\min_{v \in \partial g(x)} \|v\| \geq \lambda' \sqrt{g(x) - \min_z g(z)}$$

for some $\lambda' > 0$ and for all x (see [17] and references therein). In our setting (coercive, semialgebraic, unique critical value), these conditions are equivalent and are sufficient for (4.2) to hold true [17]. On the other hand, (4.2) implies a quadratic Łojasiewicz inequality (KL with exponent $1/2$) or a global quadratic error bound in our setting [37, Proposition 4.19]. Since all these conditions are equivalent the results of this section actually hold true replacing (4.2) by a quadratic error bound or a quadratic Łojasiewicz inequality showing a fundamental limit to the extension of Theorem 4.2 beyond strong convexity.

5 Applications to saddle point problems and duality

We demonstrate how to apply the previous sections' results to several different parametric optimization problems in which one seeks to differentiate the solution mapping as a function of the parameters θ. In each of the following subsections, $\Theta \subset \mathbb{R}^p$ is a connected open set on which Assumption 4.1 will be required to hold for various operators.

5.1 Differentiating the dual solution of a convex composite problem

Consider the following composite minimization problem

$$\min_{x\in\mathbb{R}^n} f_\theta(x) + g_\theta(K_\theta x) \tag{5.1}$$

where, for each $\theta \in \Theta$, $f_\theta \in C^{1,1}(\mathbb{R}^n)$ is a strongly convex function, $g_\theta \colon \mathbb{R}^m \to \bar{\mathbb{R}}$ is a proper closed convex function, $K_\theta \colon \mathbb{R}^n \to \mathbb{R}^m$ is a surjective linear operator, and $x^\star(\theta)$ is the unique solution (the objective is proper by surjectivity).

The goal is to differentiate the solution $x^\star$ with respect to θ, for which we assume that ∇f_θ and $\operatorname{prox}_{\gamma g_\theta}$ are path differentiable, jointly in (θ, x), so that Assumption 4.1 holds here. It is then possible to directly apply Theorem 4.2 to differentiate $x^\star$ since Assumption 4.1 holds and f_θ is strongly convex. However, because of the coupling between g_θ and K_θ in (5.1), computing $\mathcal{J}_{x^\star}$ through this approach would necessitate computing $\operatorname{prox}_{g_\theta \circ K_\theta}$, which is nontrivial even when $\operatorname{prox}_{g_\theta}$ is known (unless K_θ is a (semi)orthogonal matrix [30, Lemma 2.8]).

We can instead use the generalized duality of Fenchel-Rockafellar, which will decouple the linear operator K_θ from g_θ in a way that is especially useful if K_θ is surjective, which we will assume. The dual problem of (5.1) is given, for each $\theta \in \Theta$, by

$$-\min_{y\in\mathbb{R}^m} f_\theta^*(-K_\theta^* y) + g_\theta^*(y) \tag{5.2}$$

to which we can apply our results, with $f_\theta^*(-K_\theta^* \cdot)$ and g_θ^* taking the role of f and g in the assumptions and theorems. Note that $y \mapsto f_\theta^*(-K_\theta^* y)$ is indeed strongly convex. To be explicit, we take $T(\theta, y) = \operatorname{prox}_{\gamma g_\theta^*}(y)$ and $S(\theta, y) = (\theta, y + \gamma K_\theta \nabla f_\theta^*(-K_\theta^* y))$ with $H(\theta, y) = T(S(\theta, y))$, so that the fixed point equation we are considering the dual solution $y^\star(\theta)$ to satisfy is

$$\operatorname{res}(\theta, y^\star(\theta)) = y^\star(\theta) - \operatorname{prox}_{\gamma g_\theta^*}(y^\star(\theta) + \gamma K_\theta \nabla f_\theta^*(-K_\theta^* y^\star(\theta))) = 0. \tag{5.3}$$

Theorem 5.1 (Path differentiability of the dual solution of a composite problem). *Consider (5.2) where, for each $\theta \in \Theta$, $f_\theta \in C^{1,1}(\mathbb{R}^n)$ is α-strongly convex with β-Lipschitz continuous gradient, $g_\theta \colon \mathbb{R}^m \to \bar{\mathbb{R}}$ is a closed convex proper function and $K_\theta \in \mathbb{R}^{m\times n}$ is surjective with singular values in $[\underline{\lambda}, \bar{\lambda}]$ for some $0 < \underline{\lambda} \le \bar{\lambda}$, uniformly in θ. Then $y \mapsto f_\theta^*(-K_\theta^* y)$ is $\underline{\lambda}^2/\beta$ strongly convex and has a gradient which is $\bar{\lambda}^2/\alpha$-Lipschitz continuous, uniformly in θ.*

Assume furthermore that $\operatorname{prox}_{\gamma g_\theta^*}$, ∇f_θ^* and K_θ are path differentiable so that Assumption 4.1 holds with $f_\theta^* \circ [-K_\theta^*]$ and g_θ^*. Then, the unique dual solution $y^\star(\theta)$ of (5.2) is path differentiable on Θ with a conservative Jacobian given for all $\theta \in \Theta$ by

$$\mathcal{J}_{y^\star} \colon \theta \rightrightarrows \left\{ \left(\mathrm{Id}_n - V(\mathrm{Id}_n - \gamma Z)\right)^{-1} (U - \gamma VW) : \ [U\ V] \in \mathrm{Jac}_1^c, [W\ Z] \in \mathrm{Jac}_2^c \right\}$$

where

$$\mathrm{Jac}_1^c := \mathrm{Jac}_T^c(\theta, y^\star(\theta) + \gamma K_\theta \nabla f_\theta^*(-K_\theta^* y^\star(\theta))) \qquad \textit{and} \qquad \mathrm{Jac}_2^c := \mathrm{Jac}^c_{\nabla(f_\theta \circ [-K_\theta^*])}(\theta, y^\star(\theta)).$$

and γ is any number in $(0, \frac{2\underline{\lambda}^2/\beta}{(\underline{\lambda}^2/\beta + \bar{\lambda}^2/\alpha)})$.

Proof. For each $\theta \in \Theta$, the function $g_\theta^* \colon \mathbb{R}^m \to \bar{\mathbb{R}}$ is a closed convex proper function since g_θ is, meanwhile the function $f_\theta^* \colon \mathbb{R}^n \to \mathbb{R}$ is $\frac{1}{\beta}$-strongly convex and differentiable with $\frac{1}{\alpha}$-Lipschitz continuous gradient [7,

Theorem 18.15(vii)]. The function $f_\theta^*(-K_\theta^* y)$ has gradient $-K_\theta \circ \nabla f_\theta^* \circ -K_\theta^*$ which is $\frac{\bar{\lambda}^2}{\alpha}$-Lipschitz continuous because ∇f_θ is Lipschitz continuous with constant $\frac{1}{\alpha}$ and both K_θ and K_θ^* are $\bar{\lambda}$-Lipschitz continuous. To demonstrate that $f_\theta^* \circ -K_\theta^*$ is $\frac{\underline{\lambda}^2}{\beta}$-strongly convex, we have for all y

$$f_\theta^*(-K_\theta^* y) - \frac{\underline{\lambda}^2}{2\beta}\|y\|^2 = \left(f_\theta^*(-K_\theta^* y) - \frac{1}{2\beta}\|K_\theta^* y\|^2\right) + \left(\frac{1}{2\beta}\|K_\theta^* y\|^2 - \frac{\underline{\lambda}^2}{2\beta}\|y\|^2\right).$$

The first term, $f_\theta^*(-K_\theta^* y) - \frac{1}{2\beta}\|K_\theta^* y\|^2$, is convex as it is the composition of a convex function $f(\cdot) - \frac{1}{2\beta}\|\cdot\|^2$ and a linear map $-K_\theta^*$. Indeed, by $\frac{1}{\beta}$-strong convexity of f_θ^*, the function $f_\theta^*(\cdot) - \frac{1}{2\beta}\|\cdot\|^2$ is necessarily convex [7, Proposition 10.6]. The second term, $\frac{1}{2\beta}\left(\|K_\theta^* y\|^2 - \underline{\lambda}^2\|y\|^2\right)$, is convex because the smallest eigenvalue of $K_\theta K_\theta^*$ is $\underline{\lambda}^2$. Hence the claimed strong convexity modulus of $\frac{\underline{\lambda}^2}{\beta}$, justifying the first part of the theorem claiming regularity of $f_\theta^* \circ -K_\theta^*$. Then, using the assumption that $\operatorname{prox}_{\gamma g_\theta}$, ∇f_θ^*, and K_θ are all path differentiable so that Assumption 4.1 holds, we are finally able to apply Theorem 4.2 to (5.2) and its fixed point formulation (5.3) and the desired results follow. □

Remark 5.2 (Path differentiability of the primal solution). We can recover the primal solution from the dual solution through the equation $x^\star(\theta) = \nabla f_\theta^*(-K_\theta^* y^\star(\theta))$, coming from the primal-dual optimality conditions, since ∇f_θ^* is path differentiable. Indeed, the functions $\operatorname{prox}_{\gamma g_\theta^*}$ and ∇f_θ^* are path differentiable if $\operatorname{prox}_{\gamma g_\theta}$ and ∇f_θ are assumed to be path differentiable. By the Moreau decomposition [7, Theorem 14.3(ii)] we can express $\operatorname{prox}_{\gamma g_\theta^*}(y) = y - \operatorname{prox}_{g_\theta/\gamma}(y/\gamma)$. Meanwhile for ∇f_θ^*, we can invoke the path differentiable inverse function theorem [16, Corollary 2] with $\nabla f_\theta^* = (\nabla f_\theta)^{-1}$, the assumptions of which hold due to the fact that ∇f_θ is path differentiable and f_θ^* is both Lipschitz-smooth and strongly convex.

Example 5.3 (Learning sparsity priors). The problem of learning a sparsity prior can be seen as a bilevel optimization problem [38, 61, 63] which fits the framework of this subsection. Given some set of training data $\{(u_1, \hat{u}_1), \ldots, (u_q, \hat{u}_q)\}$ where u_i is the ground truth for some signal (e.g., an image) and $\hat{u}_i$ is a noisy observation of u_i, we seek to find an optimal linear operator $K_\theta \in \mathbb{R}^{s\times n}$, the so-called sparsity prior. The general form of the problem can be cast as the following bilevel optimization problem,

$$\min_{\theta\in\mathbb{R}^p} \sum_{i=1}^{q} \frac{1}{2}\|u_i - x_i(\theta)\|_2^2 \quad \text{such that, } \forall i \in \{1,\ldots,q\}, \quad x_i(\theta) \in \operatorname*{argmin}_{x_i\in\mathbb{R}^n} \frac{1}{2}\|x_i - \hat{u}_i\|_2^2 + \|K_\theta x_i\|_1$$

where $\theta \in \mathbb{R}^{sn}$ with $K_\theta := \begin{bmatrix} \theta_{1,1} & \ldots & \theta_{1,n} \\ \vdots & \ddots & \vdots \\ \theta_{s,1} & \ldots & \theta_{s,n} \end{bmatrix}$. Assume that $\Theta \subset \mathbb{R}^{sn}$ is a connected open set such that, for all $\theta \in \Theta$, K_θ is surjective and its singular values are contained in $[\underline{\lambda}, \bar{\lambda}]$ for some $0 < \underline{\lambda} \leq \bar{\lambda}$. Then the lower level problem matches exactly that of (5.1) with $f_\theta(x) = \sum_{i=1}^{q} \frac{1}{2}\|x_i - \hat{u}_i\|_2^2$ and $g_\theta(x) = \sum_{i=1}^{q} \frac{1}{2}\|K_\theta x_i\|_1$, which we write here as sums even though they are separable in x_i. Note that K_θ is obviously not surjective for all $\theta \in \mathbb{R}^{sn}$ because of the general parametrization chosen. Instead of fixing the required open set Θ, one could employ a different parameterization of K_θ with constraints on parameters ensuring that K_θ remains surjective. Regardless, by assuming surjectivity, Theorem 5.1 applies and we can continue.

The dual of the inner problem, is given, for each $i \in \{1,\ldots,q\}$, by

$$y_i(\theta) \in \operatorname*{argmin}_{\{y_i\in\mathbb{R}^s : \|y_i\|_\infty \leq 1\}} \frac{1}{2}\|K_\theta^* y_i - \hat{u}_i\|_2^2$$

which has a fixed point equation

$$y_i^\star = \mathrm{P}_{\mathcal{D}}(y_i^\star + \gamma K_\theta(K_\theta^* y_i^\star - \hat{u}_i)),$$

where $\mathrm{P}_{\mathcal{D}}$ is the projection onto the ℓ^∞ unit ball in $\mathbb{R}^s$, i.e., the mapping whose coordinates are given by $z \mapsto \operatorname{sign}(z)\min(1, |z|)$ component-wise. Using the notation of Section 3, we have $T(\theta, y_i) = \mathrm{P}_{\mathcal{D}}(y_i)$ and $S(\theta, y) = y_i^\star + \gamma K_\theta(K_\theta^* y_i^\star - \hat{u}_i)$. The primal solution $x^\star$ can be recovered from the dual solution through the relationship given in Remark 5.2, for each $\theta \in \Theta$ and $i \in \{1, \dots, q\}$,

$$x_i^\star(\theta) = \nabla f_\theta^*(-K_\theta^* y_i^\star(\theta)) \implies x_i^\star(\theta) = \hat{u}_i - K_\theta^* y_i^\star(\theta).$$

We emphasize the difference in our approach to those taken in previous works [38, 61, 63]. While [61] relies on a smoothing process for the ℓ^1-norm in the lower level problem, [63] assumes that K_θ is an orthogonal matrix in contrast to our assumption that K_θ is surjective. In [38], the authors use unrolling on the algorithm used to solve the lower-level problem rather than implicit differentiation as we do.

5.2 Differentiating the solutions of min-max problems

Consider the following min-max problem

$$\min_{x \in X} \max_{y \in Y} \Phi_\theta(x, y) \tag{5.4}$$

where $X \subset \mathbb{R}^n$ is closed and convex, and $Y \subset \mathbb{R}^m$ is convex compact and, for each $\theta \in \Theta$, $\Phi_\theta \colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is continuous such that $-\Phi_\theta(x, \cdot) \colon \mathbb{R}^m \to \mathbb{R}$ and $\Phi_\theta(\cdot, y) \colon \mathbb{R}^n \to \mathbb{R}$ are α-strongly convex for each x and for each y, respectively. Assume also that $\Phi_{(\cdot)}(x, y)$ is Lipschitz continuous on Θ for all $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$. The very general form of this problem encompasses min-max problems with nonlinear couplings of the form considered in [41, 42]. The solution mapping for this problem incorporates the primal and dual variables together, $\theta \mapsto (x^\star(\theta), y^\star(\theta))$. The optimality condition can be written for each $\theta \in \Theta$ as

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \partial_x \Phi_\theta + N_X & 0 \\ 0 & -\partial_y \Phi_\theta + N_Y \end{pmatrix} \begin{pmatrix} x^\star(\theta) \\ y^\star(\theta) \end{pmatrix}$$

where N_X and N_Y denote respectively the normal cones to X and Y. This is a special case of ($\mathcal{P}_{\mathrm{mono}}$) with $\mathcal{A}_\theta = \begin{pmatrix} \partial_x \Phi_\theta + N_X & 0 \\ 0 & -\partial_y \Phi_\theta + N_Y \end{pmatrix}$ and $\mathcal{B}_\theta \equiv 0$. Indeed from strong convexity, for each $\theta \in \Theta$, $\mathcal{A}_\theta$ is α strongly monotone, and from closedness of X and compactness of Y it can be shown that the range of $I + \mathcal{A}_\theta$ is $\mathbb{R}^n \times \mathbb{R}^m$ so that $\mathcal{A}_\theta$ is maximal [69, Theorem 12.12].

For the function H_θ defined in Assumption 3.1 applied to this problem, we have $H_\theta(x) = \mathcal{R}_{\mathcal{A}_\theta}(x)$ so that $\operatorname{res}(\theta, x) = x - \mathcal{R}_{\mathcal{A}_\theta}(x)$, leading to the following result.

Theorem 5.4 (Path differentiability of min-max solutions). *Consider (5.4) where, for each $\theta \in \Theta$, $-\Phi_\theta(x, \cdot) \colon \mathbb{R}^m \to \bar{\mathbb{R}}$ and $\Phi_\theta(\cdot, y) \colon \mathbb{R}^n \to \bar{\mathbb{R}}$ are both closed α-strongly convex proper functions. Assume that $\mathcal{R}_{\mathcal{A}_\theta}$ is path differentiable so that* Assumption 3.1 *holds with the conservative Jacobians defined in (3.3) with $\gamma \in (0, 1/\alpha)$. Then, the solution mapping $\theta \mapsto (x^\star(\theta), y^\star(\theta))$ is unique and path differentiable on Θ with a conservative Jacobian given for each $\theta \in \Theta$ by*

$$\mathcal{J}_{(x^\star, y^\star)} \colon \theta \rightrightarrows \{\left(\mathrm{Id}_{(n+m)} - [V_1\ V_2]\right)^{-1} U \colon [U\ V_1\ V_2] \in \mathrm{Jac}_T^c(\theta, (x^\star(\theta), y^\star(\theta)))\}.$$

Proof. For each $\theta \in \Theta$, the maximal monotone operator $\mathcal{A}_\theta$ is α-strongly monotone by the α-strong convexity and α-strong concavity of $\Phi_\theta(\cdot, y)$ and $\Phi_\theta(x, \cdot)$ respectively. Using this fact with Assumption 3.1 and the conservative Jacobians defined in (3.3), we can apply Theorem 3.5 with any $\beta \in (0, \infty)$ since $\mathcal{B}_\theta \equiv 0$. In particular, we can take $\beta = (\sqrt{2} - 1)\alpha$ so that applying Theorem 3.5 requires $\gamma \in (0, 1/\alpha)$. Therefore, the solution $(x^\star(\theta), y^\star(\theta))$ is path differentiable with a conservative Jacobian given by

$$\mathcal{J}_{(x^\star, y^\star)} \colon \theta \rightrightarrows \{\left(\mathrm{Id}_{(n+m)} - [V_1\ V_2]\right)^{-1} U \colon [U\ V_1\ V_2] \in \mathrm{Jac}^c_T(\theta, (x^\star(\theta), y^\star(\theta)))\}.$$

□

5.3 Differentiating the solutions of primal-dual problems

The min-max problem (5.4) from the previous subsection is general in that it does not assume a particular coupling between x and y. We turn now to primal-dual optimization with linear coupling, a well-known problem template that was studied, for instance, in [24, 25, 26, 31, 48, 72] and allows to model many different problems coming from computer vision and machine learning. Consider the parametrized primal-dual problem

$$\min_{x \in \mathbb{R}^n} g_\theta(x) + \max_{y \in \mathbb{R}^m} \langle K_\theta x, y \rangle - f^*_\theta(y) \tag{5.5}$$

where, for all $\theta \in \Theta$, $K_\theta \in \mathbb{R}^{m \times n}$ is a linear operator and both $f^*_\theta \colon \mathbb{R}^m \to \bar{\mathbb{R}}$ and $g_\theta \colon \mathbb{R}^n \to \bar{\mathbb{R}}$ are closed convex proper functions. In contrast to [27, 14], we allow for the functions g_θ and f^*_θ to be parametrized by θ, in addition to the linear operator K_θ. For each $\theta \in \Theta$, the optimality conditions for a solution $(x^\star(\theta), y^\star(\theta))$ to this problem are

$$K_\theta x^\star(\theta) \in \partial f^*_\theta(y^\star(\theta)) \quad \text{and} \quad -K^*_\theta y^\star(\theta) \in \partial g_\theta(x^\star(\theta))$$

which can be equivalently written as

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \underbrace{\begin{pmatrix} \partial g_\theta & 0 \\ 0 & \partial f^*_\theta \end{pmatrix}}_{\mathcal{A}_\theta} \begin{pmatrix} x^\star(\theta) \\ y^\star(\theta) \end{pmatrix} + \underbrace{\begin{pmatrix} 0 & K^*_\theta \\ -K_\theta & 0 \end{pmatrix}}_{\mathcal{B}_\theta} \begin{pmatrix} x^\star(\theta) \\ y^\star(\theta) \end{pmatrix}$$

with $\mathcal{A}_\theta$ maximal monotone and $\mathcal{B}_\theta$ maximal monotone and β-Lipschitz for some $\beta = \|K^*_\theta\|_{\mathrm{op}} > 0$. Despite the fact that the operator $\mathcal{B}_\theta$ is not cocoercive, we can still apply the results developed in Section 3 because it is β-Lipschitz.

Theorem 5.5 (Path differentiability for primal-dual problems). *Consider* (5.5) *where, for each* $\theta \in \Theta$, $g_\theta \colon \mathbb{R}^n \to \bar{\mathbb{R}}$ *and* $f^*_\theta \colon \mathbb{R}^m \to \bar{\mathbb{R}}$ *are closed* α*-strongly convex proper functions and* $\beta \geq \|K_\theta\|_{\mathrm{op}} > 0$ *for some* $\beta > 0$. *Assume also, for each* $\theta \in \Theta$, *that* prox_{g_θ} *and* prox_{f_θ} *are path differentiable with conservative Jacobians chosen according to* (3.3) *so that* Assumption 3.1 *holds with* $\gamma \in (0, \frac{2\alpha}{(\alpha+\beta)^2})$. *Then the mapping* $\theta \mapsto (x^\star(\theta), y^\star(\theta))$ *is unique and path differentiable on* Θ *with a conservative Jacobian given for each* $\theta \in \Theta$ *by*

$$\mathcal{J}_{(x^\star, y^\star)} \colon \theta \rightrightarrows \left\{-[V_1\ V_2]^{-1} U \colon U \in \mathcal{J}_{\theta, \mathrm{res}}\,(\theta, x^\star(\theta), y^\star(\theta)), [V_1\ V_2] \in \mathcal{J}_{(x,y), \mathrm{res}}\,(\theta, x^\star(\theta), y^\star(\theta))\right\}$$

where

$$\mathcal{J}_{(x,y), \mathrm{res}} : (\theta, x^\star(\theta), y^\star(\theta)) \rightrightarrows$$
$$\begin{bmatrix} \mathrm{Id}_m - \mathrm{Jac}^c_{\mathrm{prox}_{\gamma g_\theta}}(\theta, x^\star(\theta) - \gamma K^*_\theta y^\star(\theta)) & -\mathrm{Jac}^c_{\mathrm{prox}_{\gamma g_\theta}}(\theta, x^\star(\theta) - \gamma K^*_\theta y^\star(\theta)) \times (-\gamma K^*_\theta) \\ -\mathrm{Jac}^c_{\mathrm{prox}_{\gamma f^*_\theta}}(\theta, y^\star(\theta) + \gamma K_\theta x^\star(\theta)) \times (\gamma K_\theta) & \mathrm{Id}_n - \mathrm{Jac}^c_{\mathrm{prox}_{\gamma f^*_\theta}}(\theta, y^\star(\theta) + \gamma K_\theta x^\star(\theta)) \end{bmatrix}.$$

Proof. For each $\theta \in \Theta$, due to the assumed strong convexity of g_θ and f_θ^*, the operator $\mathcal{A}_\theta = \begin{pmatrix} \partial g_\theta & 0 \\ 0 & \partial f_\theta^* \end{pmatrix}$ is strongly monotone with some constant $\alpha > 0$ and, since K_θ is a linear operator, $\mathcal{B}_\theta = \begin{pmatrix} 0 & K_\theta^* \\ -K_\theta & 0 \end{pmatrix}$ is maximal monotone and Lipschitz with constant β. Combining this with Assumption 3.1 and the conservative Jacobians chosen in (3.3), the conditions to apply Theorem 3.5 with $\gamma \in (0, \frac{2\alpha}{(\alpha+\beta)^2})$ are met and the function $\theta \mapsto (x^\star(\theta), y^\star(\theta))$ is path differentiable. More explicitly, we have for all $\theta \in \Theta$, $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$,

$$H(\theta, x, y) = \begin{pmatrix} \operatorname{prox}_{\gamma g_\theta}(x - \gamma K_\theta^* y) \\ \operatorname{prox}_{\gamma f_\theta^*}(y + \gamma K_\theta x) \end{pmatrix}$$

so that

$$\operatorname{res}(\theta, x^\star(\theta), y^\star(\theta)) = \begin{pmatrix} x^\star(\theta) \\ y^\star(\theta) \end{pmatrix} - \begin{pmatrix} \operatorname{prox}_{\gamma g_\theta}(x^\star(\theta) - \gamma K_\theta^* y^\star(\theta)) \\ \operatorname{prox}_{\gamma f_\theta^*}(y^\star(\theta) + \gamma K_\theta x^\star(\theta)) \end{pmatrix}.$$

Using (3.3), the resulting conservative Jacobian for $(x^\star(\theta), y^\star(\theta))$ on Θ is

$$\mathcal{J}_{(x^\star, y^\star)} \colon \theta \rightrightarrows$$
$$\{(\mathrm{Id}_{n+m} - [V_1\ V_2](\mathrm{Id}_{n+m} - \gamma[Z_1\ Z_2]))^{-1}(U - \gamma[V_1\ V_2]W) : [U\ V_1\ V_2] \in \mathrm{Jac}_1^c, [W\ Z_1\ Z_2] \in \mathrm{Jac}_2^c\}.$$

where

$$\mathrm{Jac}_1^c = \mathrm{Jac}_T^c(\theta, x^\star(\theta) - \gamma K_\theta^* y^\star(\theta), y^\star(\theta) + \gamma K_\theta x^\star(\theta))$$

and

$$\mathrm{Jac}_2^c = \mathrm{Jac}_{\mathcal{B}}^c(\theta, x^\star(\theta), y^\star(\theta)).$$

Alternatively, we can write

$$\mathcal{J}_{(x^\star, y^\star)} \colon \theta \rightrightarrows \{[V_1\ V_2]^{-1} U : U \in \mathcal{J}_{\theta, \mathrm{res}}(\theta, x^\star(\theta), y^\star(\theta)), [V_1\ V_2] \in \mathcal{J}_{(x,y), \mathrm{res}}(\theta, x^\star(\theta), y^\star(\theta))\}$$

where

$$\mathcal{J}_{(x,y),\mathrm{res}} : (\theta, x^\star(\theta), y^\star(\theta)) \rightrightarrows$$
$$\begin{bmatrix} \mathrm{Id}_m - \mathrm{Jac}^c_{\operatorname{prox}_{\gamma g_\theta}}(\theta, x^\star(\theta) - \gamma K_\theta^* y^\star(\theta)) & -\mathrm{Jac}^c_{\operatorname{prox}_{\gamma g_\theta}}(\theta, x^\star(\theta) - \gamma K_\theta^* y^\star(\theta)) \times (-\gamma K_\theta^*) \\ -\mathrm{Jac}^c_{\operatorname{prox}_{\gamma f_\theta^*}}(\theta, y^\star(\theta) + \gamma K_\theta x^\star(\theta)) \times (\gamma K_\theta) & \mathrm{Id}_n - \mathrm{Jac}^c_{\operatorname{prox}_{\gamma f_\theta^*}}(\theta, y^\star(\theta) + \gamma K_\theta x^\star(\theta)) \end{bmatrix}.$$

□

As in [14] and [27], we have assumed that both g_θ and f_θ^* are strongly convex; in contrast to [14] we do not assume that $\operatorname{prox}_{g_\theta}$ or $\operatorname{prox}_{f_\theta^*}$ are differentiable at any specific points, nor do we assume that g_θ or f_θ are twice differentiable as in [27].

In [27] the authors consider using a bilevel optimization problem to learn the best discretization of the total variation for inverse problems in imaging. The proposed bilevel problem was further studied in [14] where the authors noted the theoretical difficulties in differentiating the solution to a nonsmooth optimization problem with respect to some parameters. Indeed, they have no guarantees that their algorithm will avoid the set of points where the prox operator is not differentiable, nor can they show that the solutions will be points of differentiability. On the other hand, so long as the prox operator is path differentiable, our results apply despite these obstacles, and one can compute implicit conservative gradients.

5.4 Automatic differentiation of algorithms

The results of Theorem 3.5 (in particular, the fact that res is contractive) imply that the forward-backward splitting algorithm applied to solve monotone inclusions of the form in the theorem will converge linearly for $\gamma \in \left(0, \frac{2\alpha^2}{(\alpha+\beta)^2}\right)$. We illustrate our results with implicit differentiation which only requires the solution of the inclusion problem and does not depend on the algorithm used to solve it. Yet, it is worth emphasizing that under Assumption 3.1, the contractivity property in Definition 3.3 is sufficient to apply the convergence result of [19] to the forward-backward algorithm in our context. More precisely, Definition 3.3 is precisely the same as [19, Assumption 1] applied to the forward-backward algorithm to solve (1.1). Therefore, in addition to path differentiability of the solution map, assumptions of Theorem 3.5 provide a sufficient condition to ensure that automatic differentiation of the forward-backward algorithm generates a sequence of conservative jacobians such that conservativity is preserved asymptotically: the limits of iterative differentiation conservative jacobians form a conservative jacobian for the solution of (1.1) [19, Corollary 1 and 2].

As a consequence, Theorem 3.5 implies that the iterative derivative convergence results of [19] apply to the forward-backward algorithm in all special cases described in the paper: Theorem 4.2 for strongly convex optimization problems, Theorem 5.1 for solutions to primal problems obtained from their dual, Theorem 5.4 for general strongly monotone saddle point problems and Theorem 5.5 for strongly monotone structured saddle point problems.

6 Conclusion

We have presented sufficient conditions in the form of strong monotonicity, under which the path differentiability of the solution to a monotone inclusion problem is satisfied. As special cases, we have derived conditions that ensure path differentiability of solutions to a large class of nonsmooth parametric convex optimization problems - those which can be written as the sum of two parametric convex functions, one smooth and one possibly nonsmooth. By expressing the monotone inclusions as equivalent fixed point equations using the resolvent mapping, we were able to leverage path differentiability and the recently developed nonsmooth implicit path differentiation theorem of [16, Corollary 1] to deduce regularity of $x^\star$. Most importantly, we were able to characterize and give a formula for a conservative Jacobian of $x^\star$ with respect to θ using only the Clarke Jacobians associated with the resolvent mapping $\mathcal{R}_{\gamma\mathcal{A}_\theta}$ and the operator $\mathcal{B}_\theta$.

While this work is primarily theoretical, our results also lend insight to practical applications, e.g., the design of implicit neural network layers defined using convex optimization problems. Ensuring that an implicit layer is compatible with training is an important part of implicit layer design and, consequently, ensuring the invertibility condition is a necessary part of guaranteeing that training will work (c.f., [16, Section 5]). Besides this, our results highlight the relevancy of the typical strong convexity assumption made on the lower-level problem of a bilevel optimization problem to ensure implicit conservative Jacobians will exist when this lower-level problem is nonsmooth.

It is important also to understand the dependence of $x^\star$ on the step size γ taken in the definition of H. In the smooth case, where $\mathcal{A}_\theta \equiv 0$ and $\mathcal{B}_\theta = \nabla f_\theta$ for some twice differentiable convex function f_θ, there is no dependence on γ which is to be expected since $0 = \mathcal{B}_\theta(x^\star)$ can be differentiated directly without introducing γ. We delay exploring such questions for future work.

At a few points in the paper, we encounter objects which are not canonical or which can be chosen in multiple ways. We collect here these instances and elaborate on why we've chosen the way we have and the possible consequences of choosing to formulate things differently.

Choice of H Our selection for H in res can be attributed to the additive structure of $\mathcal{A}_\theta + \mathcal{B}_\theta$ and the Lipschitz continuity of $\mathcal{B}_\theta$. There are alternatives, for instance, considering the resolvent directly $\mathcal{R}_{\mathcal{A}_\theta+\mathcal{B}_\theta}$ or considering Douglas-Rachford, Peaceman-Rachford, etc, and extensions of this flavor are a matter for future research.

Choice of $\mathcal{J}_H$ We use multiplication of Clarke Jacobians in the formula for $\mathcal{J}_H$ given in (3.3) because it allows to obtain sufficient condition on problem data in ($\mathcal{P}_{\text{mono}}$) to ensure the contractivity of the residual equation. On the other hand, all the corollaries of the paper would hold similarly with arbitrary conservative Jacobians for $\mathcal{R}_{\gamma\mathcal{A}}$ and $\mathcal{B}$ combined in a similar way, provided that this is compatible with the contractivity in Definition 3.3. However, in this case, the contractivity defined in Definition 3.3 has to be explicitly assumed, not deduced from properties of problem data, because conservative Jacobians can be changed on a set of measure zero.

Choice of Θ The set Θ could be the whole space $\mathbb{R}^p$ or possibly a subset, for example if one of the operators $\mathcal{A}_\theta$ or $\mathcal{B}_\theta$ is not defined for every $\theta \in \mathbb{R}^p$, or if some of conditions (Lipschitz continuity, strong monotonicity, etc) can only be ensured to hold on some subset. The set Θ can also be taken as a neighborhood of some point $\bar{\theta} \in \mathbb{R}^p$ of interest, in which case it's possible to obtain local versions of all of our later results. For these local versions, we need only to assume that the contractivity in Definition 3.3 holds at the single point $\bar{\theta}$, from which it follows that there is some open neighborhood Θ on which the inequality must hold.

Acknowledgements

The authors acknowledge the support of the Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant numbers FA9550-19-1-7026. JB and EP acknowledge the support of ANR-3IA Artificial and Natural Intelligence Toulouse Institute, and thank Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant numbers FA8655-22-1-7012 ANR MasDol. JB acknowledges the support of ANR Chess, grant ANR-17-EURE-0010.

A Appendix

The next lemma is a general result that gives a bound on the Lipschitz constant of the forward mapping $\mathrm{Id} - \gamma\mathcal{B}$ in terms of the Lipschitz constant β of the maximal monotone operator $\mathcal{B}$. This bound is slightly better than the more obvious bound $1 + \gamma\beta$, and this improvement is crucial to prove Theorem 3.5.

Lemma A.1. *Let $\mathcal{B}\colon \mathbb{R}^n \to \mathbb{R}^n$ be a β-Lipschitz continuous maximal monotone operator, then the map $\mathrm{Id} - \gamma\mathcal{B}_\theta$ is $\sqrt{1+\gamma^2\beta^2}$-Lipschitz continuous.*

Proof. For all $x, y \in \mathbb{R}^n$,

$$\begin{aligned}\|(\mathrm{Id} - \gamma\mathcal{B})(x) - (\mathrm{Id} - \gamma\mathcal{B})(y)\|^2 &= \|x - y\|^2 - 2\gamma\,\langle x - y, \mathcal{B}(x) - \mathcal{B}(y)\rangle + \|\gamma(\mathcal{B}(x) - \mathcal{B}(y))\|^2 \\ &\leq \|x - y\|^2 + \gamma^2\,\|\mathcal{B}(x) - \mathcal{B}(y)\|^2 \\ &\leq (1 + \gamma^2\beta^2)\,\|x - y\|^2\end{aligned}$$

where we have used the monotonicity of $\mathcal{B}$ followed by the β-Lipschitz continuity of $\mathcal{B}$ for the first and second inequalities, respectively. Thus, taking square roots, the mapping $\mathrm{Id} - \gamma\mathcal{B}$ is $\sqrt{1+\gamma^2\beta^2}$-Lipschitz continuous. □

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. {TensorFlow}: a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.

[2] Samir Adly and R Tyrrell Rockafellar. Sensitivity analysis of maximally monotone inclusions via the proto-differentiability of the resolvent operator. *Mathematical Programming*, 189(1):37–54, 2021.

[3] Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning*, pages 136–145. PMLR, 2017.

[4] Hedy Attouch. Modern maximal monotone operator theory: From nonsmooth optimization to differential inclusions. 2019.

[5] Alfred Auslender and Marc Teboulle. Lagrangian duality and related multiplier methods for variational inequality problems. *SIAM Journal on Optimization*, 10(4):1097–1115, 2000.

[6] Heinz H Bauschke, Patrick L Combettes, and D Russell Luke. Finding best approximation pairs relative to two closed convex sets in hilbert spaces. *Journal of Approximation theory*, 127(2):178–192, 2004.

[7] Heinz H. Bauschke and Patrick Louis Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2nd edition, 2011.

[8] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[9] Thomas Beck. Automatic differentiation of iterative processes. *Journal of Computational and Applied Mathematics*, 50(1-3):109–118, 1994.

[10] Aaron Berk, Simone Brugiapaglia, and Tim Hoheisel. Lasso reloaded: a variational analysis perspective with applications to compressed sensing. *arXiv preprint arXiv:2205.06872*, 2022.

[11] 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.

[12] Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

[13] Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. *arXiv preprint arXiv:2105.15183*, 2021.

[14] Lea Bogensperger, Antonin Chambolle, and Thomas Pock. Convergence of a Piggyback-style method for the differentiation of solutions of standard saddle-point problems. working paper or preprint, January 2022.

[15] Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

[16] Jérôme Bolte, Tam Le, Edouard Pauwels, and Antonio Silveti-Falls. Nonsmooth implicit differentiation for machine learning and optimization. *arXiv preprint arXiv:2106.04350*, 2021.

[17] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

[18] Jerome Bolte and Edouard Pauwels. A mathematical model for automatic differentiation in machine learning. *arXiv preprint arXiv:2006.02080*, 2020.

[19] Jérôme Bolte, Edouard Pauwels, and Samuel Vaiter. Automatic differentiation of nonsmooth iterative algorithms. *arXiv preprint arXiv:2206.00457*, 2022.

[20] Jonathan M Borwein. Fifty years of maximal monotonicity. *Optimization Letters*, 4(4):473–490, 2010.

[21] Radu Ioan Boţ and Ernö Robert Csetnek. An inertial forward-backward-forward primal-dual splitting algorithm for solving monotone inclusion problems. *Numerical Algorithms*, 71(3):519–540, 2016.

[22] Radu Ioan Boţ, Ernö Robert Csetnek, and Christopher Hendrich. Inertial douglas–rachford splitting for monotone inclusion problems. *Applied Mathematics and Computation*, 256:472–487, 2015.

[23] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, et al. Jax: composable transformations of python+ numpy programs. *Version 0.2*, 5:14–24, 2018.

[24] Antonin Chambolle and Juan Pablo Contreras. Accelerated bregman primal-dual methods applied to optimal transport and wasserstein barycenter problems. *arXiv preprint arXiv:2203.00802*, 2022.

[25] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

[26] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1):253–287, 2016.

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

[28] Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[29] Patrick L Combettes*. Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization*, 53(5-6):475–504, 2004.

[30] Patrick L Combettes and Valérie R Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale modeling & simulation*, 4(4):1168–1200, 2005.

[31] Laurent Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of optimization theory and applications*, 158(2):460–479, 2013.

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

[33] Iain J Day. On the inversion of diffusion nmr data: Tikhonov regularization and optimal choice of the regularization parameter. *Journal of Magnetic Resonance*, 211(2):178–185, 2011.

[34] Jalal M. Fadili, Jérôme Malick, and Gabriel Peyré. Sensitivity Analysis for Mirror-Stratifiable Convex Functions. *SIAM Journal on Optimization*, 28(4):2975–3000, 2018.

[35] Mohamed-Jalal Fadili and J-L Starck. Monotone operator splitting for optimization problems in sparse recovery. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 1461–1464. IEEE, 2009.

[36] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[37] Guillaume Garrigos, Lorenzo Rosasco, and Silvia Villa. Convergence of the forward-backward algorithm: beyond the worst-case with the help of geometry. *Mathematical Programming*, pages 1–60, 2022.

[38] Hashem Ghanem, Joseph Salmon, Nicolas Keriven, and Samuel Vaiter. Supervised learning of analysis-sparsity priors with automatic differentiation. *arXiv preprint arXiv:2112.07990*, 2021.

[39] Jean Charles Gilbert. Automatic differentiation and iterative processes. *Optimization methods and software*, 1(1):13–21, 1992.

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

[41] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm for general convex-concave saddle point problems. *arXiv preprint arXiv:1803.01401*, 2, 2018.

[42] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021.

[43] J. B. Hiriart-Urruty. Tangent cones, generalized gradients and mathematical programming in banach spaces. *Mathematics of Operations Research*, 4(1):79–97, 1979.

[44] J Ye Jane. Necessary and sufficient optimality conditions for mathematical programs with equilibrium constraints. *Journal of Mathematical Analysis and Applications*, 307(1):350–369, 2005.

[45] Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*, 21, 2008.

[46] Alan J King and Ralph Tyrrell Rockafellar. Sensitivity analysis for nonsmooth generalized equations. *Mathematical Programming*, 55(2):193–212, 1992.

[47] Adam B Levy and Ralph Tyrrell Rockafellar. Sentitivity analysis of solutions to generalized equations. *Transactions of the American Mathematical Society*, 345(2):661–671, 1994.

[48] Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Local linear convergence analysis of primal–dual splitting methods. *Optimization*, 67(6):821–853, 2018.

[49] Pierre-Louis Lions and Bertrand Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[50] Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[51] Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, pages 1540–1552. PMLR, 2020.

[52] Jérôme Malick and Hristo S Sendov. Clarke generalized jacobian of the projection onto the cone of positive semidefinite matrices. *Set-Valued Analysis*, 14(3):273–293, 2006.

[53] Yura Malitsky and Matthew K Tam. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472, 2020.

[54] 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.

[55] Sheheryar Mehmood and Peter Ochs. Fixed-point automatic differentiation of forward–backward splitting algorithms for partly smooth functions. *arXiv preprint arXiv:2208.03107*, 2022.

[56] Peter Ochs, René Ranftl, Thomas Brox, and Thomas Pock. Bilevel optimization with nonsmooth lower level problems. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 654–665. Springer, 2015.

[57] Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. *Advances in Neural Information Processing Systems*, 29, 2016.

[58] Jong-Shi Pang. Error bounds in mathematical programming. *Mathematical Programming*, 79(1):299–332, 1997.

[59] 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.

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

[61] Gabriel Peyré and Jalal M Fadili. Learning analysis sparsity priors. In *Sampta'11*, pages 4–pp, 2011.

[62] René A Poliquin and R Tyrrell Rockafellar. Generalized hessian properties of regularized nonsmooth functions. *SIAM Journal on Optimization*, 6(4):1121–1137, 1996.

[63] Saiprasad Ravishankar and Yoram Bresler. Sparsifying transform learning with efficient optimal updates and convergence guarantees. *IEEE Transactions on Signal Processing*, 63(9):2389–2404, 2015.

[64] Erlend Skaldehaug Riis. *Geometric numerical integration for optimisation*. PhD thesis, University of Cambridge, 2020.

[65] Stephen M. Robinson. *Generalized equations and their solutions, Part I: Basic theory*, pages 128–141. Springer Berlin Heidelberg, Berlin, Heidelberg, 1979.

[66] Stephen M Robinson. Strongly regular generalized equations. *Mathematics of Operations Research*, 5(1):43–62, 1980.

[67] Stephen M Robinson. Generalized equations. In *Mathematical Programming The State of the Art*, pages 346–367. Springer, 1983.

[68] R Tyrrell Rockafellar. Proto-differentiability of set-valued mappings and its applications in optimization. In *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, volume 6, pages 449–482. Elsevier, 1989.

[69] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer Verlag, Heidelberg, Berlin, New York, 1998.

[70] Ralph Rockafellar. On the maximal monotonicity of subdifferential mappings. *Pacific Journal of Mathematics*, 33(1):209–216, 1970.

[71] Alexander Shapiro. Sensitivity analysis of generalized equations. *Journal of Mathematical Sciences*, 115(4), 2003.

[72] Antonio Silveti-Falls, Cesare Molinari, and Jalal Fadili. A stochastic bregman primal-dual splitting algorithm for composite optimization. *arXiv preprint arXiv:2112.11928*, 2021.

[73] Defeng Sun. The strong second-order sufficient condition and constraint nondegeneracy in nonlinear semidefinite programming and their implications. *Mathematics of Operations Research*, 31(4):761–776, 2006.

[74] Samuel Vaiter, Charles Deledalle, Jalal Fadili, Gabriel Peyré, and Charles Dossal. The degrees of freedom of partly smooth regularizers. *Annals of the Institute of Statistical Mathematics*, 69(4):791–832, 2017.

[75] Gerd Wachsmuth. From resolvents to generalized equations and quasi-variational inequalities: existence and differentiability. *Journal of Nonsmooth Analysis and Optimization*, Volume 3, January 2022.

[76] Xiangfeng Wang, Jane J Ye, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. Perturbation techniques for convergence analysis of proximal gradient method and other first-order algorithms via variational analysis. *Set-Valued and Variational Analysis*, 30(1):39–79, 2022.

[77] Jürgen Weese. A reliable and fast method for the solution of fredhol integral equations of the first kind based on tikhonov regularization. *Computer physics communications*, 69(1):99–111, 1992.

[78] Ezra Winston and J Zico Kolter. Monotone operator equilibrium networks. *Advances in neural information processing systems*, 33:10718–10728, 2020.