Differentiating Minimal-Norm Solutions to Parametric Optimization Problems

Baptiste Plaquevent-Jourdain BAPTISTE.PLAQUEVENT-JOURDAIN@USHERBROOKE.CA
CVN, OPIS
CentraleSupélec
Gif-sur-Yvette, 91190, France
Postdoctoral researcher, corresponding author

Jalal Fadili JALAL.FADILI@ENSICAEN.FR
GREYC CNRS
ENSICAEN
Caen, 14000, France

Antonio Silveti-Falls TONYS.FALLS@GMAIL.COM
CVN, OPIS
CentraleSupélec
Gif-sur-Yvette, 91190, France

Abstract

Differentiating through parametric optimization problems is central to bilevel programming and meta-learning, often accomplished using approximate implicit differentiation. The implicit function theorem requires inverting a partial Jacobian of the optimality condition, which fails when there are many solutions. Nonetheless, in such cases it is possible to relax invertibility to a strictly weaker uniform range condition, under which it is shown that the minimal-norm solution mapping admits generalized derivatives by using a limiting Tikhonov regularization argument and conservative set-valued field theory. With additional control on the eigenvalues of the generalized Hessians, a pseudoinverse formula is justified. This is established for a class of smooth convex objectives and extended to nonsmooth composite problems. These assumptions are verified for Least-Squares, Huber regression and LASSO. The resulting extension of nonsmooth implicit differentiation to ill-posed settings is examined experimentally on data transformation problems.

Keywords: bilevel programming, implicit differentiation, conservative mappings, minimal-norm solutions, Tikhonov regularization

1 Introduction

Parametric optimization is pervasive in modern machine learning, including bilevel programming problems. Examples include hyperparameter tuning (Pedregosa, 2016; Lorraine et al., 2020; Bertrand et al., 2020), dictionary learning (Peyré and Fadili, 2011; Mairal et al., 2012), meta-learning (Franceschi et al., 2018; Engstrom et al., 2025), data transformations (cleaning, poisoning, etc) (Wang et al., 2018; Blondel et al., 2021), model selection in statistics (Vaiter et al., 2015, 2017), and deep learning (Winston and Kolter, 2020). All of these can be cast

© Baptiste Plaquevent-Jourdain, Jalal Fadili and Antonio Silveti-Falls.
License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/.

into the following template bilevel optimization problem, which we consider in this paper:

$$\min_{\theta\in\Theta} \mathcal{L}(x^\star(\theta),\theta) \tag{Q}$$

$$\text{s.t.} \quad x^\star(\theta) \in \mathcal{S}(\theta) := \operatorname*{argmin}_{x\in\mathbb{R}^n} f(x,\theta). \tag{P}$$

Here $\theta \in \Theta \subset \mathbb{R}^m$ is the *upper-level* variable and x is the *lower-level* variable. We work under the standing assumptions:

(Reg) [Regularity] For every $\theta \in \Theta$, $f(\cdot,\theta)$ is $\mathcal{C}^{1,1}(\mathbb{R}^n)$, convex, and $\mathcal{S}(\theta) \neq \emptyset$. Furthermore, $F := \nabla_x f$ is locally Lipschitz continuous jointly in (x,θ).

(D) [Definability] The function f and the set Θ are *definable*. By definability of f, $F := \nabla_x f$ is also definable. Since F is assumed locally Lipschitz, it is path-differentiable.

Throughout, definability is understood relative to some fixed o-minimal geometry (van den Dries and Miller, 1996; Coste, 1999) and will be used in all of our arguments. We also give a short overview to this theory in Section C.

First-order methods for (Q)-(P) often require differentiating a selection of the lower-level solution map $\mathcal{S}$, this differentiation being the main question discussed in this paper. Under sufficient regularity assumptions and strong convexity of the lower-level objective f, the implicit function theorem (IFT) applied to the first-order optimality condition $F(x,\theta) = 0$ gives a Jacobian of the solution map $\theta \mapsto x^\star(\theta) = \mathcal{S}(\theta)$ in closed-form as $-H^{-1}M$, where $(H\ M) \in J_F(x^\star(\theta),\theta)$ is an element of the generalized Jacobian of F (Bolte et al., 2021, 2024).[1] The nonsmooth IFT given in Bolte et al. (2021) justifies this formula with conservative Jacobians (Bolte and Pauwels, 2021) but it still requires invertibility of the H matrices.

When the lower-level problem is ill-posed and admits multiple solutions, such as when $\mathcal{S}(\theta)$ is not a singleton, H is no longer guaranteed to be invertible and the formula $-H^{-1}M$ can fail. This is a common occurrence in practice, especially for overparameterized problems (Vicol et al., 2022). Even a simple Least-Squares problem with $f(x,\theta) = \frac{1}{2}\|Ax-\theta\|_2^2$ for $A \in \mathbb{R}^{m\times n}$ with $m < n$ has a nontrivial affine subspace of solutions; the same degeneracy arises in Huber regression with a fat design matrix and in the LASSO (Tibshirani, 1996, 2011) with redundant features.

In practice, invertibility is treated as an afterthought despite some evidence showing that its absence could lead to serious training instabilities (Bolte et al., 2021, Section 5). A solution seen in practice is to apply the pseudoinverse $H^\dagger$ if H is singular (Vicol et al., 2022). Our goal is to rigorously study aspects of these practical approaches by connecting them to the *minimal-norm solution* $x^\star_{\min}(\theta) := \operatorname{argmin}\{\|x\|_2 : x \in \mathcal{S}(\theta)\}$, which is selected by *Tikhonov regularization* (Tikhonov, 1963). Namely, for $\varepsilon > 0$, define the ε-regularized problem

$$x^\star_\varepsilon(\theta) := \operatorname*{argmin}_{x\in\mathbb{R}^n}\{f(x,\theta) + \tfrac{\varepsilon}{2}\|x\|_2^2\}. \tag{P_ε}$$

Classical results (Browder, 1966, 1967) guarantee that $x^\star_\varepsilon(\theta) \to x^\star_{\min}(\theta)$ as $\varepsilon \to 0^+$. Since $f(\cdot,\theta)$ is convex for any θ, then for each fixed $\varepsilon > 0$, the regularized objective in (P_ε) is strongly convex with (generalized) gradient F_ε. Hence nonsmooth implicit differentiation (Bolte et al., 2024) gives a conservative Jacobian of $x^\star_\varepsilon$ of the form $J_{x^\star_\varepsilon}(\theta) = \{-(H_\varepsilon + \varepsilon I)^{-1}M_\varepsilon \colon (H_\varepsilon\ M_\varepsilon) \in J_F(x^\star_\varepsilon(\theta),\theta)\}$. The central question is therefore

1. We pick the letters H and M to represent the Hessian and Mixed derivative, since $F(x,\theta) = \nabla_x f(x,\theta)$, in parallel with the smooth case.

Do these regularized conservative Jacobians have a meaningful limit as $\varepsilon \to 0^+$, and how does this limit relate to $x^\star_{\min}$ (cf. Figure 1) ?

We answer this question affirmatively using recent results in the theory of path-differentiable functions (Schechtman, 2026), which give a precise characterization of limits of gradients and conservative mappings for definable mappings. Our results show that invertibility can be relaxed to a uniform *range condition*, $M_\varepsilon \in \operatorname{range}(H_\varepsilon)$ around the solution, together with a uniform condition on the eigenvalues of the H_ε. Under these assumptions, the Tikhonov implicit derivatives converge to $-H_0^\dagger M_0$ as $\varepsilon \to 0^+$ for some pairs $(H_0, M_0) \in J_F(x^\star_{\min}(\theta), \theta)$, and this limit defines an element of a conservative mapping of the minimal-norm solution $x^\star_{\min}(\theta)$. In turn, this element can be used in first-order methods for the upper-level. This extends implicit differentiation using conservative calculus from well-posed lower-level problems to ill-posed convex problems with many solutions. One could also consider the anchored term $\varepsilon\|x - \hat{x}\|^2/2$ or other viscosity-type regularizers.

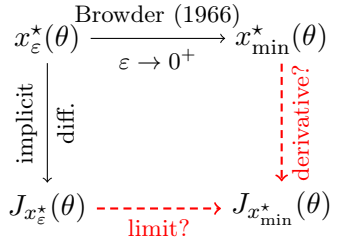


Figure 1: Solid arrows are known; dashed arrows are studied in this work. Standard IFT applies to the Tikhonov regularized solution $x^\star_\varepsilon$ but fails in general if H is singular. Under suitable assumptions, we show that the cluster points of the conservative Jacobians along the bottom path are conservative for the minimal-norm solution, justifying in addition the formula $-H^\dagger M \in J_{x^\star_{\min}}(\theta)$ for $x^\star_{\min}$ along the right path when the blocks H_ε have controlled eigenvalues.

The range condition is a strict relaxation of the usual invertibility condition required since, if the matrices H_0 are invertible, so are the matrices H_ε for ε small enough thus $M_\varepsilon \in \operatorname{range}(H_\varepsilon)$ (by continuity of the determinant). The eigenvalues condition is not necessary for path-differentiability but only for the pseudoinverse formula, which must be corrected without this condition. In general, the matrices H_ε shall be symmetric, which simplifies the statement of the eigenvalue condition: for functions that are twice differentiable (even without continuity of their Hessians), the Hessian is symmetric (Dieudonné, 1969); for nonsmooth functions, every matrix in the Clarke Jacobian of a proximal operator is also symmetric and positive semidefinite, (Patrinos et al., 2014, Theorem 3.2, see Theorem 35).

1.1 Main Contributions

- We show path-differentiability of the minimal-norm solution to a parametric convex optimization problem: under definability and local boundedness of the regularized implicit Jacobians, the limits of sequences of regularized conservative Jacobians are conservative for the minimal-norm solution.
- Under a uniform range condition on the second-order derivatives, we justify local boundedness and thus path-differentiability of the minimal-norm solution. This condition is verified on Least-Squares, Huber regression, and LASSO.

- If in addition an eigenvalue condition holds, we justify an explicit formula for the implicit Jacobians in terms of the pseudoinverse of the Hessian. In general, a residual term appears resulting from vanishing curvature alongside the trajectory – even in a smooth setting with unique solution.
- We experimentally verify our work with problems involving ill-posed lower-level objectives. The benefits of differentiating the minimal-norm solution are showcased on data poisoning experiments with Huber loss and LASSO lower-level costs, displaying the relevance of the minimal-norm solution compared to guaranteed nonminimal solutions. Finally, we consider an experiment illustrating some of the theoretical hypotheses.

1.2 Related Work

Differentiating solutions is intrinsically related to sensitivity analysis (Bonnans and Shapiro, 2000). Nonsmooth versions of the IFT first appear in Clarke (1990); Robinson (1991); Dontchev and Rockafellar (2009). In Adly and Rockafellar (2021), one-dimensional parametric variational inclusions are considered via graphical convergence, semi- and proto-differentiability. The role of (Tikhonov) regularization is also discussed in Attouch (1996). In Arbel and Mairal (2022) the authors introduce a parametric Morse-Bott property (see also Bolte et al., 2026b) which implies that f behaves quadratically around critical points. This is a related property as it implies some range condition (Arbel and Mairal, 2022, Proposition 6) though it captures the more general case where the lower-level problem is nonconvex, in which case Tikhonov regularization and Browder convergence cannot be applied. Recently, in Masiha et al. (2026), the authors consider a similar setting under a Polyak-Łojasiewicz assumption (uniform in θ) alongside third-order smoothness and nondegeneracy. This allows them to differentiate an optimistic selection mapping and to obtain formulas also employing a kind of pseudoinverse. Another approach to ensure finding the minimal-norm solution is to use vanishing Tikhonov regularization ε_k with a first-order solver. This technique was studied for unrolled differentiation in Mehmood and Ochs (2024). Closest to our work is Vicol et al. (2022), which ensures convergence in a setting with quadratic lower and upper objectives. Despite this limitation, they validate the behavior of several algorithms on multiple problems. A related contribution is Pauwels and Vaiter (2023), which showed the convergence of derivatives of the Sinkhorn algorithm in optimal transport without contractivity by using the spectral inverse (a refined inverse identical to pseudoinverse for symmetric matrices). In Bambade et al. (2024), the authors differentiate the solutions of a quadratic problem with parametric cost and constraints using Lagrangian-based techniques and minimal-norm solutions. We also mention that there are many works that use value functions or adjoint methods to transform the bilevel problem into a single-loop problem, for example Suonperä and Valkonen (2026).

The paper is organized as follows. In Section 2, we present the main theoretical tools we employ, conservative Jacobians and nonsmooth implicit function theorem, and how to apply them to ill-posed settings via Tikhonov regularization, see also Section C. Then, Section 3 presents our framework on smooth convex objectives, some of the proofs being in Section D. Under stronger assumptions, we present extensions for nonsmooth composite problems in Section 4, completed in Section E. Section 4.4 briefly discusses how regularization and accuracy counterbalance each other. Finally, Section 5 presents experimental results, with additional details for reproducibility given in Section A. The Supplementary Material gives a

dicontinuous minimal-norm solution example (Section SM1), the Huber duality and solution-set derivation (Section SM2), limit calculations for three different regimes (Section SM3), and quantitative regularization and iteration bounds (Section SM4).

Main Notation For a matrix $A \in \mathbb{R}^{m\times n}$, the transpose of A is denoted by $A^\mathsf{T} \in \mathbb{R}^{n\times m}$, its null space by $\ker(A)$, its range space by $\mathrm{range}(A)$ and its pseudoinverse (Moore, 1919; Penrose, 1955) by $A^\dagger \in \mathbb{R}^{n\times m}$. $A \in \mathbb{R}^{n\times n}$ is orthogonal if $A^\mathsf{T}A = I_n = AA^\mathsf{T}$ with I_n (or I) the identity matrix of size $n \times n$. A matrix A that is positive semidefinite (PSD) will be denoted by $A \succeq 0$. The spectrum of a square matrix is denoted spec. For $q \in \mathbb{R}^n$, $\mathrm{diag}(q) \in \mathbb{R}^{n\times n}$ denotes the diagonal matrix whose entries are equal to q_i. For A and B with suitable dimensions, $(A\ B)$ is the horizontal concatenation of A and B, $(A; B)$ or $(A^\mathsf{T}\ B^\mathsf{T})^\mathsf{T}$ the vertical one. Set-valued mappings are denoted by $\rightrightarrows$. A function $F : \mathbb{R}^n \to \mathbb{R}^m$ is *locally Lipschitz* if, for each $x \in \mathbb{R}^n$, there exists a neighborhood $\mathcal{U}$ of x such that F is Lipschitz on $\mathcal{U}$. The set of points at which F is differentiable is denoted by $\mathrm{Diff}(F)$; at these points, $\nabla F(x) \in \mathbb{R}^{n\times m}$ is the gradient of F and $\mathrm{Jac}F(x) = \nabla F(x)^\mathsf{T}$ its Jacobian. The class of extended real-valued convex, lower semicontinuous, and proper functions on $\mathbb{R}^n$ is denoted $\Gamma_0(\mathbb{R}^n)$. For $f \in \Gamma_0(\mathbb{R}^n)$, $\mathrm{prox}_f(x) := \mathrm{argmin}_y\{f(y) + \|y - x\|_2^2/2\}$ is the proximal operator of f (Moreau, 1965; Bauschke and Combettes, 2017). We refer to Section B for additional information.

2 Implicit Differentiation and Conservative Mappings

This section details the main analytical tools used in the paper, which are the conservative implicit function theorem (Bolte et al., 2021) and a recent result from Schechtman (2026) about limits of path-differentiable functions. We combine these in Theorem 7 to give a sufficient condition that ensures the path-differentiability of the minimal-norm solution.

2.1 Background

The notions of *conservative Jacobian* (used interchangeably with *conservative mapping*) and *path-differentiable function* were introduced in the seminal paper of Bolte and Pauwels (2021) as a way to faithfully model automatic differentiation.

Definition 1 (Conservative mappings (Bolte et al., 2022, section 2))**.** Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be locally Lipschitz, then $J_F : \mathbb{R}^n \rightrightarrows \mathbb{R}^{m\times n}$ is a *conservative Jacobian* for the *path-differentiable* function F if J_F is never empty, is locally bounded, has a closed graph, and satisfies, for any absolutely continuous curve $\mu : [0, 1] \to \mathbb{R}^n$, for almost all $t \in [0, 1]$, for all $V \in J_F(\mu(t))$, $\frac{\mathrm{d}}{\mathrm{d}t}F(\mu(t)) = V\dot{\mu}(t)$.

A conservative mapping appears as a generalized derivative of a locally Lipschitz function along absolutely continuous curves and generalizes the Clarke Jacobian (Clarke, 1990) $\partial F(x) := \mathrm{conv}\{\lim_{y\to x} \mathrm{Jac}F(y) : y \in \mathrm{Diff}(F)\}$, where Diff is the differentiable domain of F, the set of points at which it is differentiable.

Contrary to the Clarke Jacobian (see Bolte et al., 2021, Example 1), conservative mappings enjoy exact calculus rules such as sum and composition, which notably justifies automatic differentiation: its outputs are elements of a certain conservative Jacobian of the solution mapping (see Bolte and Pauwels, 2021, section 5). For convex conservative mappings, the Clarke Jacobian is a minimal conservative mapping (Bolte and Pauwels, 2021, Corollary 1).

Almost everywhere, $J_F(x) = \{\mathrm{Jac}F(x)\} = \partial F(x)$ by Rademacher's theorem since F is locally Lipschitz (Rademacher, 1919). If J_F takes convex values, $\partial F(x) \subset J_F(x)$; conservative mappings are preserved by pointwise convexification but need not be convex in general.

Functions that have a conservative Jacobian are said to be *path-differentiable*. Important examples include locally Lipschitz convex or concave functions and, crucially for this paper, locally Lipschitz definable mappings on open definable domains (Davis et al., 2020; Bolte and Pauwels, 2021). Here definability is relative to some o-minimal structure fixed throughout the paper; see Section C. Semialgebraic problems are therefore covered as a special case. The main property of conservative mappings employed in this paper is a nonsmooth IFT, which requires the existence of a locally Lipschitz implicit function. Another version (Bolte et al., 2021, Corollary 1) proves this existence but requires the conservative mapping to be convex.

Theorem 2 (Path-differentiable IFT (Bolte et al., 2021))**.** *Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be path-differentiable on an open set $\mathcal{U} \times \mathcal{V} \subset \mathbb{R}^n \times \mathbb{R}^m$ and $G : \mathcal{V} \to \mathcal{U}$ a locally Lipschitz function such that, $\forall\ y \in \mathcal{V}$, $F(G(y), y) = 0$. Furthermore, assume that $\forall\ y \in \mathcal{V}$, $\forall\ (H\ M) \in J_F(G(y), y)$ the matrix H is invertible where J_F is a conservative Jacobian for F. Then, G is path-differentiable with conservative Jacobian given, $\forall\ y \in \mathcal{V}$, by $J_G : y \rightrightarrows \{-H^{-1}M : (H\ M) \in J_F(G(y), y)\}$.*

For strongly convex problems, it is possible to proceed as in Bolte et al. (2024). Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ such that, for each $\theta \in \Theta \subset \mathbb{R}^m$, $f(\cdot, \theta)$ is definable, $\mathcal{C}^{1,1}$ and strongly convex. Consider

$$\theta \mapsto x^\star(\theta) := \operatorname*{argmin}_{x \in \mathbb{R}^n} f(x, \theta). \tag{1}$$

For fixed θ, this strongly convex problem has a unique solution $x^\star(\theta)$ verifying the optimality condition $F(x^\star(\theta), \theta) = 0$ where $F = \nabla_x f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$. Then, the H-blocks of elements $(H\ M) \in J_F(x^\star(\theta), \theta)$ are invertible because they are generalized Hessians of a strongly convex $\mathcal{C}^{1,1}$ function — assuming we take the Clarke Jacobian or a suitable conservative field. A way to proceed (the strategy of Bolte et al., 2024) is to convexify the conservative Jacobian J_F, apply Corollary 1 of Bolte et al. (2021), which gives existence of the implicit mapping, and then to apply Theorem 2 using the given implicit mapping with the nonconvex conservative mapping.

In this paper, we concern ourselves with ill-posed problems that are not strongly convex (the matrices H can be singular) such as Least-Squares (Fahrmeir et al., 2013) or LASSO (Tibshirani, 1996; Osborne et al., 2000; Tibshirani, 2013). The approach we follow is to use Tikhonov regularization as in (P_ε) to circumvent ill-posedness. While the convergence of the solutions $x^\star_\varepsilon(\theta) \to x^\star_{\min}(\theta)$ is well-known (Tikhonov, 1963; Browder, 1966, 1967), the convergence of generalized derivatives studied in this paper is far more intricate and requires much more care (see Figure 1).

2.2 Limits of Conservative Jacobians

To pass from the regularized solutions to $x^\star_{\min}$, we use a limiting result for definable conservative Jacobians from Schechtman (2026). We define $x^\star$ on $\Theta \times \mathbb{R}_+$ by $x^\star(\theta, \varepsilon) := x^\star_\varepsilon(\theta)$ if $\varepsilon > 0$ and $x^\star(\theta, \varepsilon) := x^\star_{\min}(\theta)$ if $\varepsilon = 0$, with $x^\star_{\min}$ the minimal-norm solution mapping of (P). Given an open set $O \subset \Theta$, for every $\varepsilon > 0$ let $J_{x^\star_\varepsilon} : O \rightrightarrows \mathbb{R}^{n \times m}$ be a chosen conservative Jacobian of $\theta \mapsto x^\star_\varepsilon(\theta)$ on O. We collect these mappings in the family $J_{x^\star}(\theta, \varepsilon) := J_{x^\star_\varepsilon}(\theta)$ for $(\theta, \varepsilon) \in O \times \mathbb{R}^*_+$. We emphasize that ε is only an index and $J_{x^\star}(\theta, \varepsilon)$ contains derivatives

with respect to θ only, not with respect to ε (which is how they are used by Schechtman (2026)). For $\theta \in O$, define

$$J_{x^\star_{\min}}(\theta) := \left\{ V \in \mathbb{R}^{n\times m} : \begin{array}{l} \exists\ \theta_k \in O,\ \theta_k \to \theta,\ \varepsilon_k \downarrow 0,\ V_k \to V \text{ with} \\ V_k \in J_{x^\star}(\theta_k, \varepsilon_k), \text{and } x^\star(\theta_k, \varepsilon_k) \to x^\star_{\min}(\theta) \end{array} \right\}. \tag{2}$$

Theorem 3 (Limit properties (Schechtman, 2026, Lemma 4.2, Theorem 4.3 and Corollary 4.5))**.** *Fix $\bar{\varepsilon} > 0$. Suppose that (i) the restrictions of $x^\star$ and $J_{x^\star}$ to $O \times [0, \bar{\varepsilon}]$ and $O \times]0, \bar{\varepsilon}]$, respectively, are definable, and (ii) for every $0 < \varepsilon \leq \bar{\varepsilon}$, $J_{x^\star_\varepsilon}$ is a conservative mapping of $x^\star_\varepsilon$ on O. Suppose in addition that, for every compact set $K \subset O$,*

$$(iii) \quad \sup_{\theta\in K} \|x^\star(\theta,\varepsilon) - x^\star_{\min}(\theta)\| \underset{\varepsilon\to 0}{\to} 0,$$

$$(iv) \quad \limsup_{\varepsilon_0 \downarrow 0} \sup\{\|V\| : V \in J_{x^\star_\varepsilon}(\theta),\ 0 < \varepsilon \leq \varepsilon_0,\ \theta \in K\} < \infty.$$

Then $x^\star_{\min}$ and $J_{x^\star_{\min}}$ are definable on O, and $J_{x^\star_{\min}}$ is a conservative mapping of $x^\star_{\min}$ on O.

In Theorem 3, (iii) is a form of continuity, while (iv) relates to local boundedness of the derivatives which must hold for conservative mappings. As we will show later in Theorem 5, the setting we consider guarantees that $(iv) \implies (iii)$, so that we may focus solely on the boundedness of the sequence of elements coming from the conservative mappings.

3 Problems with $\mathcal{C}^{1,1}(\mathbb{R}^n)$ Objectives

We study in this section the conservative Jacobians of minimal-norm solutions to problems with $\mathcal{C}^{1,1}$ lower-level objectives. We first introduce the general setting and illustrate it concretely on Least-Squares; this example motivates the uniform range condition, after which we study the pseudoinverse formula and apply the results to Huber regression.

We consider problem (P) under Assumptions **(Reg)** and **(D)** on an open definable set $O \subset \Theta$. Let $F := \nabla_x f$, since $f(\cdot, \theta)$ is convex, the solutions of (P) are characterized by $F(x, \theta) = 0$. For $\varepsilon > 0$, define $F_\varepsilon(x, \theta) := F(x, \theta) + \varepsilon x = \nabla_x f(x, \theta) + \varepsilon x$. The regularized objective in (P_ε) is strongly convex and therefore has a unique solution $x^\star_\varepsilon(\theta)$, characterized by $F_\varepsilon(x^\star_\varepsilon(\theta), \theta) = 0$. Let J_F be a definable conservative Jacobian of F. At the regularized solution $x^\star_\varepsilon(\theta)$, write $(H_\varepsilon\ M_\varepsilon) \in J_F(x^\star_\varepsilon(\theta), \theta)$. Although F has no explicit dependence on ε, the blocks are evaluated at $x^\star_\varepsilon(\theta)$, and both may therefore vary along the regularized trajectory. We thus take the following conservative Jacobian for F_ε, evaluated along the regularized solution: $J_{F_\varepsilon}(x^\star_\varepsilon(\theta), \theta) := \{(H_\varepsilon + \varepsilon I_n\ M_\varepsilon) : (H_\varepsilon\ M_\varepsilon) \in J_F(x^\star_\varepsilon(\theta), \theta)\}$. When $J_F = \partial F$, the convexity of $f(\cdot, \theta)$ guarantees that every H_ε is symmetric PSD, since these x-blocks are generalized Hessians of f. If we use a more general conservative Jacobian, we will require its x-blocks to be symmetric PSD. In either case, $H_\varepsilon + \varepsilon I_n$ is symmetric positive definite. Theorem 2 then gives the conservative Jacobian

$$J_{x^\star_\varepsilon}(\theta) = \{-(H_\varepsilon + \varepsilon I_n)^{-1} M_\varepsilon \colon (H_\varepsilon\ M_\varepsilon) \in J_F(x^\star_\varepsilon(\theta), \theta)\}. \tag{3}$$

For the remainder of this section, $J_{x^\star_\varepsilon}$ denotes this implicit conservative Jacobian and $J_{x^\star}(\theta, \varepsilon) := J_{x^\star_\varepsilon}(\theta)$ denotes the associated family. To apply Theorem 3 and conclude about $x^\star_{\min}$, we do not need every selection in (3) to converge. What we need is a local uniform bound: for every compact set $K \subset O$, there must exist $C_K > 0$ and $\varepsilon_K > 0$ such that

$$\sup \left\{ \| (H_\varepsilon + \varepsilon I_n)^{-1} M_\varepsilon \| : \begin{array}{l} \theta \in K, \quad 0 < \varepsilon \leq \varepsilon_K, \\ (H_\varepsilon\ M_\varepsilon) \in J_F(x^\star_\varepsilon(\theta), \theta) \end{array} \right\} \leq C_K. \tag{4}$$

3.1 A Motivating Example

We start with an example that shows how a bound like (4) can hold even when the limits are singular. Here $\Theta = O = \mathbb{R}^m$. We will obtain a bound for every $\varepsilon > 0$, so both ε_K in (4) and the global upper bound $\bar{\varepsilon}$ used below may be chosen arbitrarily and independently of K. Let $A \in \mathbb{R}^{m\times n}$ with $m < n$ and consider the Least-Squares problem

$$\min_{x\in\mathbb{R}^n} \tfrac{1}{2}\|Ax - \theta\|_2^2. \tag{LS}$$

For this problem, $F(x,\theta) = A^\mathsf{T}(Ax-\theta)$, and we take $J_F = \mathrm{Jac}F$. Along the regularized trajectory, the blocks introduced above are $H_\varepsilon \equiv A^\mathsf{T}A$ and $M_\varepsilon \equiv -A^\mathsf{T}$. The set of solutions to (LS) is $A^\dagger\theta + \ker(A)$, so its minimal-norm solution is $x^\star_{\min}(\theta) = A^\dagger\theta$. The Jacobian of $x^\star_{\min}$ is $A^\dagger$. We now recover the same formula from the Tikhonov-regularized problem. Its solution and Jacobian are $x^\star_\varepsilon(\theta) = (A^\mathsf{T}A + \varepsilon I_n)^{-1}A^\mathsf{T}\theta$ and $\mathrm{Jac}\, x^\star_\varepsilon(\theta) = (A^\mathsf{T}A + \varepsilon I_n)^{-1}A^\mathsf{T}$. Set $W_\varepsilon \equiv -A^\dagger$. The identity $A^\mathsf{T}AA^\dagger = A^\mathsf{T}$ gives $M_\varepsilon = H_\varepsilon W_\varepsilon$. Consequently, $\mathrm{Jac}\, x^\star_\varepsilon(\theta) = -(H_\varepsilon + \varepsilon I_n)^{-1}M_\varepsilon = -(H_\varepsilon + \varepsilon I_n)^{-1}H_\varepsilon W_\varepsilon$. Since $H_\varepsilon = A^\mathsf{T}A \succeq 0$, the eigenvalues of $(H_\varepsilon + \varepsilon I_n)^{-1}H_\varepsilon$ are of the form $\lambda/(\lambda+\varepsilon)$ with $\lambda \geq 0$. Hence $0 \preceq (H_\varepsilon + \varepsilon I_n)^{-1}H_\varepsilon \preceq I_n$, and therefore $\|\mathrm{Jac}\, x^\star_\varepsilon(\theta)\| \leq \|W_\varepsilon\| = \|A^\dagger\|$, uniformly in $\varepsilon > 0$. Since $H_\varepsilon \equiv A^\mathsf{T}A$ is fixed along the trajectory, $(H_\varepsilon + \varepsilon I_n)^{-1}H_\varepsilon \underset{\varepsilon\to 0}{\to} (A^\mathsf{T}A)^\dagger A^\mathsf{T}A$. It follows that $\mathrm{Jac}\, x^\star_\varepsilon(\theta) \underset{\varepsilon\to 0}{\to} -(A^\mathsf{T}A)^\dagger(-A^\mathsf{T}) = A^\dagger$. Thus the singularity of $H_\varepsilon = A^\mathsf{T}A$ is not an obstruction per se. The inverse $(H_\varepsilon + \varepsilon I_n)^{-1}$ is never applied to an element of $\ker(H_\varepsilon)$, where it would blow up as $\varepsilon \to 0$, since $M_\varepsilon \in \mathrm{range}(H_\varepsilon)$. Least-Squares is particularly nice because its blocks are fixed and do not depend on ε; in general a uniform control will be necessary.

3.2 Uniform Range Condition

The Least-Squares calculation rests on two ingredients: $H_\varepsilon \succeq 0$ and a uniformly bounded factorization $M_\varepsilon = H_\varepsilon W_\varepsilon$. This inspires the following assumption requiring both along the regularized solutions.

(R) [Uniform range condition] Let $O \subset \Theta$ be open and definable, fix $\bar{\varepsilon} > 0$, and let J_F be a definable conservative Jacobian for F. For every compact set $K \subset O$, there exists $C_K > 0$ such that, for every $\theta \in K$, $0 < \varepsilon \leq \bar{\varepsilon}$, and $(H_\varepsilon \; M_\varepsilon) \in J_F(x^\star_\varepsilon(\theta), \theta)$, the block H_ε is symmetric PSD and $M_\varepsilon = H_\varepsilon W_\varepsilon$ for some $W_\varepsilon \in \mathbb{R}^{n\times m}$ satisfying $\|W_\varepsilon\| \leq C_K$.

The factorization says that the columns of M_ε lie in $\mathrm{range}(H_\varepsilon)$, while the bound controls a choice of preimage uniformly in θ, ε, and all admissible blocks. For $J_F = \partial F$, convexity of $f(\cdot,\theta)$ ensures that the H-blocks are PSD. For a more general conservative Jacobian, symmetry and positive semidefiniteness of its x-blocks are required.

Example 4 (Assumption **(R)** fails)**.** Consider the one-dimensional problem with $f(x,\theta) = x^4/4 + \theta x$. The optimality condition is $F(x,\theta) = x^3 + \theta = 0$, so $H_\varepsilon = 3(x^\star_\varepsilon(\theta))^2$ and $M_\varepsilon = 1$. Thus, for $\theta = 0$, $x^\star_\varepsilon(0) = 0$ so $1 = M_\varepsilon \notin \mathrm{range}(H_\varepsilon) = \{0\}$. This agrees with the fact that $x^\star_{\min}(\theta) = -\theta^{1/3}$ is not locally Lipschitz at 0, thus not path-differentiable. This cost can be made $\mathcal{C}^{1,1}$ by a method similar to Example 8.

Theorem 5 (Regularity from uniformly bounded implicit Jacobians)**.** *Let $O \subset \Theta$ be open and definable and assume that* **(Reg)** *and* **(D)** *hold. For every $\varepsilon > 0$, let $J_{x^\star_\varepsilon}$ be a conservative Jacobian of $x^\star_\varepsilon$ on O. Suppose that, for every compact set $K \subset O$, there exist $C_K > 0$ and $\varepsilon_K > 0$ such that* $\sup\{\|V\| : \theta \in K,\ 0 < \varepsilon \leq \varepsilon_K,\ V \in J_{x^\star_\varepsilon}(\theta)\} \leq C_K$*. Then $x^\star_{\min}$ is locally*

Lipschitz on O, the joint selection $x^\star$ is continuous on $O \times \mathbb{R}_+$, and $x^\star_\varepsilon \to x^\star_{\min}$ uniformly on compact subsets of O as $\varepsilon \to 0$. If **(R)** *holds and $J_{x^\star_\varepsilon}$ is chosen as the implicit conservative Jacobian in* (3)*, then the bound holds with $\varepsilon_K = \bar{\varepsilon}$ and all subsequent conclusions follow.*

Proof We first verify the last claim. Assume **(R)**, take $J_{x^\star_\varepsilon}$ from (3), fix a compact set $K \subset O$, and let $\theta \in K$, $0 < \varepsilon \leq \bar{\varepsilon}$, and $V \in J_{x^\star_\varepsilon}(\theta)$. There is a block $(H_\varepsilon \; M_\varepsilon) \in J_F(x^\star_\varepsilon(\theta), \theta)$ such that $V = -(H_\varepsilon + \varepsilon I_n)^{-1} M_\varepsilon = -(H_\varepsilon + \varepsilon I_n)^{-1} H_\varepsilon W_\varepsilon$. The eigenvalues of $(H_\varepsilon + \varepsilon I_n)^{-1} H_\varepsilon$ are of the form $\lambda/(\lambda + \varepsilon) \in [0, 1]$. Hence $0 \preceq (H_\varepsilon + \varepsilon I_n)^{-1} H_\varepsilon \preceq I_n$ and $\|V\| \leq \|W_\varepsilon\| \leq C_K$ as claimed.

We now use only the bound on $J_{x^\star_\varepsilon}$. Fix $\theta_0 \in O$ and choose $r > 0$ such that the compact convex set $K_0 := \overline{B}(\theta_0, r)$ ($\overline{B}$ is the closed ball) is contained in O. For $\theta, \eta \in K_0$, let $\mu(t) = \eta + t(\theta - \eta)$. If $0 < \varepsilon \leq \varepsilon_{K_0}$, path-differentiability gives, for almost every $t \in [0, 1]$, for all $V(t) \in J_{x^\star_\varepsilon}(\mu(t))$, $\frac{d}{dt} x^\star_\varepsilon(\mu(t)) = V(t)(\theta - \eta)$. Using this and invoking the bound gives

$$\|x^\star_\varepsilon(\theta) - x^\star_\varepsilon(\eta)\| = \left\| \int_0^1 \tfrac{d}{dt} x^\star_\varepsilon(\mu(t)) dt \right\| \leq \int_0^1 \|\tfrac{d}{dt} x^\star_\varepsilon(\mu(t))\| dt \leq C_{K_0} \|\theta - \eta\|.$$

Since $x^\star_\varepsilon \to x^\star_{\min}$ as $\varepsilon \to 0$ pointwise in θ (see Theorem 36), continuity of $\|\cdot\|$ finally gives $\|x^\star_{\min}(\theta) - x^\star_{\min}(\eta)\| = \lim_{\varepsilon \to 0} \|x^\star_\varepsilon(\theta) - x^\star_\varepsilon(\eta)\| \leq C_{K_0} \|\theta - \eta\|$. So $x^\star_{\min}$ is locally Lipschitz on O.

Consider now a sequence $(\theta_k, \varepsilon_k) \to (\theta_0, 0)$ with $\varepsilon_k \geq 0$. For k large enough, θ_k and θ_0 lie in K_0, hence

$$\begin{aligned} \|x^\star(\theta_k, \varepsilon_k) - x^\star(\theta_0, 0)\| &\leq \|x^\star_{\varepsilon_k}(\theta_k) - x^\star_{\varepsilon_k}(\theta_0)\| + \|x^\star_{\varepsilon_k}(\theta_0) - x^\star_{\min}(\theta_0)\| \\ &\leq C_{K_0} \|\theta_k - \theta_0\| + \|x^\star_{\varepsilon_k}(\theta_0) - x^\star_{\min}(\theta_0)\| \to 0 \end{aligned}$$

because $\theta_k \to \theta_0$ and $x^\star(\theta_0, \varepsilon_k) \to x^\star_{\min}(\theta_0)$. This proves continuity on $O \times \{0\}$. At points with $\varepsilon > 0$, joint continuity in (θ, ε) follows from strong convexity and the implicit function theorem applied with both variables as parameters (Bolte et al., 2024). Hence $x^\star$ is continuous on $O \times \mathbb{R}_+$.

Finally, let $K \subset O$ be compact and fix $a > 0$. The restriction of $x^\star$ to the compact set $K \times [0, a]$ is uniformly continuous. Since $x^\star(\theta, 0) = x^\star_{\min}(\theta)$, it follows that $\sup_{\theta \in K} \|x^\star_\varepsilon(\theta) - x^\star_{\min}(\theta)\| \to 0$ as $\varepsilon \downarrow 0$ as desired. ■

We point out that the arguments above give joint continuity in (θ, ε), but not joint local Lipschitz continuity.

Remark 6 (Definability of the construction)**.** Proposition 40 gives the definability of $x^\star$, while Proposition 42 gives the definability of $J_{x^\star}$ and the conservativity of each fixed-ε slice $J_{x^\star_\varepsilon}$. These results verify (*i*) and (*ii*) of Theorem 3. The compact-uniform convergence in (*iii*) follows from Theorem 5, while Assumption **(R)** gives the local bound in (*iv*).

Theorem 7 (Path-differentiability under the uniform range condition)**.** *Assume that* **(Reg)**, **(D)** *and* **(R)** *hold, and let $J_{x^\star_\varepsilon}$ be the implicit conservative Jacobian in* (3) *for $0 < \varepsilon \leq \bar{\varepsilon}$. Then $x^\star_{\min}$ is path-differentiable on O. Moreover, $J_{x^\star_{\min}}$ defined in* (2) *is a definable conservative Jacobian of $x^\star_{\min}$.*

Proof By Remark 6, all the hypotheses of Theorem 3 hold for the restricted family on $O \times]0, \bar{\varepsilon}]$. Hence $J_{x^\star_{\min}}$ is a definable conservative Jacobian of $x^\star_{\min}$, which is therefore path-differentiable on O. ■

By joint continuity, the condition $x^\star(\theta_k, \varepsilon_k) \to x^\star_{\min}(\theta)$ used in (2) holds automatically in our settings. Thus $J_{x^\star_{\min}}$ can be rewritten as

$$J_{x^\star_{\min}} : \theta \rightrightarrows \left\{ V \in \mathbb{R}^{n \times m} : \begin{array}{l} \exists\ (\theta_k, \varepsilon_k, V_k)_{k \in \mathbb{N}} \text{ with } \theta_k \in O,\ 0 < \varepsilon_k \leq \bar{\varepsilon}, \\ \theta_k \to \theta,\ \varepsilon_k \downarrow 0,\ V_k \to V,\ V_k \in J_{x^\star}(\theta_k, \varepsilon_k) \end{array} \right\}.$$

3.3 Pseudoinverse Formula

Theorem 7 bounds the regularized conservative Jacobians in a way that ensures path-differentiability but does not identify their cluster points. The Least-Squares example suggests the formula $-H^\dagger M$ but, as the following example shows, this formula can fail when curvature vanishes along the regularized trajectory.

Example 8 (Slowly vanishing curvature)**.** Let $O :=]-1, 1[$ and $f(x, \theta) = \psi(x - \theta)$, where $\psi(t) := t^4/4$ if $|t| \leq 1$ and $\psi(t) := |t| - 3/4$ if $|t| > 1$. The function ψ is convex, semialgebraic, and belongs to $\mathcal{C}^{1,1}(\mathbb{R})$. Its unique minimizer is zero, so $x^\star_{\min}(\theta) = \theta$ and $\partial x^\star_{\min}/\partial\theta \equiv 1$. At the unregularized solution, however, the two blocks H and M are both 0 (since $\theta \in O =]-1, +1[$), and hence $-H^\dagger M = 0$.

Fix $\theta \neq 0$. The regularized solution lies between 0 and θ, so $|x^\star_\varepsilon(\theta) - \theta| < 1$ and the quartic branch of ψ is active. The optimality condition is $(x^\star_\varepsilon(\theta) - \theta)^3 + \varepsilon x^\star_\varepsilon(\theta) = 0$ and the corresponding blocks are $H_\varepsilon = 3(x^\star_\varepsilon(\theta) - \theta)^2$, $M_\varepsilon = -H_\varepsilon$, and $W_\varepsilon = -1$. Rearranging the optimality condition and dividing by ε gives $(\frac{x^\star_\varepsilon(\theta)-\theta}{\varepsilon^{1/3}})^3 = -x^\star_\varepsilon(\theta)$. Since $x^\star_\varepsilon(\theta) \to \theta$ and the real cube-root function is continuous, it follows that $\frac{x^\star_\varepsilon(\theta)-\theta}{\varepsilon^{1/3}} \to -\sqrt[3]{\theta}$ or, equivalently, $x^\star_\varepsilon(\theta) - \theta \sim -\sqrt[3]{\theta}\,\varepsilon^{1/3}$. Therefore, $H_\varepsilon \sim 3|\theta|^{2/3}\varepsilon^{2/3}$ and $-(H_\varepsilon + \varepsilon)^{-1}M_\varepsilon \to 1$. Thus the regularized derivatives recover the derivative of $x^\star_{\min}$, although the limiting blocks alone give 0. The mismatch is due to H_ε vanishing more slowly than ε.

In general, consider a sequence $\varepsilon_k \downarrow 0$ and convergent selected blocks H_k. After passing to a subsequence, we may suppose that each ratio $\lambda_i(H_k)/\varepsilon_k$ has a limit in $[0, +\infty]$. For an eigenvalue such that $\lambda_i(H_k) \to 0$, the corresponding term has one of the following limits:

$$\frac{\lambda_i(H_k)}{\varepsilon_k + \lambda_i(H_k)} \to \begin{cases} 0, & \lambda_i(H_k) = o(\varepsilon_k), & \text{fast vanishing,} \\ c/(1+c), & \lambda_i(H_k)/\varepsilon_k \to c \in]0, +\infty[, & \text{critically vanishing,} \\ 1, & \varepsilon_k = o(\lambda_i(H_k)), & \text{slow vanishing.} \end{cases}$$

Eigenvalues with a positive limit also have multiplier one, as prescribed by $H^\dagger H$, while identically zero eigenvalues belong to the fast regime and therefore cause no difficulty. Thus agreement with $H^\dagger H$ requires every eigenvalue converging to zero to vanish faster than ε_k.

To summarize, if $p > 1$, $0 < q < 1$, $c > 0$, and $H_\varepsilon = \operatorname{diag}(0, \varepsilon^p, c\varepsilon, \varepsilon^q, 1)$, then $(H_\varepsilon + \varepsilon I_n)^{-1}H_\varepsilon \to \operatorname{diag}(0, 0, c/(1+c), 1, 1)$ but $H^\dagger H = \operatorname{diag}(0, 0, 0, 0, 1)$. The critical and slow regimes can retain a nonzero contribution in directions that belong to $\ker H$ at the limit. This is the contribution represented by the residual S below.

Remark 9 (Reachable set)**.** Fix $\theta \in O$. Since $x^\star_\varepsilon(\theta) \to x^\star_{\min}(\theta)$ and J_F is locally bounded, every sequence of blocks along the regularized trajectory admits a convergent subsequence. The closed graph property of J_F then guarantees its limit in $J_F(x^\star_{\min}(\theta), \theta)$. We call the set of all such limits the *reachable* set denoted by

$$\mathcal{R}(\theta) := \left\{ (H\ M) \in J_F(x^\star_{\min}(\theta), \theta) : \begin{array}{l} \exists\, \varepsilon_k \downarrow 0,\ (H_k\ M_k) \in J_F(x^\star_{\varepsilon_k}(\theta), \theta), \\ (H_k\ M_k) \to (H\ M) \end{array} \right\}.$$

A corresponding cluster point of the regularized derivatives belongs to $J_{x^\star_{\min}}(\theta)$ by taking $\theta_k \equiv \theta$ in (2). However, (2) also permits $\theta_k \to \theta$ so that the full set $J_{x^\star_{\min}}(\theta)$ may contain additional elements.

Example 8 shows that vanishing curvature from the limiting block can affect the $-H^\dagger M$ formula. The next theorem makes this precise by introducing a residual S; its proof is given in Section D.1.

Theorem 10 (Residual formula)**.** *Assume that* **(Reg)**, **(D)** *and* **(R)** *hold, and fix $\theta \in O$. Let $\varepsilon_k \downarrow 0$ with $0 < \varepsilon_k \leq \bar{\varepsilon}$, and select $(H_k\ M_k) \in J_F(x^\star_{\varepsilon_k}(\theta), \theta)$. Set $R_k := (H_k + \varepsilon_k I_n)^{-1} H_k$, $W_k := H_k^\dagger M_k$, and $V_k := -(H_k + \varepsilon_k I_n)^{-1} M_k = -R_k W_k$. Then $V_k \in J_{x^\star}(\theta, \varepsilon_k)$, and the sequences $(W_k)_k$ and $(V_k)_k$ are bounded. If $V_k \to V$ along a subsequence, then, after passing to a further common subsequence, $(H_k\ M_k) \to (H\ M) \in \mathcal{R}(\theta)$, $W_k \to W$, and $R_k \to R$. Moreover, $M = HW$ and $S := R - H^\dagger H = \Pi_{\ker H} R\, \Pi_{\ker H}$ with $0 \preceq S \preceq \Pi_{\ker H}$, while the limit satisfies $V = -H^\dagger M - SW \in J_{x^\star_{\min}}(\theta)$ and $HV + M = 0$.*

The matrix $H^\dagger H$ is the identity when restricted to range(H), whereas S acts only on $\ker(H)$. Thus S accounts for the regularized curvature that survives in directions where the limiting curvature is zero. The correction $-SW$ does not affect the limiting equation because $HS = 0$. We give conditions in the following subsection to ensure $S = 0$ although, for a given limit, the pseudoinverse formula already holds once the weaker condition $SW = 0$ is satisfied. Let us mention this result (and all following analogue ones) concerns limits corresponding to reachable blocks and not any element at the unregularized solution.

Remark 11. [Residual in the quartic example] In Example 8 with fixed $\theta \neq 0$, we have $H = M = 0$, $W = -1$, and $R = S = 1$, so the residual formula gives $V = 1$. At $\theta = 0$, the regularized solution is identically zero, hence the fixed trajectory gives $R = S = V = 0$ and therefore $0 \in J_{x^\star_{\min}}(0)$. On the other hand, $1 \in J_{x^\star_{\min}}(\theta)$ for every θ. Since $J_{x^\star_{\min}}$ has closed graph, letting $\theta \to 0$ shows that $1 \in J_{x^\star_{\min}}(0)$ as well. The theorem therefore identifies some elements of $J_{x^\star_{\min}}$ without necessarily characterizing the whole set. In practice, one admissible element is sufficient for a gradient-type method applied to (Q), see Section 3.3.2.

3.3.1 Conditions Yielding the Formula

We now give two pointwise conditions that force $S = 0$, both stated with respect to a fixed $\theta \in O$. The first rules out small positive eigenvalues, while the second allows them provided they vanish faster than ε.

- **(G)** [Spectral gap] there exist $\lambda^* > 0$ and $\varepsilon_G \in]0, \bar{\varepsilon}]$ such that, for every $0 < \varepsilon \leq \varepsilon_G$ and every $(H_\varepsilon\ M_\varepsilon) \in J_F(x^\star_\varepsilon(\theta), \theta)$, $\operatorname{spec}(H_\varepsilon) \subset \{0\} \cup [\lambda^*, +\infty[$.
- **(V)** [Fast vanishing] for every sequence $\varepsilon_k \downarrow 0$ with $0 < \varepsilon_k \leq \bar{\varepsilon}$ and every selection $(H_k\ M_k) \in J_F(x^\star_{\varepsilon_k}(\theta), \theta)$ such that $(H_k)_k$ converges, the ordered eigenvalues $\lambda_1(H_k) \geq \cdots \geq \lambda_n(H_k)$ satisfy $\lambda_i(H_k) \to 0 \implies \lambda_i(H_k) = o(\varepsilon_k)$.

Theorem 12 (Reachable pseudoinverse formula)**.** *Assume* **(Reg)**, **(D)** *and* **(R)**, *fix $\theta \in O$, and suppose that either* **(G)** *or* **(V)** *holds at θ. Then the conservative Jacobian $J_{x^\star_{\min}}$ defined in* (2) *satisfies $\{-H^\dagger M \colon (H\ M) \in \mathcal{R}(\theta)\} \subset J_{x^\star_{\min}}(\theta)$.*

The proof is given in Section D.2. Condition **(G)** includes the uniformly strongly convex setting but also permits eigenvalues that are identically zero. It may also be stated as an *eventual rank stability*, in the sense that $\operatorname{rank}(H_\varepsilon) = \operatorname{rank}(H)$ for ε small enough. Without either spectral condition, Assumptions **(Reg)**, **(D)** and **(R)** still give bounded derivatives and path-differentiability, but the residual S may be nonvanishing. This significantly generalizes Vicol et al. (2022, Statement 3.1), which considered only the fully quadratic setting.

The relationship between these conditions and standard nonsingularity assumptions is direct. Earlier implicit-differentiation results assume nonsingular H-blocks at the solution (Bolte et al., 2021) (thus near the solution by continuity of the determinant), or uniform nonsingularity (Blondel et al., 2021), while related range and spectral conditions appear

in Pauwels and Vaiter (2023). If the H_ε-blocks are uniformly positive definite on compact sets and the corresponding M_ε-blocks are uniformly bounded, then $W_\varepsilon = H_\varepsilon^{-1} M_\varepsilon$ verifies Assumption **(R)**, and the same lower bound gives **(G)**. Unlike uniform nonsingularity, however, Assumption **(R)** and **(G)** allow rank-deficient blocks.

As a consistency check, Theorem 12 recovers the Least-Squares calculation above.

Corollary 13 (Least-Squares)**.** *For the Least-Squares problem, $A^\dagger \in J_{x^\star_{\min}}(\theta)$ for all $\theta \in \mathbb{R}^m$.*

Proof For the classical singleton Jacobian used in the motivating example, $O = \mathbb{R}^m$, $\mathcal{R}(\theta) = \{(A^\mathsf{T} A \ -A^\mathsf{T})\}$, and **(G)** holds because the blocks do not depend on ε. Hence Theorem 12 gives $-(A^\mathsf{T} A)^\dagger(-A^\mathsf{T}) = (A^\mathsf{T} A)^\dagger A^\mathsf{T} = A^\dagger$. ■

3.3.2 Upper problem

Before discussing the case of Huber regression, we discuss what can be achieved for the upper problem once elements of $J_{x^\star_{\min}}$ are obtained. Let $\mathcal{T}(\theta) := \mathcal{L}(x^\star_{\min}(\theta), \theta)$.

By Theorem 5, $\theta \mapsto (x^\star_{\min}(\theta), \theta)$ is locally Lipschitz and definable thus path-differentiable. Assume $\mathcal{L}$ is locally Lipschitz and definable with conservative Jacobian $J_\mathcal{L}$. Then $\mathcal{T}$ is also locally Lipschitz and definable thus path-differentiable using the composition rules of conservative calculus (Bolte and Pauwels, 2021), with conservative Jacobian given by $J_\mathcal{T} : \theta \rightrightarrows \{(L_x \ L_\theta)(V; I_m) = L_x V + L_\theta : (L_x \ L_\theta) \in J_\mathcal{L}(x^\star_{\min}(\theta), \theta), V \in J_{x^\star_{\min}}(\theta)\}$. From there, the stochastic subgradient algorithm for conservative fields (Davis et al., 2020; Bolte and Pauwels, 2021) can be employed for the global problem.

3.4 Application to Huber Regression

Huber regression (Huber, 1964) illustrates how Assumption **(R)** can hold even when the Hessian blocks change rank along the regularized trajectory. Let $A \in \mathbb{R}^{m \times n}$, $\delta > 0$, and consider

$$\min_{x \in \mathbb{R}^n} \phi(Ax - \theta), \qquad \phi(y) := \sum_{i=1}^m h(y_i), \qquad h(t) := \tfrac{t^2}{2} \text{ if } |t| \le \delta \text{ otherwise } \delta|t| - \tfrac{1}{2}\delta^2. \tag{Hub}$$

Concretely, fix $\delta = 1$; the function $\phi\colon \mathbb{R}^m \to \mathbb{R}$ is convex and semialgebraic, with $\phi \in \mathcal{C}^{1,1}(\mathbb{R}^m) \setminus \mathcal{C}^2(\mathbb{R}^m)$. For $y \in \mathbb{R}^m$, every $Q = \operatorname{diag}(q) \in \partial(\nabla\phi)(y)$ is diagonal by separability, with $q_i = 1$ if $|y_i| < 1$, $q_i = 0$ if $|y_i| > 1$, and $q_i \in [0,1]$ if $|y_i| = 1$. Writing $F(x, \theta) = A^\mathsf{T} \nabla\phi(Ax - \theta)$, we use the definable conservative Jacobian $J_F(x, \theta) := \{(A^\mathsf{T} Q A \ -A^\mathsf{T} Q) : Q \in \partial(\nabla\phi)(Ax - \theta)\}$. Along the regularized solutions, we can write $Q_\varepsilon \in \partial(\nabla\phi)(Ax^\star_\varepsilon(\theta) - \theta)$ and $H_\varepsilon := A^\mathsf{T} Q_\varepsilon A$ and $M_\varepsilon := -A^\mathsf{T} Q_\varepsilon$. The matrices H_ε are symmetric PSD, although their rank may change with ε. We can nonetheless state a bound that is uniform over all the possible matrices Q_ε by defining

$$\chi(A) := \max\{\|(A_{R,C})^{-1}\| : |R| = |C| \ge 1, \ A_{R,C} \text{ nonsingular}\}, \tag{5}$$

where $R \subset \{1, \dots, m\}$ and $C \subset \{1, \dots, n\}$, taking $\chi(A) = 0$ when $A = 0$. This constant is finite because A has only finitely many square submatrices. This leads to two estimates, given in the next corollary; the first estimate controls the regularized systems uniformly in both δ and Q, while the second gives the corresponding bound at $\delta = 0$. The proof in Section D.3 first establishes a general diagonal-Gram estimate that also serves for the LASSO.

Corollary 14 (Uniform bound for weighted Gram systems)**.** *Let $A \in \mathbb{R}^{m \times n}$. For every diagonal matrix Q with diagonal in $\mathbb{R}^m_+$ and every $\delta > 0$, $\|(\delta I_n + A^\mathsf{T} Q A)^{-1} A^\mathsf{T} Q\| \le \chi(A)$ and $\|(\sqrt{Q}A)^\dagger \sqrt{Q}\| \le \chi(A)$.*

If one chooses $B_\varepsilon := \sqrt{Q_\varepsilon}A$ and $W_\varepsilon := -B_\varepsilon^\dagger\sqrt{Q_\varepsilon}$ then, since $B_\varepsilon^\mathsf{T} B_\varepsilon B_\varepsilon^\dagger = B_\varepsilon^\mathsf{T}$, $H_\varepsilon W_\varepsilon = -B_\varepsilon^\mathsf{T} B_\varepsilon B_\varepsilon^\dagger\sqrt{Q_\varepsilon} = -B_\varepsilon^\mathsf{T}\sqrt{Q_\varepsilon} = M_\varepsilon$. The second estimate in Corollary 14 gives then $\|W_\varepsilon\| = \|(\sqrt{Q_\varepsilon}A)^\dagger\sqrt{Q_\varepsilon}\| \le \chi(A)$. Thus both the range factorization and its bound are uniform over θ, ε, and every admissible Q_ε. This is summarized in the next corollary, whose proof is given in Section D.4.

Corollary 15 (Huber regression)**.** *For every fixed $A \in \mathbb{R}^{m\times n}$, Assumptions* **(Reg)**, **(D)** *and* **(R)** *hold for* (Hub) *with $\Theta = O = \mathbb{R}^m$ and any $\bar\varepsilon > 0$; the range factors satisfy the global bound $\|W_\varepsilon\| \le \chi(A)$. Consequently, $x^\star_{\min}\colon \mathbb{R}^m \to \mathbb{R}^n$ is path-differentiable, and $J_{x^\star_{\min}}$ defined in* (2) *is a definable conservative Jacobian.*

For $\theta \in \mathbb{R}^m$, define the reachable Huber matrices by

$$\mathcal{R}^{\mathrm{Hub}}(\theta) := \left\{ Q \in \partial(\nabla\phi)(Ax^\star_{\min}(\theta) - \theta) : \begin{array}{l} \exists\ \varepsilon_k \downarrow 0,\ Q_{\varepsilon_k} \in \partial(\nabla\phi)(Ax^\star_{\varepsilon_k}(\theta) - \theta), \\ Q_{\varepsilon_k} \to Q \end{array} \right\}. \quad (6)$$

These are precisely the matrices that can be obtained by keeping θ fixed and following the regularized solutions.

Proposition 16 (Limiting derivatives for Huber regression)**.** *For every $\theta \in \mathbb{R}^m$ and every $Q \in \mathcal{R}^{\mathrm{Hub}}(\theta)$, $(A^\mathsf{T} QA)^\dagger A^\mathsf{T} Q = (\sqrt{Q}A)^\dagger\sqrt{Q} \in J_{x^\star_{\min}}(\theta)$.*

The proof is given in Section D.5. Proposition 16 thus gives the pseudoinverse element associated with every reachable Q. It does not identify every cluster point of the regularized derivatives because the entries of Q_ε at the transition points of the Huber loss may vary with ε, and the eigenvalues of $A^\mathsf{T} Q_\varepsilon A$ may enter the critical or slow regimes described in subsection 3.3.1.

4 Composite Problems

We move on to composite problems involving sums of parametric functions. Parallel to our presentation of Section 3, we first introduce the notation and regularized implicit-differentiation calculus in the general setting. We then specialize it to the LASSO, whose uniform derivative bound will motivate the reduced range conditions used in the general results. Finally, we give some comments on potential convergence rates.

Let $O \subset \Theta$ be open and definable, and consider two functions $f : \mathbb{R}^n \times O \to \mathbb{R}$ and $g : \mathbb{R}^n \times O \to \mathbb{R} \cup \{+\infty\}$. For every $\theta \in O$, consider

$$\min_{x\in\mathbb{R}^n} \{f(x,\theta) + g(x,\theta)\} \quad \text{(P-Comp)}$$

under the following two assumptions.

(D') [Composite definability] The functions f and g as well as the set Θ are definable. This implies $\nabla_x f$ and $\mathrm{prox}_{\gamma g}$ for $\gamma > 0$ are definable, see Remark 17.

(Reg') [Composite regularity] For every $\theta \in O$, $f(\cdot,\theta) \in \mathcal{C}^{1,1}(\mathbb{R}^n)$ and $f(\cdot,\theta) \in \Gamma_0(\mathbb{R}^n)$, $g(\cdot,\theta) \in \Gamma_0(\mathbb{R}^n)$, and the solution set $\mathcal{S}_c(\theta) := \mathrm{argmin}_{x\in\mathbb{R}^n}\{f(x,\theta) + g(x,\theta)\}$ is nonempty. The map $\nabla_x f : \mathbb{R}^n \times O \to \mathbb{R}^n$ is jointly locally Lipschitz and there exists some $L \ge 0$ such that, for all $x, x' \in \mathbb{R}^n$ and $\theta \in O$, $\|\nabla_x f(x,\theta) - \nabla_x f(x',\theta)\| \le L\|x - x'\|$ Fix $\bar\varepsilon > 0$ and a positive stepsize $\gamma < 2/(L + 2\bar\varepsilon)$. For this γ, the parametric proximal map $G^\gamma(z,\theta) := \mathrm{prox}_{\gamma g(\cdot,\theta)}(z)$ is jointly locally Lipschitz.

Remark 17 (Definability of $\nabla_x f$ and G^γ)**.** The graphs of $\nabla_x f$ and G^γ admit first-order descriptions in terms of f and g. By closure of definable sets under first-order quantification, or equivalently under Boolean operations and coordinate projection, both maps are definable; in the semialgebraic case, this is the Tarski–Seidenberg theorem (see Section C.1 and Coste (1999); van den Dries and Miller (1996)). Their local Lipschitz continuity in **(Reg')** therefore makes their joint Clarke Jacobians conservative Jacobians.

Under **(Reg')**, $\mathcal{S}_c(\theta)$ is a nonempty closed convex set. We denote its minimal-norm element by $x^\star_{\min}(\theta) := \operatorname{argmin}_{x\in\mathcal{S}_c(\theta)} \frac{1}{2}\|x\|_2^2$. For $0 < \varepsilon \le \bar{\varepsilon}$, define the Tikhonov-regularized solution

$$x^\star_\varepsilon(\theta) := \operatorname*{argmin}_{x\in\mathbb{R}^n}\{f(x,\theta) + g(x,\theta) + \tfrac{\varepsilon}{2}\|x\|_2^2\}. \tag{P-Comp$_\varepsilon$}$$

The regularized objective is ε-strongly convex, so this solution is single-valued. For the fixed γ in **(Reg')**, define the forward step, the forward–backward operator, and its fixed-point residual by $z_\varepsilon(x,\theta) := x - \gamma\big(\nabla_x f(x,\theta) + \varepsilon x\big)$, $\varphi^\gamma_\varepsilon(x,\theta) := G^\gamma\big(z_\varepsilon(x,\theta),\theta\big)$, and $\Phi^\gamma_\varepsilon(x,\theta) := x - \varphi^\gamma_\varepsilon(x,\theta)$, respectively. Fermat's rule and the standard proximal characterization (see Bauschke and Combettes, 2017) ensures x solves (P-Comp$_\varepsilon$) iff $x = \varphi^\gamma_\varepsilon(x,\theta)$ iff $\Phi^\gamma_\varepsilon(x,\theta) = 0$.

We use the residual equation $\Phi^\gamma_\varepsilon(x,\theta) = 0$ to characterize the solutions whose dependence on θ we wish to differentiate. Write the joint Clarke Jacobians blockwise as $(H^f\ M^f) \in J_{\nabla_x f}(x,\theta) := \partial(\nabla_x f)(x,\theta)$ and $(Q\ P) \in J_{G^\gamma}(z,\theta) := \partial G^\gamma(z,\theta)$ where $H^f, Q \in \mathbb{R}^{n\times n}$ and $M^f, P \in \mathbb{R}^{n\times m}$. To keep the x- and θ-blocks from each joint Clarke selection paired, we set

$$\mathcal{D}_\varepsilon(x,\theta) := \left\{ D = (Q,P,H^f,M^f) \colon \begin{array}{l} (Q\ P) \in J_{G^\gamma}(z_\varepsilon(x,\theta),\theta), \\ (H^f\ M^f) \in J_{\nabla_x f}(x,\theta) \end{array} \right\}.$$

For $D \in \mathcal{D}_\varepsilon(x,\theta)$, we then define

$$H_\varepsilon(D) := I_n - Q\big((1-\gamma\varepsilon)I_n - \gamma H^f\big) \quad \text{and} \quad M_\varepsilon(D) := \gamma Q M^f - P. \tag{7}$$

The chain rule applied to these conservative mappings gives a conservative Jacobian of the residual

$$J_{\Phi^\gamma_\varepsilon}(x,\theta) := \{(H_\varepsilon(D)\ M_\varepsilon(D)) : D \in \mathcal{D}_\varepsilon(x,\theta)\}. \tag{8}$$

Proposition 18 (Conservative Jacobian of the regularized composite problem)**.** *Assume* **(D')** *and* **(Reg')**. *For every $0 < \varepsilon \le \bar{\varepsilon}$, the map $\theta \mapsto x^\star_\varepsilon(\theta)$ is path-differentiable on O, with the definable conservative Jacobian*

$$J_{x^\star_\varepsilon}(\theta) := \{-H_\varepsilon(D)^{-1} M_\varepsilon(D) : D \in \mathcal{D}_\varepsilon(x^\star_\varepsilon(\theta),\theta)\}. \tag{9}$$

The proof is given in Section E.1 and is a straightforward application of Bolte et al. (2024, Theorem 3.5 and Remark 3.6).

4.1 LASSO

Fix a matrix $A \in \mathbb{R}^{m\times n}$ and a penalty parameter $\lambda \in \mathbb{R}$, and regard the observation vector $\theta \in \mathbb{R}^m$ as the parameter. For every θ, consider the LASSO problem (Tibshirani, 1996, 2011; Osborne et al., 2000; Tibshirani, 2013) $\min_{x\in\mathbb{R}^n}\{\frac{1}{2}\|Ax-\theta\|_2^2 + e^\lambda\|x\|_1\}$. This is an instance of (P-Comp) with $O = \mathbb{R}^m$, $f(x,\theta) = \frac{1}{2}\|Ax-\theta\|_2^2$, and $g(x,\theta) = e^\lambda\|x\|_1$. Both functions are semialgebraic, hence definable, and their x-sections belong to $\Gamma_0(\mathbb{R}^n)$. Since $e^\lambda > 0$, the LASSO objective is coercive and therefore attains its minimum. Thus **(D')** holds.

For $0 < \varepsilon \le \bar{\varepsilon}$, its Tikhonov regularization, also called elastic net (Zou and Hastie, 2005; Mairal et al., 2012), admits a unique solution $x^\star_\varepsilon(\theta) := \operatorname{argmin}_{x\in\mathbb{R}^n}\{\frac{1}{2}\|Ax-\theta\|_2^2 + e^\lambda\|x\|_1 +$

$\frac{\varepsilon}{2}\|x\|_2^2\}$. The gradient $\nabla_x f(x,\theta) = A^\mathsf{T}(Ax-\theta)$ is jointly affine and is $\|A\|^2$-Lipschitz in x, while $G^\gamma = \mathrm{prox}_{\gamma e^\lambda\|\cdot\|_1}$ is the soft-thresholding operator and is jointly globally Lipschitz. Thus **(Reg')** also holds, and the fixed stepsize chosen above must satisfy $0 < \gamma < 2/(\|A\|^2 + 2\bar{\varepsilon})$. Substituting the blocks $H^f = A^\mathsf{T}A$, $M^f = -A^\mathsf{T}$, and $P = 0$ into Proposition 18 gives the following formula.

Corollary 19 (Conservative Jacobian of the regularized LASSO)**.** *For every $0 < \varepsilon \le \bar{\varepsilon}$, the map $\theta \mapsto x_\varepsilon^\star(\theta)$ is path-differentiable, with conservative Jacobian*

$$J_{x_\varepsilon^\star}(\theta) := \left\{ \Big(I_n - Q\big(I_n - \gamma(A^\mathsf{T}A + \varepsilon I_n)\big)\Big)^{-1} \gamma Q A^\mathsf{T} : Q \in \partial \mathrm{prox}_{\gamma e^\lambda\|\cdot\|_1}\big(z_\varepsilon(x_\varepsilon^\star(\theta),\theta)\big) \right\}. \tag{10}$$

Matrices Q in the Clarke Jacobian of the soft-thresholding operator $\mathrm{prox}_{\gamma e^\lambda\|\cdot\|_1}$ at a point z are diagonal with entries $(Q)_{ii} = 1$ if $|z_i| > \gamma e^\lambda$, $(Q)_{ii} = 0$ if $|z_i| < \gamma e^\lambda$, and the set $[0,1]$ if $|z_i| = \gamma e^\lambda$. This formula follows from the separability of soft-thresholding (Bauschke and Combettes, 2017). Along the regularized trajectory, an admissible selection $Q_\varepsilon \in \partial \mathrm{prox}_{\gamma e^\lambda\|\cdot\|_1}\big(z_\varepsilon(x_\varepsilon^\star(\theta),\theta)\big)$ may vary with ε. In particular, neither its support nor the positive diagonal entries in Q_ε need remain fixed as $\varepsilon \downarrow 0$. This is not an issue if the estimate is uniform over all admissible selections. The following proposition proves a stronger bound over the entire set $\mathrm{diag}([0,1]^n)$, which always contains the admissible matrices Q_ε.

Proposition 20 (LASSO uniform derivative bound)**.** *Let $Q \in \mathrm{diag}([0,1]^n)$ be nonzero and let $S_+ := \{i : (Q)_{ii} > 0\}$ be its support. For every $\varepsilon \ge 0$,*

$$\big\| \big(\tfrac{1}{\gamma}(Q^{-1}_{S_+,S_+} - I_{|S_+|}) + \varepsilon I_{|S_+|} + A^\mathsf{T}_{:,S_+}A_{:,S_+}\big)^\dagger A^\mathsf{T}_{:,S_+} \big\| \le \chi(A_{:,S_+}) \le \chi(A) \tag{11}$$

with $\chi(A)$ defined in (5) *and the pseudoinverse being an inverse when $\varepsilon > 0$. For $\varepsilon > 0$, the matrix $\big(I_n - Q(I_n - \gamma(A^\top A + \varepsilon I_n))\big)^{-1}\gamma Q A^\top$ vanishes on the rows outside S_+ and coincides on the rows indexed by S_+ with the matrix in* (11)*. In particular, condition (iv) of* Theorem 3 *holds for the LASSO with the bound $C_K = \chi(A)$, for every compact set K.*

The proof is given in Section E.2. Although the x-block being inverted is not symmetric, the computation reduces to a symmetric system on S_+; Section 4.2 performs this reduction in general.

4.2 Uniform Control in the General Case

We now return to the general composite problem. The LASSO calculation shows that the x-blocks in (8) need not be symmetric, even though they arise from convex data. We therefore expose the positive semidefinite system carried by $\mathrm{range}(Q)$ before formulating the uniform assumptions. The blocks inherit structure from the assumed convexity of the problem. By Theorem 35, Q is symmetric with $0 \preceq Q \preceq I_n$. At every point where $\nabla_x f$ is differentiable, its x-block is a symmetric Hessian satisfying $0 \preceq \nabla^2_{xx} f \preceq L I_n$ by convexity and the L-Lipschitz continuity of $\nabla_x f$. These properties are preserved under limits and convex combinations, so every H^f is symmetric with $0 \preceq H^f \preceq L I_n$. In particular, $\Pi_Q := QQ^\dagger = Q^\dagger Q$ is the orthogonal projection onto $\mathrm{range}(Q)$, while $I_n - \Pi_Q$ is the orthogonal projection onto $\ker(Q)$. For $D = (Q, P, H^f, M^f)$, denote $\Sigma(D) := Q^\dagger - \Pi_Q + \gamma\Pi_Q H^f \Pi_Q$, $T_\varepsilon(D) := \Sigma(D) + \gamma\varepsilon\Pi_Q$, and $V(D) := Q^\dagger M_\varepsilon(D) - \gamma\Pi_Q H^f (I_n - \Pi_Q) M_\varepsilon(D)$. These matrices control the inversion of the x-blocks through the following lemma, proven in Section E.3.

Lemma 21 (Projected inversion)**.** *Let $\gamma, \varepsilon > 0$, let $Q, H^f \in \mathbb{R}^{n\times n}$ be symmetric with $0 \preceq Q \preceq I_n$ and $H^f \succeq 0$, and let $M \in \mathbb{R}^{n\times m}$. Set $H := I_n - Q\big((1-\gamma\varepsilon)I_n - \gamma H^f\big)$ and $\Pi := QQ^\dagger$. Then Σ is symmetric positive semidefinite and $\Pi\Sigma = \Sigma\Pi = \Sigma$. Moreover, T_ε vanishes on $\ker(Q)$ and is positive definite on $\mathrm{range}(Q)$, with $T_\varepsilon|_{\mathrm{range}(Q)} \succeq \gamma\varepsilon I_{\mathrm{range}(Q)}$. The matrix H is invertible and satisfies both $(I_n - \Pi)H^{-1}M = (I_n - \Pi)M$ and $\Pi H^{-1}M = T_\varepsilon^\dagger V$. Consequently, $H^{-1}M = (I_n - \Pi)M + T_\varepsilon^\dagger V$. If $\mathrm{range}(V) \subseteq \mathrm{range}(\Sigma)$, then*

$$\|T_\varepsilon^\dagger V\| \le \|\Sigma^\dagger V\| \quad \text{and} \quad \|H^{-1}M\| \le \|(I_n - \Pi)M\| + \|\Sigma^\dagger V\|. \tag{12}$$

In particular, if $V = \Sigma W$, then $\|\Sigma^\dagger V\| \le \|W\|$, so the two right-hand bounds in (12) may be replaced by $\|W\|$ and $\|(I_n - \Pi)M\| + \|W\|$, respectively.

We now apply Lemma 21 to the regularized solution map. For $D \in \mathcal{D}_\varepsilon(x_\varepsilon^\star(\theta), \theta)$, Proposition 18 identifies $-H_\varepsilon(D)^{-1}M_\varepsilon(D)$ as an element of the conservative Jacobian (9). Taking $H = H_\varepsilon(D)$ and $M = M_\varepsilon(D)$ in Lemma 21 gives $(I_n - \Pi_Q)H_\varepsilon(D)^{-1}M_\varepsilon(D) = (I_n - \Pi_Q)M_\varepsilon(D)$ and $\Pi_Q H_\varepsilon(D)^{-1}M_\varepsilon(D) = T_\varepsilon(D)^\dagger V(D)$. These identities determine the $\ker(Q)$ component without an inversion and reduce the $\mathrm{range}(Q)$ component to the symmetric positive definite system with $T_\varepsilon(D)$. Consequently,

$$-H_\varepsilon(D)^{-1}M_\varepsilon(D) = -(I_n - \Pi_Q)M_\varepsilon(D) - T_\varepsilon(D)^\dagger V(D). \tag{13}$$

For the limiting result, the matrices in (13) must remain uniformly bounded as $\varepsilon \downarrow 0$. The Tikhonov regularization does not damp the $\ker(Q)$ component (the first term), which must therefore be bounded directly. On $\mathrm{range}(Q)$, (12) gives the required control when $V(D)$ belongs to $\mathrm{range}(\Sigma(D))$ and $\Sigma(D)^\dagger V(D)$ remains bounded. These requirements motivate the following assumptions.

The decomposition is pointwise in Q. In particular, it does not assert that Π_Q varies continuously when $\mathrm{rank}(Q)$ changes; we return to this issue when studying cluster points.

(R') [Composite uniform range assumption] With O and $\bar\varepsilon$ as above, suppose that for every compact set $K \subset O$ there exists $C_K > 0$ such that, for every $\theta \in K$, every $\varepsilon \in]0, \bar\varepsilon]$, and every $D \in \mathcal{D}_\varepsilon(x_\varepsilon^\star(\theta), \theta)$, $\mathrm{range}(V(D)) \subseteq \mathrm{range}(\Sigma(D))$ and $\|\Sigma(D)^\dagger V(D)\| \le C_K$.

The range inclusion gives $\Sigma(D)\Sigma(D)^\dagger V(D) = V(D)$. Conversely, if $V(D) = \Sigma(D)W$, then $\Sigma(D)^\dagger V(D) = \Sigma(D)^\dagger\Sigma(D)W$ is obtained by projecting each column of W orthogonally onto $\mathrm{range}(\Sigma(D))$, and hence $\|\Sigma(D)^\dagger V(D)\| \le \|W\|$. Thus **(R')** is equivalent to the existence of uniformly bounded solutions of $\Sigma(D)W = V(D)$, and is the exact analog of **(R)** for the reduced system on $\mathrm{range}(Q)$.

(M') [Composite uniform kernel assumption] With O and $\bar\varepsilon$ as above, suppose that for every compact set $K \subset O$ there exists $B_K \ge 0$ such that, for every $\theta \in K$, every $\varepsilon \in]0, \bar\varepsilon]$, and every $D = (Q, P, H^f, M^f) \in \mathcal{D}_\varepsilon(x_\varepsilon^\star(\theta), \theta)$, $\|(I_n - \Pi_Q)M_\varepsilon(D)\| = \|(I_n - \Pi_Q)P\| \le B_K$.

Since $(I_n - \Pi_Q)Q = 0$ and $M_\varepsilon(D) = \gamma QM^f - P$, it holds $(I_n - \Pi_Q)M_\varepsilon(D) = -(I_n - \Pi_Q)P$ and $V(D) = \gamma\Pi_Q M^f - Q^\dagger P + \gamma\Pi_Q H^f(I_n - \Pi_Q)P$. Thus **(M')** controls only the θ-block P of J_{G^γ}.

Remark 22 (Sufficient conditions for **(M')**)**.** Assumption **(M')** holds in either of the following situations. (a) If g does not depend on θ, then $P = 0$, so one may take $B_K = 0$; moreover, $V(D) = \gamma\Pi_Q M^f$. (b) Suppose that $x_{\min}^\star$ is independently known to be locally bounded on O. The Tikhonov estimate $\|x_\varepsilon^\star(\theta)\| \le \|x_{\min}^\star(\theta)\|$ (Theorem 36) and **(Reg')** then

imply that, for $\theta \in K$ and $0 < \varepsilon \le \bar{\varepsilon}$, the points $(z_\varepsilon(x^\star_\varepsilon(\theta), \theta), \theta)$ remain in a compact subset of $\mathbb{R}^n \times O$. The Clarke Jacobian of the locally Lipschitz map G^γ is bounded on this set, and hence so is $\|(I_n - \Pi_Q)P\| \le \|P\|$.

With those assumptions in mind, we can state the path-differentiability theorem. Its proof is given in Section E.4.

Theorem 23 (Path-differentiability of the composite case)**.** *Assume that* **(D')**, **(Reg')**, **(R')** *and* **(M')** *hold. Then $J_{x^\star_{\min}}$ defined in* (2) *is a definable conservative Jacobian of the minimal-norm solution $x^\star_{\min}$ of* (P-Comp)*. In particular, $x^\star_{\min}$ is path-differentiable on O.*

The uniform bound underlying the theorem ensures that derivative sequences have cluster points but does not identify their form. Two losses of continuity must be distinguished. The projector $Q \mapsto QQ^\dagger$ may jump when the rank of Q changes, and, even after this projector stabilizes, eigenvalues of the reduced matrix Σ may vanish relative to ε.

Under the assumptions of Theorem 23, fix $\theta \in O$, let $\varepsilon_k \downarrow 0$ with $0 < \varepsilon_k \le \bar{\varepsilon}$, and select $D_k = (Q_k, P_k, H^f_k, M^f_k) \in \mathcal{D}_{\varepsilon_k}(x^\star_{\varepsilon_k}(\theta), \theta)$ so $D_k \to D_0 = (Q_0, P_0, H^f_0, M^f_0)$. For the results below, we write $\Pi_k := \Pi_{Q_k}$, $\mathcal{M}_k := \gamma Q_k M^f_k - P_k$, $\Sigma_k := \Sigma(D_k)$, $V_k := V(D_k)$, $T_k := T_{\varepsilon_k}(D_k)$, $J_k := -H_{\varepsilon_k}(D_k)^{-1}\mathcal{M}_k = -(I_n - \Pi_k)\mathcal{M}_k - T^\dagger_k V_k$, and define Π_0, $\mathcal{M}_0$, Σ_0, and V_0 analogously from D_0.

Proposition 24 (Rank-changing cluster points)**.** *The limit satisfies $D_0 \in \mathcal{D}_0(x^\star_{\min}(\theta), \theta)$, the sequence $(J_k)_k$ is bounded, and every cluster point belongs to $J_{x^\star_{\min}}(\theta)$. If J is such a cluster point, then, after passing to a subsequence converging to J and extracting further, there exist an orthogonal projector $\bar{\Pi}$ and a matrix $\bar{Y}$ such that $\Pi_k \to \bar{\Pi}$, $T^\dagger_k V_k \to \bar{Y}$, and $J = -(I_n - \bar{\Pi})\mathcal{M}_0 - \bar{Y}$. Moreover, $\bar{\Pi}Q_0 = Q_0\bar{\Pi} = Q_0$,* $\mathrm{range}(Q_0) \subseteq \mathrm{range}(\bar{\Pi})$*, and $\bar{\Pi}\bar{Y} = \bar{Y}$.*

The proof is given in Section E.5.

Corollary 25 (Stable-Q residual formula)**.** *Suppose that* $\mathrm{rank}(Q_k) = \mathrm{rank}(Q_0)$ *for all sufficiently large k, and let J be a cluster point of $(J_k)_k$. Set $W_k := \Sigma^\dagger_k V_k$ and $R_k := T^\dagger_k \Sigma_k$. Then $Q^\dagger_k \to Q^\dagger_0$, $\Pi_k \to \Pi_0$, $\Sigma_k \to \Sigma_0$, and $V_k \to V_0$. Along every subsequence converging to J, a further subsequence satisfies $W_k \to W$ and $R_k \to R$. Set $S := R - \Sigma^\dagger_0\Sigma_0$. Then $\Sigma_0 W = V_0$, $0 \preceq S \preceq \Pi_0 - \Sigma^\dagger_0\Sigma_0$, $\Sigma_0 S = S\Sigma_0 = 0$, and $J = -(I_n - \Pi_0)\mathcal{M}_0 - \Sigma^\dagger_0 V_0 - SW$.*

The proof is given in Section E.6.

For a sequence satisfying the stable-rank condition in Corollary 25, consider either of the following spectral alternatives:

$$\mathrm{spec}(\Sigma_k) \subset \{0\} \cup [\lambda^*, +\infty[\quad \text{eventually, for some } \lambda^* > 0, \tag{G_Σ}$$

or, writing the eigenvalues in nonincreasing order,

$$\lambda_i(\Sigma_k) \to 0 \quad \Longrightarrow \quad \lambda_i(\Sigma_k) = o(\varepsilon_k), \qquad i = 1, \dots, n. \tag{V_Σ}$$

Corollary 26 (Pseudoinverse cluster formula)**.** *Under the stable-rank hypothesis of* Corollary 25*, if either* (G_Σ) *or* (V_Σ) *holds, then the representation in* Corollary 25 *may be taken with $S = 0$. Equivalently, every cluster point J of $(J_k)_k$ satisfies $J = -(I_n - \Pi_0)\mathcal{M}_0 - \Sigma^\dagger_0 V_0$.*

The proof is given in Section E.7.

Stable reduced rank. If, in addition to stable rank of Q_k, $\mathrm{rank}(\Sigma_k) = \mathrm{rank}(\Sigma_0)$ for all sufficiently large k, then (G_Σ) holds. Indeed, $\Sigma_k \to \Sigma_0$, so their positive eigenvalues stay uniformly away from zero.

4.3 Return to the LASSO

Corollary 27 (LASSO path-differentiability)**.** *The minimal-norm solution $x^\star_{\min}$ of the LASSO problem is path-differentiable as a function of θ.*

Proof For the LASSO, $P = 0$, $H^f = A^\mathsf{T}A$ and $M^f = -A^\mathsf{T}$, so **(M')** holds with $B_K = 0$ and $V(D) = -\gamma\Pi_Q A^\mathsf{T}$. If $Q = 0$ then $\Sigma(D) = V(D) = 0$. Otherwise, with S_+ as in Proposition 20, Π_Q is the projection onto S_+, $\Sigma(D)_{S_+,S_+} = Q^{-1}_{S_+,S_+} - I_{|S_+|} + \gamma A^\mathsf{T}_{:,S_+}A_{:,S_+}$ and $V(D)_{S_+,:} = -\gamma A^\mathsf{T}_{:,S_+}$, all other rows and columns vanishing. Since $\operatorname{range}(A^\mathsf{T}_{:,S_+}) = \operatorname{range}(A^\mathsf{T}_{:,S_+}A_{:,S_+}) \subseteq \operatorname{range}\Sigma(D)_{S_+,S_+}$ (the kernel of a sum of positive semidefinite matrices is the intersection of their kernels), we have $\operatorname{range} V(D) \subseteq \operatorname{range}\Sigma(D)$, and $\Sigma(D)^\dagger V(D)$ vanishes outside the rows indexed by S_+, where it equals minus the matrix in (11) with $\varepsilon = 0$. Its norm is therefore at most $\chi(A)$ by Proposition 20, so **(R')** holds with $C_K = \chi(A)$, and Theorem 23 applies. ■

We now identify explicit elements of $J_{x^\star_{\min}}(\theta)$ obtained from binary soft-threshold selections along the regularized trajectory. Fix $\theta \in \mathbb{R}^m$, set $\tau := \gamma e^\lambda$, and define $\mathcal{Q}_\varepsilon(\theta) := \partial\operatorname{prox}_{\gamma e^\lambda\|\cdot\|_1}\big(z_\varepsilon(x^\star_\varepsilon(\theta),\theta)\big)$ for $0 < \varepsilon \le \bar\varepsilon$ and $\mathcal{Q}_0(\theta) := \partial\operatorname{prox}_{\gamma e^\lambda\|\cdot\|_1}\big(z_0(x^\star_{\min}(\theta),\theta)\big)$. The fixed-parameter reachable set is

$$\mathcal{R}^{\mathrm{LASSO}}(\theta) := \{Q \in \mathcal{Q}_0(\theta) : \exists\, \varepsilon_k \downarrow 0\colon Q_k \in \mathcal{Q}_{\varepsilon_k}(\theta), Q_k \to Q\}\,. \tag{14}$$

Since (2) also permits $\theta_k \to \theta$, this fixed-parameter set need not describe every element of $J_{x^\star_{\min}}(\theta)$. For $i \in [1:n]$, let $\rho_i(\varepsilon;\theta) := \big|\big[z_\varepsilon(x^\star_\varepsilon(\theta),\theta)\big]_i\big| - \tau$. Each $\rho_i(\cdot;\theta)$ is definable, so, for all sufficiently small $\varepsilon > 0$, it is either positive, negative, or identically zero.

Corollary 28 (Reachable binary LASSO derivatives)**.** *Fix $\theta \in \mathbb{R}^m$ and set $\mathcal{R}^{\mathrm{LASSO}}_{\mathrm{bin}}(\theta) := \mathcal{R}^{\mathrm{LASSO}}(\theta) \cap \operatorname{diag}(\{0,1\}^n)$. A binary diagonal matrix Q belongs to $\mathcal{R}^{\mathrm{LASSO}}_{\mathrm{bin}}(\theta)$ if and only if, for every $i \in [1:n]$,*

$$(Q)_{ii} \in \begin{cases} \{1\}, & \rho_i(\varepsilon;\theta) > 0 \text{ for all sufficiently small } \varepsilon,\\ \{0\}, & \rho_i(\varepsilon;\theta) < 0 \text{ for all sufficiently small } \varepsilon,\\ \{0,1\}, & \rho_i(\varepsilon;\theta) = 0 \text{ for all sufficiently small } \varepsilon. \end{cases}$$

In particular, $\mathcal{R}^{\mathrm{LASSO}}_{\mathrm{bin}}(\theta) \neq \varnothing$, and $\big\{(AQ)^\dagger : Q \in \mathcal{R}^{\mathrm{LASSO}}_{\mathrm{bin}}(\theta)\big\} \subseteq J_{x^\star_{\min}}(\theta)$.

The proof is given in Section E.8.

Remark 29 (Differentiation with respect to θ and λ)**.** All LASSO derivatives above are taken with respect to the observation θ, with λ fixed. In this parameterization, the proximal map is independent of θ, hence $P = 0$, and Proposition 20 gives a uniform bound without any rank assumption on A. If λ is differentiated instead, it enters the soft-threshold level: $M^f = 0$, while the proximal parameter block P need not vanish. The argument of Proposition 20 therefore does not apply directly. Bolte et al. (2021, Proposition 5 and Section C.3) obtain path-differentiability of the LASSO solution map without Tikhonov regularization by assuming that $A^\mathsf{T}_{:,E}A_{:,E}$ is invertible, where E is the equicorrelation set. A very similar assumption arises in Bertrand et al. (2022). Their condition is sufficient for that λ-sensitivity result, whereas our θ-sensitivity result for the minimal-norm selection requires no such condition.

4.4 Convergence Rates with Fixed Tikhonov Regularization

This section discusses the accuracy algorithms can reach, in the context of Tikhonov regularization, following Grazzi et al. (2024); Bolte et al. (2022). We focus on problem (P-Comp) under the same assumptions and fixed-point formulation:

$$x = \operatorname{prox}_{\gamma g(\cdot,\theta)}(z_\varepsilon(x,\theta)) = \varphi_\varepsilon^\gamma(x,\theta), \qquad z_\varepsilon(x,\theta) = x - \gamma\varepsilon x - \gamma\nabla_x f(x,\theta). \tag{15}$$

The mapping $\varphi_\varepsilon^\gamma$ is forward-backward due to the proximal (implicit/backward) step on γg composed with the gradient-descent (explicit/forward) step on γf. A key property to study the sequence of Jacobians is that the derivatives of $\varphi_\varepsilon^\gamma$ w.r.t. x are contracting. Let $\overline{\varepsilon} > 0$ be such that $\varepsilon \in]0, \overline{\varepsilon}]$, for $\gamma \in]0, 2/(2\overline{\varepsilon} + L)[$, $\varphi_\varepsilon^\gamma$ has contracting x-derivatives. Indeed, $0 \preceq H^f \preceq LI_n$ and $(1-\gamma\varepsilon-\gamma L)I_n \preceq I_n - \gamma\varepsilon I_n - \gamma H^f \preceq (1-\gamma\varepsilon)I_n$ by (7), so the eigenvalues of $I_n - \gamma\varepsilon I_n - \gamma H^f$ belong to $]-1, +1[$. One concludes using that $\|Q_\varepsilon\| \leq 1$. Once a suitable γ is set, we can use fixed-point algorithms such as in Grazzi et al. (2024, Section 4) (see also Bolte et al., 2022, Section 4.1), using (15) and leading to the iterations

$$x_\varepsilon^0(\theta) \in \mathbb{R}^n, \qquad x_\varepsilon^{k+1}(\theta) = \varphi_\varepsilon^\gamma(x_\varepsilon^k(\theta), \theta). \tag{16}$$

Now, consider the derivatives. For fixed $\varepsilon > 0$, due to the contractivity of the x-derivatives, the fixed-point approach yields a rather natural conservative Jacobian (Bolte et al., 2022): $J^{\mathrm{fix}}_{x_\varepsilon^\star}(\theta) := \mathrm{fix}(J_{\varphi_\varepsilon^\gamma}(x_\varepsilon^\star(\theta), \theta))$ where fix denotes the (set of) fixed-points of a set of matrices, to be understood as:

$$\begin{aligned}\mathrm{fix}\left(J_{\varphi_\varepsilon^\gamma}(x_\varepsilon^\star(\theta),\theta)\right) = J_{\varphi_\varepsilon^\gamma}(x_\varepsilon^\star(\theta),\theta)\left[\mathrm{fix}\left(J_{\varphi_\varepsilon^\gamma}(x_\varepsilon^\star(\theta),\theta)\right)\right], \\ J_{\varphi_\varepsilon^\gamma}(x_\varepsilon^\star(\theta),\theta)(\mathcal{Z}) = \{AZ + B : Z \in \mathcal{Z}, (A\ B) \in J_{\varphi_\varepsilon^\gamma}(x_\varepsilon^\star(\theta),\theta)\},\end{aligned}$$

where A is the x-block and B the θ-block (thus $\|A\| < 1$). Notably, $J^{\mathrm{imp}}_{x_\varepsilon^\star}(\theta) \subset J^{\mathrm{fix}}_{x_\varepsilon^\star}(\theta)$ (Bolte et al., 2022, Remark 2) when both are computed with the same $J_{\varphi_\varepsilon^\gamma}$. Elements of $J^{\mathrm{fix}}_{x_\varepsilon^\star}(\theta)$ can be computed by the following iterations (Grazzi et al., 2024, Section 4), akin to (16):

$$\text{(ITER)} \quad J^0_{x_\varepsilon^\star}(\theta) = \{0\}, \quad J^{k+1}_{x_\varepsilon^\star}(\theta) = H^k J^k_{x_\varepsilon^\star}(\theta) + M^k, \quad (H^k\ M^k) \in \partial\varphi_\varepsilon^\gamma(x_\varepsilon^k(\theta), \theta).$$

An improved version fixes $N \in \mathbb{N}$, computes $x_\varepsilon^N(\theta)$ and, for $k \leq N$, uses:

$$\text{(FIXN)} \quad J^{N,0}_{x_\varepsilon^\star}(\theta) = \{0\}, \quad J^{N,k+1}_{x_\varepsilon^\star}(\theta) = H^N J^{N,k}_{x_\varepsilon^\star}(\theta) + M^N, \quad (H^N\ M^N) \in \partial\varphi_\varepsilon^\gamma(x_\varepsilon^N(\theta), \theta).$$

The version (FIXN) yields better results since $x_\varepsilon^N(\theta)$ is typically closer to $x_\varepsilon^\star(\theta)$ than $x_\varepsilon^k(\theta)$. Indeed, Theorem 30 gives some upper bound between the excess (denoted e, $e(\mathcal{C}, \mathcal{D}) = \sup_{C\in\mathcal{C}} \inf_{D\in\mathcal{D}} \|C - D\|$) and an element computed by the iterations (FIXN).

Theorem 30 (Rates of convergence, (Grazzi et al., 2024, Theorem 4.1 simplified))**.** *Let $\varepsilon > 0$ and $\gamma > 0$ small enough. One has, denoting k the iteration number and $\Delta_0 := \|x_\varepsilon^0(\theta) - x_\varepsilon^\star(\theta)\|$,*

$$\begin{aligned}\|x_\varepsilon^k(\theta) - x_\varepsilon^\star(\theta)\| &\leq C_1(1-\gamma\varepsilon)^k = O((1-\gamma\varepsilon)^k) \\ e(J^{N,k}_{x_\varepsilon^\star}(\theta), J^{\mathrm{fix}}_{x_\varepsilon^\star}(\theta)) &\leq \tfrac{B_\theta}{\gamma\varepsilon}(1-\gamma\varepsilon)^k + \tfrac{B_\theta+\gamma\varepsilon}{\gamma\varepsilon}\left(L' + \tfrac{M_\theta}{R_\theta}\right)\tfrac{1-(1-\gamma\varepsilon)^k}{\gamma\varepsilon}(1-\gamma\varepsilon)^N\Delta_0\end{aligned} \tag{17}$$

where $\varphi_\varepsilon^\gamma$ is L'-Lipschitz smooth, B_θ, M_θ and R_θ are other constants of $\varphi_\varepsilon^\gamma$.

The following remark exposes, in the Hölderian setting (see Definition 37 and Theorem 38), via Maulen-Soto et al. (2025), a simple but important fact of regularization techniques.

Remark 31 (Rate for x in the Hölderian setting)**.** Suppose that $f+g$ verifies a Hölderian error bound $f(x,\theta)+g(x,\theta)-f(x^\star_{\min}(\theta),\theta)-g(x^\star_{\min}(\theta),\theta)\geq\kappa\mathrm{dist}(x,\mathcal{S}(\theta))^p$ for some $p\geq 1$ and $\kappa>0$ (see Theorem 38). Then there exists C_0 and ε^* such that, for any $\varepsilon\leq\varepsilon^*$, one has

$$\|x^k_\varepsilon(\theta)-x^\star_{\min}(\theta)\|\leq \underbrace{C_1(1-\gamma\varepsilon)^k}_{\text{approximation error}}+\underbrace{C_0\varepsilon^{\frac{1}{2p}}}_{\text{regularization error}}.$$

Indeed, $\|x^k_\varepsilon(\theta)-x^\star_{\min}(\theta)\|\leq\|x^k_\varepsilon(\theta)-x^\star_\varepsilon(\theta)\|+\|x^\star_\varepsilon(\theta)-x^\star_{\min}(\theta)\|\leq C_1(1-\gamma\varepsilon)^k+C_0\varepsilon^{\frac{1}{2p}}$ using the triangle inequality, (17) and then Theorem 38.

When $\varepsilon\to 0$, so does $C_0\varepsilon^{\frac{1}{2p}}$ but $C_1(1-\gamma\varepsilon)^k$ increases: the fidelity to the original problem is counterbalanced by slower convergence speed; it is easier to solve a more regularized problem; see also Section SM4 of the Supplementary Material. For the conservative mappings, we could similarly use $e(J^k_{x^\star_\varepsilon}(\theta),J_{x^\star_{\min}}(\theta))\leq e(J^k_{x^\star_\varepsilon}(\theta),J^{\mathrm{fix}}_{x^\star_\varepsilon}(\theta))+e(J^{\mathrm{fix}}_{x^\star_\varepsilon}(\theta),J_{x^\star_{\min}}(\theta))$. The first term, by contractivity, can be estimated in terms of ε and k by Theorem 30. However, $J^{\mathrm{fix}}_{x^\star_\varepsilon}$ is not given by an IFT-type formula so different assumptions would be necessary to deal with the second term. Using $e(\mathcal{A},\mathcal{C})\leq e(\mathcal{A},\mathcal{B})+e(\mathcal{B},\mathcal{C})$ (see Grazzi et al., 2024, Lemma B.1(i)), we have

$$e(J^k_{x^\star_\varepsilon}(\theta),J_{x^\star_{\min}}(\theta))\leq \underbrace{e(J^k_{x^\star_\varepsilon}(\theta),J^{\mathrm{fix}}_{x^\star_\varepsilon}(\theta))}_{\text{approximation error}}+\underbrace{e(J^{\mathrm{fix}}_{x^\star_\varepsilon}(\theta),J^{\mathrm{imp}}_{x^\star_\varepsilon}(\theta))}_{\text{nonsmoothness error}}+\underbrace{e(J^{\mathrm{imp}}_{x^\star_\varepsilon}(\theta),J_{x^\star_{\min}}(\theta))}_{\text{regularization error}}. \quad (18)$$

Here, the second term represents some "derivative gap" between two specific conservative mappings: since in general $J^{\mathrm{imp}}\subsetneq J^{\mathrm{fix}}$, this term is nonzero. Finally, the third term can be $o(1)$ by Theorem 3, but without explicit rate. Those are available for simple problems such as Least-Squares (given in the Supplementary Material, Section SM4, Proposition SM5), and in some settings where the solution is unique (see Pauwels and Vaiter (2023) in optimal transport). The estimations of the nonsmoothness and regularization errors are interesting open questions.

Remark 32. In Bolte et al. (2022) and Grazzi et al. (2024), the explicit upper bounds are for the strongly convex setting. In Blondel et al. (2021, theorem 1), the strong bounds are obtained in the smooth case, with global uniform invertibility and Lipschitz bound.

5 Numerical Experiments

Full descriptions of the experimental details and data used are given in Section A.

5.1 Data Poisoning: Ill-posed Huber Regression

We consider a data poisoning experiment using the NYC Open Data service-request records (NYC Open Data, 2026) modeled by Huber regression. The detailed data preprocessing protocol is given in Section A.1; the duality and solution-set derivation is given in Section SM2 of the Supplementary Material.

Suppose the lower data have matrix $A \in \mathbb{R}^{n_{\text{tr}} \times p}$ and clean response $\theta_0 \in \mathbb{R}^{n_{\text{tr}}}$, while the validation data have design A_{val} and response b_{val}. The adversary replaces θ_0 by θ to degrade the performance of the model (maximizing upper-level loss). Denoting h_δ as the standard Huber loss (Hub) with threshold $\delta > 0$, $\tau > 0$ a perturbation budget, $x \in \mathbb{R}^p$ the request information, the bilevel problem is

$$\max_{\|\theta - \theta_0\|_2 \leq \tau} \tfrac{1}{2n_{\text{val}}} \|A_{\text{val}} x^\star(\theta) - b_{\text{val}}\|^2 \text{ subject to } x^\star(\theta) \in \operatorname*{argmin}_{x \in \mathbb{R}^p} \sum_{i=1}^{n_{\text{tr}}} h_\delta((Ax - \theta)_i).$$

The lower-level problem is ill-posed since, depending on the data split, the number of columns of A, p, is between 131 and 149 and the kernel has dimension $p - \operatorname{rank}(A)$ between 2 and 9. The minimal-norm solution for this problem is not known analytically. However, by rewriting the Huber loss using convex duality, one can compute the lower-level solution set and solve a quadratic problem to obtain the minimal-norm solution, see Section A.1. Using directions $q_k \in \ker(A)$, we can generate the following perturbed nonminimal solutions $x^\star_{c,k}(\theta) = x^\star_{\min}(\theta) + c\,\|x^\star_{\min}(\theta_0)\|_2\, q_k$ with $\|q_k\|_2 = 1, c \in \{0.5, 1, 2, 3\}$.

We compute the derivatives using a continuation-type method. To compute an element $V \in J_{x^\star_{\min}}(\theta)$, we follow Proposition 16. At each θ, we identify a reachable Huber matrix Q as defined in (6) by solving the Tikhonov problem over a fixed grid of ε values in $[10^{-10}, 10^{-1}]$. Each solution $x^\star_\varepsilon$ determines the diagonal matrix Q_ε with diagonal elements $(Q_\varepsilon)_{ii} = 1$ if $|(Ax^\star_\varepsilon(\theta) - \theta)_i| < \delta$, $(Q_\varepsilon)_{ii} \in [0, 1]$ if $|(Ax^\star_\varepsilon(\theta) - \theta)_i| = \delta$, and $(Q_\varepsilon)_{ii} = 0$ if $|(Ax^\star_\varepsilon(\theta) - \theta)_i| > \delta$. We check that Q_ε does not change over the final three values of ε and thus select $Q = Q_{\varepsilon = 10^{-10}}$. This gives numerical evidence that Q represents a reachable Huber matrix selected by the Tikhonov path. The validity of this process is illustrated in Figure 3.

For a fixed grid of values of ε from 10^{-1} to 10^{-10}, we compare the Jacobian error with a separately solved zero reference. The panes of Figure 3 display heatmaps (left panes) where the ordinate indicates ε and the abscissa indicates the upper iteration t: for each pair (ε, t) a colored rectangle gives the median error encountered among 20 seeds: the solution error on top and the Jacobian error on the bottom. We observe the error is mostly dependent on ε and decreases linearly with ε when it is small enough. The right panes represent a single iteration $t = 50$. We can then compute an element of the conservative Jacobian $V = (A^\mathsf{T} Q A)^\dagger A^\mathsf{T} Q$ and compute the implicit hypergradient depending on (c, q_k), which impacts the upper-level $V^\mathsf{T} \frac{1}{n_{\text{val}}} A_{\text{val}}^\mathsf{T}(A_{\text{val}} x^\star_{c,k} - b_{\text{val}})$.

For each pair (c, q_k) we do 150 iterations of projected normalized gradient ascent with Polyak momentum on the constrained bilevel problem and plot the median relative poison damage (upper objective value) in Figure 2. In this problem setting, using the minimal-norm is better on all 20 data splits at every c, though it need not hold in general: there is no reason the minimal-norm should be particularly efficient for the upper-level, as it is not an optimistic selection. Section A.1 gives the full details.

5.2 Composite Data Poisoning: Ill-posed LASSO

The principle of this experiment is akin to that of Section 5.1, with additional details given in Section A.2. Here, we use the scikit-learn Digits data set so that the lower-level problem fits Corollary 27. The upper-level is a Least-Squares maximization with parameters $(A_{\text{tr}}, Y_{\text{tr}})$ and variable $\Omega \in \Theta \subset \mathbb{R}^{10 \times 10}$ as synthetic response, with $\Theta = \{I_{10} + \Delta : \|\Delta\|_F \leq \tau\}$ for a budget $\tau > 0$. The lower-level is a (regularized) LASSO with design matrix $A \in \mathbb{R}^{10 \times 64}$ and

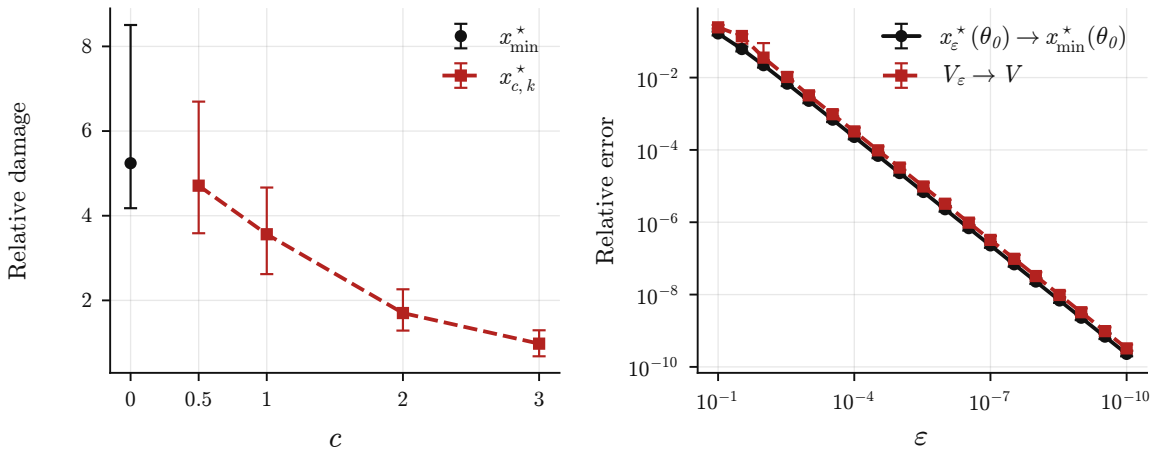


Figure 2: NYC 311 Huber poisoning. *Left:* each selection's relative poison damage $d_{c,k} = \frac{\Phi_{c,k}(\theta_{150}) - \Phi_{c,k}(\theta_0)}{\Phi_{c,k}(\theta_0)}$ with $\Phi_{c,k}$ the upper-level objective. *Right:* the relative solution error $\|x^\star_\varepsilon(\theta_0) - x^\star_{\min}(\theta_0)\|_2 / \|x^\star_{\min}(\theta_0)\|_2$ and relative derivative error $\|V_\varepsilon - V\|_{\mathrm{F}} / \|V\|_{\mathrm{F}}$ along the Tikhonov path at the clean response, where $V_\varepsilon = (A^\mathsf{T} Q_\varepsilon A + \varepsilon I_p)^{-1} A^\mathsf{T} Q_\varepsilon$ and $V = (A^\mathsf{T} Q A)^\dagger A^\mathsf{T} Q$. In both plots, markers are medians over the 20 declared splits and whiskers are 95.86% confidence intervals for the median. More details are in Section A.1.

variable $X \in \mathbb{R}^{64\times 10}$. Let $e^\lambda = 10^{-3}$, the bilevel problem reads

$$\max_{\|\Delta\|_F \le \tau} \frac{\|A_{\mathrm{tr}} X_\varepsilon(I_{10}+\Delta) - Y_{\mathrm{tr}}\|_F^2}{2n_{\mathrm{tr}}} \text{ subject to } X^\star_\varepsilon(\Omega) \in \operatorname*{argmin}_{X\in\mathbb{R}^{64\times 10}} \tfrac{1}{2\cdot 10}\|AX-\Omega\|_F^2 + e^\lambda \|X\|_1 + \tfrac{\varepsilon}{2}\|X\|_F^2. \tag{19}$$

The budget τ is defined as $\tau = \|I_{10}\|_F/10 = 1/\sqrt{10}$. At $\varepsilon = 0$, the positive ℓ_1 penalty makes the solution set nonempty and compact. To ensure existence of multiple solutions, A is designed with pairs of duplicate columns. The minimal-norm solution can be found by splitting equally within duplicate pairs.

The upper-level problem applies 100 iterations of projected normalized gradient ascent with Polyak momentum, using the hypergradient provided by the (regularized) lower-level solution. The lower-level problem is solved by cyclic coordinate descent, the scikit-learn solver for elastic net or ℓ_2-regularized LASSO (Zou and Hastie, 2005; Mairal et al., 2012).

To show the accuracy of implicit differentiation, we also use continuation, see Section 5.1. Convergence is observed after ε passes a certain threshold value, which we conjecture to be related to the active manifold identification of the regularized solution.

5.3 Showcasing Different Regimes: a LASSO Illustration

This section studies an overparameterized separable instance of LASSO whose goal is to illustrate Assumption (V_Σ) (see also Assumption **(V)**). Let $\theta \in \mathbb{R}^n_+$ with $\theta_1 \ge \cdots \ge \theta_n$. Let $\alpha > 0$, consider the following problem

$$\min_{x_1,x_2} \tilde{h}(x_1,x_2) := \tfrac{1}{2}\|x_1 + x_2 - \theta\|_2^2 + \alpha(\|x_1\|_1 + \|x_2\|_1). \tag{20}$$

It is an overparameterized version, obtained by duplicating the variable to obtain many solutions, of the following problem (see also Meyer, 2001; Aujol et al., 2005):

$$\min_{m\in\mathbb{R}^{40}} h(m) := \tfrac{1}{2}\|m-\theta\|_2^2 + \alpha\|m\|_1. \tag{21}$$

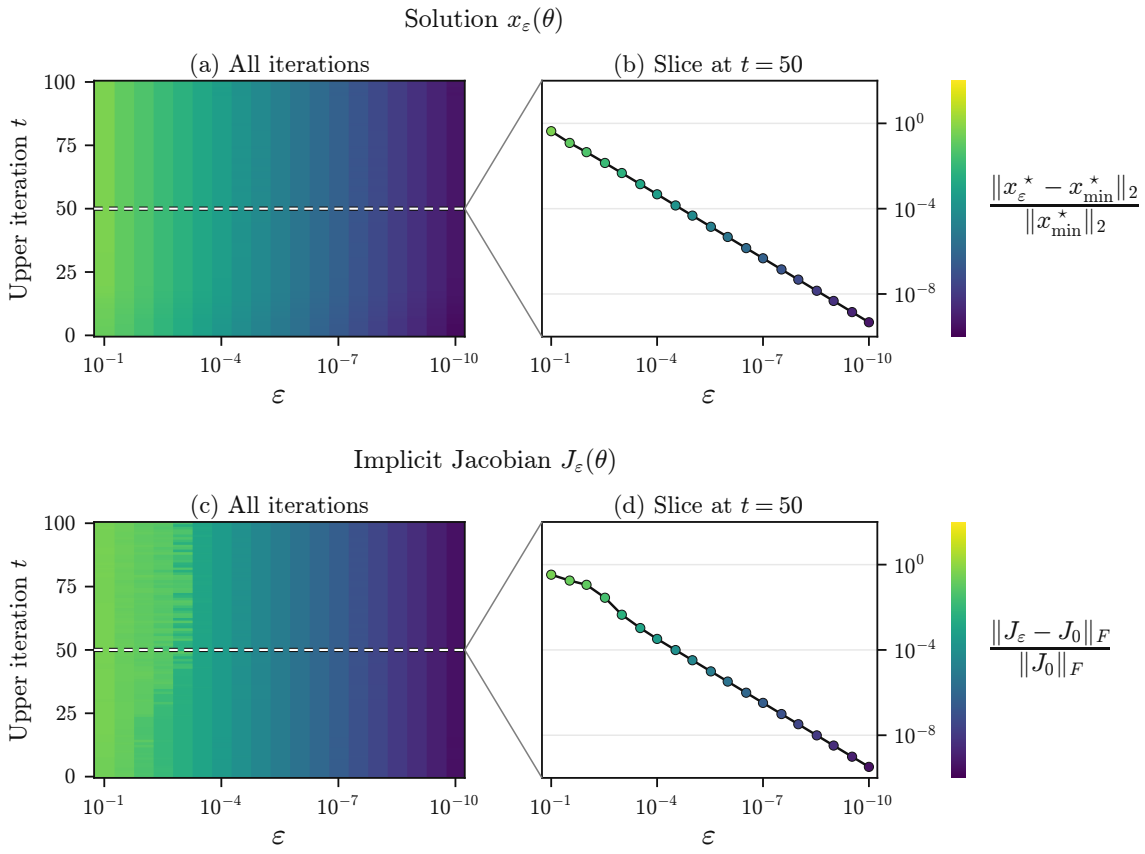


Figure 3: Huber continuation phenomenon. *Top*: convergence profile of $x_\varepsilon^\star(\theta)$. *Bottom*: convergence profile of $J_{x_\varepsilon^\star}(\theta)$. The heatmaps in (a) and (c) show the convergence of $x_\varepsilon^\star$ and J_ε as a function of ε, across 100 upper-level iterations. The plots in (b) and (d) show the behavior at iteration $t = 50$. A linear regime is observed for $x_\varepsilon^\star$; for $J_{x_\varepsilon^\star}$ this occurs only with smaller ε.

By strong convexity, (21) has a unique solution given by $m^\star = \operatorname{prox}_{\alpha\|\cdot\|_1}(\theta)$.

In Section A.3 (see Proposition 34), we detail that both (21) and (20) have same optimal value $h(m^\star)$ and the (Frobenius) minimal-norm solution of (20) is $(m^\star, m^\star)/2$. Using $x = (x_1; x_2)$, the Tikhonov regularization of (20) reads

$$\min_x f(x,\theta) + g(x,\theta) + \tfrac{\varepsilon}{2}\|x\|_2^2, \quad f(x,\theta) = \tfrac{1}{2}\|x_1 + x_2 - \theta\|_2^2,\ g(x,\theta) = \alpha\left(\|x_1\|_1 + \|x_2\|_1\right).$$

It can be shown (see Section A.3) that the regularized solution of (20) is $x_{1,\varepsilon} = m^\star/(2+\varepsilon) = x_{2,\varepsilon}$. The fixed-point formulation is, for $j \in \{1, 2\}$,

$$z_{j,\varepsilon} = x_j - \gamma(x_1 + x_2 - \theta + \varepsilon x_j), \quad (x_1, x_2) = \left(\operatorname{prox}_{\gamma\alpha}(z_{1,\varepsilon}), \operatorname{prox}_{\gamma\alpha}(z_{2,\varepsilon})\right) = G^\gamma. \tag{22}$$

By separability, we can choose α so that the solution $z_{j,\varepsilon}$ has a component equal to $\gamma\alpha$ whatever ε is. Thus, in $\operatorname{prox}_{\gamma\alpha\|\cdot\|_1}(z_{j,\varepsilon})$, $z_{j,\varepsilon}$ has a component always on the kink (see Section A.3 for explanations on this construction). In the notation of (7) from Section 4, in the $Q = \operatorname{diag}(Q_1, Q_2)$ block, the values corresponding to that coordinate can be chosen arbitrarily.

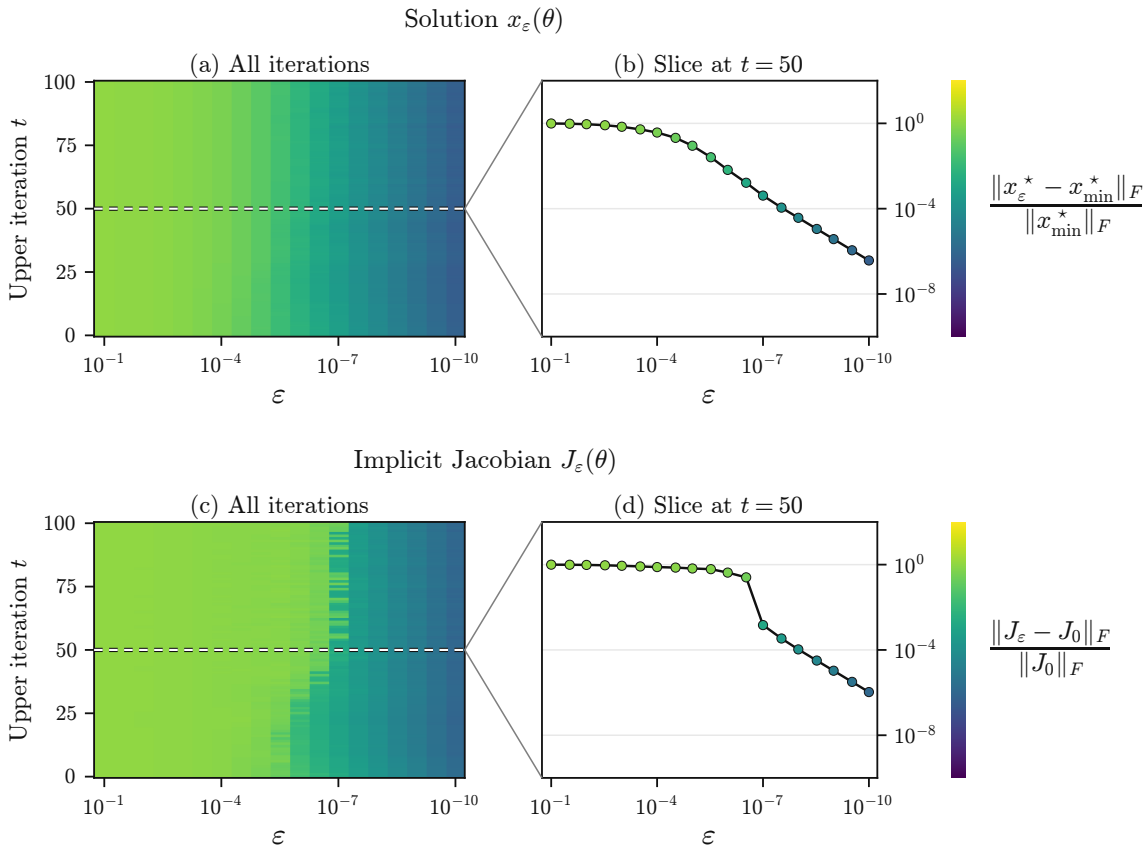


Figure 4: LASSO continuation phenomenon. *Top*: convergence profile of $x_\varepsilon^\star(\theta)$. *Bottom*: convergence profile of $J_{x_\varepsilon^\star}(\theta)$. The heatmaps in (a) and (c) show the convergence of $x_\varepsilon^\star$ and J_ε as a function of ε, across 100 upper-level iterations. The plots in (b) and (d) show the behavior at iteration $t = 50$. For both $x_\varepsilon^\star$ and $J_{x_\varepsilon^\star}$, linear convergence is observed for ε sufficiently small, possible indicating the identification of the support of $x_{\min}^\star$.

This construction allows us to create bias between x_1 and x_2 by choosing 1 for the corresponding coordinate of Q_1 and $(1 + \gamma\varepsilon^\delta)^{-1}$ for the one of Q_2, with $\delta \in \{1/2, 1, 2\}$ corresponding to the three regimes. This ε-dependent bias induces a difference between the limiting element and the element given by the pseudoinverse, first foretold in Example 8 and formally stated in Proposition 24 and Corollary 25. Complete details are given in Section A.3 and Section SM3 of the Supplementary Material.

In Figure 5 the three regimes (slow, critical and fast) display the expected behavior as $\varepsilon \to 0$. The fast regime plot showcases a perfect match between the pseudoinverse limit and what is computed along the way whereas the critical and slow regimes exhibit a difference that is smaller for the critical regime.

Let us mention this phenomenon is more a consequence of nonsmoothness than Tikhonov regularization and requires a specific choice of parameters. Nonetheless, it showcases (alongside Example 8) a situation where correctness of the pseudoinverse formula can be characterized through Assumptions (V_Σ) and **(V)**.

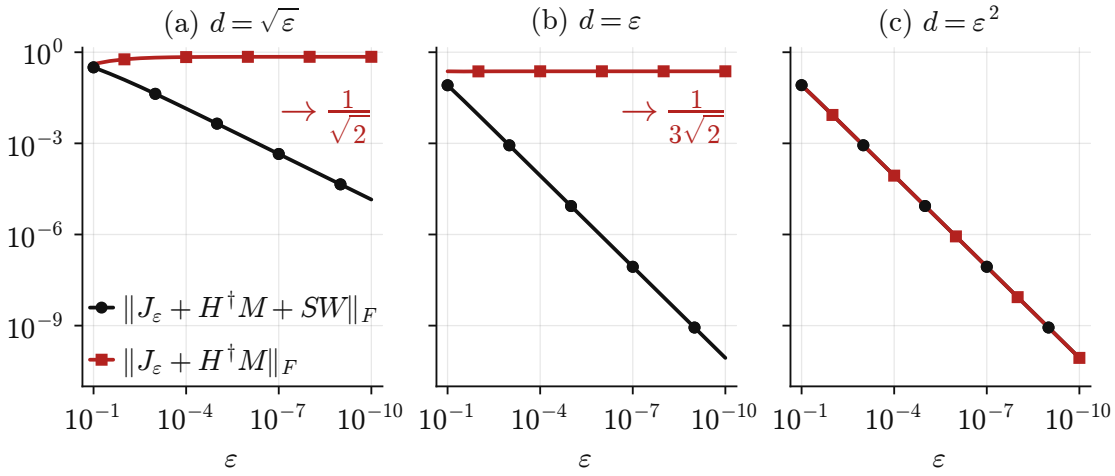


Figure 5: The plots (a), (b), and (c) correspond to the slow, critical, and fast vanishing of the eigenvalues, respectively. The two curves correspond to whether or not the correction term SW of Corollary 25 is included. As expected for the critical and slow regimes, the red curves do not go to zero because the correction term SW is nonzero.

6 Concluding Remarks

Our approach relies on the minimal-norm solution being path-differentiable, which can fail: in Section SM1 of the Supplementary Material we give a problem where it is not even continuous, in line with the hardness results of Bolte et al. (2026a). Moreover, the formulas we derive describe the derivatives obtained along the regularized path at a fixed parameter, which does not necessarily characterize all elements of the conservative Jacobian. An interesting direction for future research would be the identification of conditions yielding convergence rates of conservative Jacobians with respect to the Tikhonov regularization parameter.

Use of LLMs All of our experiments involved hyperparameter sweeps and designs whose implementations and parallelizations were aided through the use of ChatGPT-5.5 and Claude Opus 4.6. ChatGPT 5.6 Sol and Claude Fable 5 were used extensively to review and polish later drafts of the manuscript.

Acknowledgments and Disclosure of Funding

This work was supported by the French National Research Agency (ANR) under grant ANR-25-CE23-3749 (project SIMPLES). Antonio Silveti-Falls and Baptiste Plaquevent-Jourdain thank Sholom Schechtman for fruitful discussions.

Appendix A. Additional Experimental Results

The experiments use NumPy, JAX (Bradbury et al., 2018), scikit-learn (Pedregosa et al., 2011), and SciPy (Virtanen et al., 2020).

A.1 NYC 311 Huber Data Poisoning

Data and splits We use weeks 2–21 of 2023 in the NYC Open Data service-request records (NYC Open Data, 2026). Within each window, the first 250 requests form the lower-level and the next 1000 are validation candidates. Let $\bar{r}_{\mathrm{tr}}$ and s_{tr} be the mean and standard deviation of the 250 lower-level responses. The clean lower-level response is $\theta_0 = (r_{\mathrm{tr}} - \bar{r}_{\mathrm{tr}})/s_{\mathrm{tr}}$, and validation responses use the same transformation; we denote the resulting validation vector by b_{val} (they do not vary).

The lower-level design matrix A is the incidence matrix of the bipartite graph joining request types to incident ZIP codes. Its numerically computed rank ranges from 125 to 147 across the 20 windows.

Huber parameter and complete solution set We use the Huber loss defined in (Hub) and, following Huber (1964), we use $x_{\mathrm{LS}} = A^{\dagger}\theta_0$, $\widehat{\sigma} = (\frac{\|Ax_{\mathrm{LS}} - \theta_0\|_2^2}{250 - \mathrm{rank}(A)})^{1/2}$, and $\delta = 1.345\,\widehat{\sigma}$. The value of δ is computed from the clean lower partition and held fixed throughout the upper optimization, where the budget is fixed to be $\tau = 4.5$. We now present the minimal-norm computation; the dual representation and solution-set characterization are proved in Section SM2 of the Supplementary Material (Proposition SM2).

Proposition 33. *Let $\delta > 0$ be a positive constant, one has $h_\delta = \delta((\|\cdot\|_1)^* + \frac{\delta}{2}\|\cdot\|_2^2)^*$. Furthermore, the lower-level problem has a dual problem given by the following strongly convex problem: $u^\star(\theta) = \mathrm{argmin}_u \ \frac{1}{2}\|u+\theta\|_2^2$ such that $A^\top u = 0$ and $\|u\|_\infty \leq \delta$. Finally, define $I_0 = \{i : |u_i^\star| < \delta\}$, $I_+ = \{i : u_i^\star = \delta\}$, and $I_- = \{i : u_i^\star = -\delta\}$. The solution set of the (primal) lower level is given by $\mathcal{S}(\theta) = \{x \colon A_{I_0}x = \theta_{I_0} + u^\star_{I_0},\ A_{I_+}x \geq \theta_{I_+} + \delta\mathbf{1},\ A_{I_-}x \leq \theta_{I_-} - \delta\mathbf{1}\}$.*

Numerically, we can now solve the dual projection for $u^\star(\theta)$ which is a quadratic problem, construct $\mathcal{S}(\theta)$, and then compute the global minimal-norm solution from the secondary quadratic program, removing redundant equalities by SVD and using active-face polishing,

$$x^\star_{\min}(\theta) = \operatorname*{argmin}_{x \in \mathcal{S}(\theta)} \tfrac{1}{2}\|x\|_2^2. \tag{23}$$

Nonminimal solutions For the selections defined in Section 5.1, we draw ten unit directions $q_k \in \ker(A)$ per split and hold them fixed throughout each trajectory. These selections preserve the lower-level objective because $x^\star_{\min}(\theta) + \ker(A) \subseteq S(\theta)$.

Computing the implicit hypergradient At every upper iterate θ_t, we apply the continuation procedure of Section 5.1 on the grid $\{10^{-j} : j = 1, \ldots, 10\} \cup \{3 \cdot 10^{-j} : j = 2, \ldots, 10\}$, traversed in decreasing order. We use binary masks, setting $(Q_\varepsilon)_{ii} = 1$ when $|(Ax^\star_\varepsilon(\theta_t) - \theta_t)_i| < \delta$, and zero otherwise.

Solving the upper-level problem Each selection $x^\star_{c,k}$ defines a single-level objective $\Phi_{c,k}(\theta) := \frac{1}{2n_{\mathrm{val}}}\|A_{\mathrm{val}}x^\star_{c,k}(\theta) - b_{\mathrm{val}}\|_2^2$. Writing $\eta_t = \theta_t - \theta_0$ and taking $\theta_t = \theta_0 + \eta_t$ when evaluating the lower solution and hypergradient, and initialize $\eta_0 = v_0 = 0$. With h_t denoting the implicit hypergradient, the momentum algorithm reads as follows: $d_t = 0.01(4.5)\frac{h_t}{\|h_t\|_2 + 10^{-12}}$, $v_{t+1} = 0.9v_t + d_t$, and $\eta_{t+1} = \Pi(\eta_t + v_{t+1})$.

A.2 Digits LASSO Data Poisoning

Data and splits We sample a random subset of 1078 images from the scikit-learn Digits dataset to define the upper-level problem. For each of 20 seeds, we randomly sampled one image per digit class from this subset to construct the lower-level design matrix $A \in \mathbb{R}^{10\times 64}$.

Within these ten images, we averaged adjacent horizontal pixels and repeated each average, producing 32 exact duplicate column pairs.

Solving the lower-level problem The coordinate-descent solver includes active-manifold polishing once the support stabilizes.

Computing the implicit hypergradient For each output class c, we use the active set $S_c = \{j : |(X_N - \gamma[\frac{1}{10}A^\top(AX_N - \theta) + \varepsilon X_N])_{jc}| > \gamma e^\lambda\}$, taking the derivative of soft-thresholding to be zero at equality (on the kink). The gradient of the clean upper training loss with respect to the coefficients is $B = \nabla_X L_{\mathrm{tr}}(X_N) = \frac{1}{n_{\mathrm{tr}}} X_{\mathrm{tr}}^\top (X_{\mathrm{tr}} X_N - Y_{\mathrm{tr}})$. For each class c, the active-space adjoint $u_c \in \mathbb{R}^{|S_c|}$ is obtained from $(\frac{1}{10}A_{:,S_c}^\top A_{:,S_c} + \varepsilon I)u_c = B_{S_c,c}$ and $(\widehat{g}_\theta)_{:,c} = \frac{1}{10}A_{:,S_c}u_c$. If S_c is empty, we set $(\widehat{g}_\theta)_{:,c} = 0$. The ten columns form the hypergradient $\widehat{g}_\theta \in \mathbb{R}^{10\times 10}$ used in forward-backward. Since $\varepsilon > 0$, every active system is positive definite and is solved directly.

Solving the upper-level problem 100 iterations of projected normalized gradient ascent with momentum are applied like in the previous Huber data poisoning problem, with $\Delta_0 = 0$ and zero initial momentum.

Let us mention that despite the errors displaying encouraging decrease with ε, we have observed, on one seed, a Jacobian error of $\simeq 0.2$ for one iteration (that arose after the first 100 displayed ones).

A.3 Details on the Illustration of the Different Regimes

Now, we present the reasoning regarding the experiment of Section 5.3, details of the proofs being in Section SM3 of the Supplementary Material.

Data and splits We use the scikit-learn resized version of the Olivetti data set (Samaria and Harter, 1994). Every batch randomly selects one image for each of the 40 subjects, whose $64 \times 64 = 4096$ pixels are first normalized to $[0, 1]$ and then divided by $\sqrt{4096} = 64$. This leads to a matrix of size 4096×40, whose singular values are extracted, yielding $\theta \in \mathbb{R}^{40}_+$.

With h and $\tilde{h}$ defined in (21) and (20), respectively, set $v(m) := \min_{x_1+x_2=m} \tilde{h}(x_1, x_2)$ and $\tilde{h}_\varepsilon(x_1, x_2) := \tilde{h}(x_1, x_2) + \frac{\varepsilon}{2}(\|x_1\|_2^2 + \|x_2\|_2^2)$.

Proposition 34 (solution sets of the problems)**.** *Let $m^\star = \mathrm{prox}_{\alpha\|\cdot\|_1}(\theta)$ be the unique minimizer of* (21)*. Then $v = h$ and the solution set of* (20) *is $\mathcal{S}(m^\star) = \{(x_1, x_2) : x_1 + x_2 = m^\star, \|x_1\|_1 + \|x_2\|_1 = \|m^\star\|_1\}$. Moreover, $\min \tilde{h} = h(m^\star)$. Then, $\mathcal{S}(m^\star)$ is nonempty convex compact containing the segment $\{(tm^\star, (1-t)m^\star) : t \in [0, 1]\}$ and its (Euclidean) minimal-norm element is $(m^\star, m^\star)/2$. Finally, for every $\varepsilon > 0$,* argmin $\tilde{h}_\varepsilon = \{(m^\star, m^\star)/(2 + \varepsilon)\}$.

Details of the kink construction are given in Section SM3 of the Supplementary Material.

Appendix B. General Theorems

Theorem 35 (Clarke Jacobian of the proximal operator (Patrinos et al., 2014, theorem 3.2))**.** *Let $f \in \Gamma_0(\mathbb{R}^n)$ and $\gamma > 0$, for every $P \in \partial(\mathrm{prox}_{\gamma f})(x)$, P is symmetric positive semidefinite and verifies $\|P\| \leq 1$, that is, the eigenvalues of P are in $[0, 1]$.*

This supplies symmetry and spectral bounds but not definability and conservativity. When the proximal mapping is locally Lipschitz and definable, it is path-differentiable, thus its Clarke Jacobian is a conservative mapping. For a parameterized proximal mapping, the analogous conclusion requires these properties jointly in the input and parameter variables.

Theorem 36 (Hierarchical minimization). *Let $f \in \Gamma_0(\mathbb{R}^n)$ with $\mathcal{S} := \operatorname{argmin} f \neq \varnothing$ and $x^\star_{\min} := \Pi_{\mathcal{S}}(0) = \operatorname{argmin}_{x\in\mathcal{S}} \|x\|_2$. For $\varepsilon > 0$, let $x^\star_\varepsilon$ be the unique minimizer of $f + \varepsilon\|\cdot\|_2^2/2$. Then, for $\varepsilon > 0$ and $0 < \varepsilon_1 \leq \varepsilon_2$, $\|x^\star_\varepsilon\|_2 \leq \|x^\star_{\min}\|_2$, $\|x^\star_{\varepsilon_2}\|_2 \leq \|x^\star_{\varepsilon_1}\|_2 \leq \|x^\star_{\min}\|_2$, and $\lim_{\varepsilon\to 0} \|x^\star_\varepsilon - x^\star_{\min}\|_2 = 0$.*

We now present a recent result giving, under a Hölderian assumption, a worst-case upper bound on convergence speed.

Definition 37 (Hölderian error bound). A proper function f with a nonempty set of minimizers $\mathcal{S}$ is said to verify a Hölderian error bound with exponent $p \geq 1$ if there exist constants $r > \min f$ and $\kappa > 0$ such that, when $f(x) < r$, $f(x) - \min f \geq \kappa \operatorname{dist}(x, \mathcal{S})^p$.

Theorem 38 (Rate with Hölderian error bound (Maulen-Soto et al., 2025, Proposition 4.10)). *Let $f \in \Gamma_0(\mathbb{R}^n)$ have a nonempty set of minimizers $\mathcal{S}$ and satisfy a Hölderian error bound with exponent p as in* Definition 37. *Let $x^\star_{\min} := \Pi_{\mathcal{S}}(0)$ and $x^\star_\varepsilon$ be the unique minimizer of $f + \varepsilon\|\cdot\|^2/2$, then there exist constants $\bar\varepsilon$ and C_0 such that, for all $\varepsilon \in]0, \bar\varepsilon]$, $\|x^\star_\varepsilon - x^\star_{\min}\| \leq C_0 \varepsilon^{\frac{1}{2p}}$.*

Appendix C. Definability of the Tikhonov Construction

We briefly recall definability, then prove that the Tikhonov solutions and regularized Jacobians form definable families under our assumptions. See van den Dries and Miller (1996); Coste (1999); Ioffe (2009) for detailed introductions to definability and o-minimal geometry.

C.1 A Glimpse of Definability

Definition 39 (o-minimal structure). An o-minimal structure on $(\mathbb{R}, +, \times)$ is a collection $\mathcal{O} = (\mathcal{O}_p)_{p\in\mathbb{N}}$ such that each $\mathcal{O}_p$ is a collection of subsets of $\mathbb{R}^p$ verifying the following axioms:

- $\mathcal{O}_1$ consists exactly of the finite unions of points and open intervals;
- $\mathcal{O}_p$ contains $\mathbb{R}^p$ and is stable by complement, finite union, and finite intersection;
- $A \in \mathcal{O}_p \Rightarrow A \times \mathbb{R} \in \mathcal{O}_{p+1}$ and $\mathbb{R} \times A \in \mathcal{O}_{p+1}$;
- if $A \in \mathcal{O}_{p+1}$ and $\pi : \mathbb{R}^{p+1} \to \mathbb{R}^p$ is the canonical projection onto $\mathbb{R}^p$, then $\pi(A) \in \mathcal{O}_p$;
- $\mathcal{O}_p$ contains the real algebraic subsets of $\mathbb{R}^p$, that is, the sets of the form $\{x \in \mathbb{R}^p : g(x) = 0\}$ where g is a polynomial.

A subset of $\mathbb{R}^p$ is definable if it belongs to $\mathcal{O}_p$. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is definable if its graph is definable.

Important examples include semialgebraic sets, which are finite unions of sets of the form $\bigcap_{i=1}^k \{x \in \mathbb{R}^p : g_i(x) < 0, h_i(x) = 0\}$ with g_i, h_i polynomial functions. The logarithm and exponential are also definable in suitable o-minimal structures. Definability is preserved by standard operations, including composition. The definable class therefore contains semialgebraic functions, neural networks built from definable elementary functions, and outputs of backpropagation (Bolte and Pauwels, 2021, Section 5).

C.2 Definability of the Tikhonov Solution Maps

Proposition 40 (Definability of the Tikhonov solution maps). *Assume* **(Reg)** *and* **(D)**. *Then $(\theta, \varepsilon) \mapsto x^\star_\varepsilon(\theta)$ is definable on $\Theta \times \mathbb{R}^*_+$. The minimal-norm solution $x^\star_{\min}$ and the joint selection $x^\star$ defined in* Section 2.2 *are also definable.*

Proof For $\varepsilon \geq 0$, set $f_\varepsilon(x,\theta) := f(x,\theta) + \varepsilon\|x\|_2^2/2$ and consider the full solution graph $\mathcal{T} := \{(\theta,\varepsilon,x) \in \Theta \times \mathbb{R}_+ \times \mathbb{R}^n : x \in \operatorname{argmin}_y f_\varepsilon(y,\theta)\}$. The set $\mathcal{A} := \{(\theta,\varepsilon,x,y) \in \Theta\times\mathbb{R}_+\times\mathbb{R}^n\times\mathbb{R}^n : f_\varepsilon(y,\theta) < f_\varepsilon(x,\theta)\}$ is definable, and $\mathcal{T} = (\Theta\times\mathbb{R}_+\times\mathbb{R}^n)\setminus\pi_{\theta,\varepsilon,x}(\mathcal{A})$. Since definable sets are stable under projection and complement, $\mathcal{T}$ is definable. For every $\theta\in\Theta$ and $\varepsilon > 0$, the regularized objective $f_\varepsilon(\cdot,\theta)$ is ε-strongly convex and therefore has a unique minimizer. Thus $\operatorname{gph}\big((\theta,\varepsilon)\mapsto x^\star_\varepsilon(\theta)\big) = \mathcal{T}\cap(\Theta\times\mathbb{R}^*_+\times\mathbb{R}^n)$ is definable. The unregularized solution correspondence has graph $\{(\theta,x) : (\theta,0,x)\in\mathcal{T}\}$. Its values are nonempty, closed, and convex, so each contains a unique element of minimal norm. The graph of $x^\star_{\min}$ is $\{(\theta,x) : (\theta,0,x)\in\mathcal{T}$ and there is no y such that $(\theta,0,y)\in\mathcal{T}$ and $\|y\|_2^2 < \|x\|_2^2\}$, which is definable. Finally, $\operatorname{gph}(x^\star)$ is the union of the graph of $(\theta,\varepsilon)\mapsto x^\star_\varepsilon(\theta)$ on $\varepsilon>0$ and the graph of $x^\star_{\min}$ at $\varepsilon = 0$, hence it is definable as well. ■

C.3 Definability of the Regularized Conservative Jacobians

We now give the corresponding definability statement for the conservative Jacobians obtained by implicit differentiation. We first recall an elementary fact used in the proof of Theorem 23 (see Bolte et al., 2007).

Lemma 41 (Definability of Clarke Jacobians)**.** *Let $U\subset\mathbb{R}^d$ be open and definable, and let $G : U\to\mathbb{R}^q$ be locally Lipschitz and definable. Then the Clarke Jacobian $\partial G : U \rightrightarrows \mathbb{R}^{q\times d}$ has definable graph.*

Proposition 42 (Definability of the regularized implicit conservative Jacobians)**.** *Assume* **(Reg)**, **(D)** *and* **(R)**. *For $0<\varepsilon\leq\bar\varepsilon$, let $J_{x^\star_\varepsilon}$ be the implicit conservative Jacobian in* (3)*, and let $J_{x^\star}(\theta,\varepsilon) := J_{x^\star_\varepsilon}(\theta)$ be the associated family defined in* Section 2.2. *Then $J_{x^\star}$ has definable graph on $O\times]0,\bar\varepsilon]$, and every fixed-$\varepsilon$ slice $J_{x^\star_\varepsilon}$ is a conservative Jacobian of $\theta\mapsto x^\star_\varepsilon(\theta)$ on O.*

Proof By Proposition 40, $(\theta,\varepsilon)\mapsto x^\star_\varepsilon(\theta)$ is definable on $O\times]0,\bar\varepsilon]$. Moreover, $D\in J_{x^\star}(\theta,\varepsilon)$ iff there exists $(H_\varepsilon\ M_\varepsilon)\in J_F(x^\star_\varepsilon(\theta),\theta)$ such that $(H_\varepsilon+\varepsilon I_n)D + M_\varepsilon = 0$. By Assumption **(R)**, J_F has definable graph, and the matrix equation above is polynomial in $(\varepsilon, H_\varepsilon, M_\varepsilon, D)$. Thus $\operatorname{gph}(J_{x^\star})$ is the projection of a definable set and is therefore definable.

For fixed $\varepsilon>0$, the regularized objective is definable and ε-strongly convex. Hence Bolte et al. (2024, Theorem 4.2) shows that $x^\star_\varepsilon$ is path-differentiable, in particular locally Lipschitz. Assumption **(R)** makes every $H_\varepsilon+\varepsilon I_n$ along the regularized solutions positive definite, so Theorem 2 gives (3). Thus $J_{x^\star_\varepsilon}$ is a conservative Jacobian of $x^\star_\varepsilon$ on O. ■

Appendix D. Results from Section 3

This section first justifies the general formula for limit elements then the pseudoinverse case, then details the computations used for the problem with the Huber loss.

D.1 Proof of the Residual Formula Theorem 10

Proof (Theorem 10) Fix the sequence $\varepsilon_k\downarrow 0$ and the blocks $(H_k\ M_k)$ from the theorem. Assumption **(R)**, applied with $K=\{\theta\}$, gives a constant C_θ and matrices $\widehat{W}_k$ such that $M_k = H_k\widehat{W}_k$ and $\|\widehat{W}_k\|\leq C_\theta$. Since H_k is symmetric PSD, $P_k := H_k^\dagger H_k$ is the orthogonal projector onto $\operatorname{range}(H_k)$, and therefore $W_k = H_k^\dagger M_k = P_k\widehat{W}_k$, $\|W_k\|\leq C_\theta$, and $H_kW_k = M_k$. The eigenvalues of $R_k = (H_k+\varepsilon_k I_n)^{-1}H_k$ are $\lambda_i(H_k)/(\lambda_i(H_k)+\varepsilon_k)\in[0,1]$. Hence $0\preceq R_k\preceq I_n$, and $V_k = -R_kW_k$ is bounded as well.

Let V be a cluster point and pass to a subsequence along which $V_k\to V$. By Theorem 36, $x^\star_{\varepsilon_k}(\theta)\to x^\star_{\min}(\theta)$. Using the local boundedness of J_F and the boundedness of $(W_k)_k$ and

$(R_k)_k$, we may pass to a further subsequence such that $(H_k\ M_k) \to (H\ M)$, $W_k \to W$, and $R_k \to R$. The chosen sequence and the closed graph property give $(H\ M) \in \mathcal{R}(\theta)$, and therefore $M = HW$, $V = -RW$, and $0 \preceq R \preceq I_n$.

The matrices H_k and R_k commute, and $\|H_kR_k - H_k\| = \max_i \frac{\varepsilon_k\lambda_i(H_k)}{\lambda_i(H_k)+\varepsilon_k} \leq \varepsilon_k$. Passing to the limit yields $HR = H = RH$. Thus R acts as the identity on $\mathrm{range}(H)$ and leaves $\ker(H)$ invariant. With $P := H^\dagger H = \Pi_{\mathrm{range}(H)}$, we obtain $S := R - P = \Pi_{\ker H} R\, \Pi_{\ker H}$ and $0 \preceq S \preceq \Pi_{\ker H}$. Since $H^\dagger M = PW$, we conclude that $V = -RW = -H^\dagger M - SW$. Then, since $V_k \in J_{x^\star}(\theta, \varepsilon_k)$, taking $\theta_k \equiv \theta$ in (2) gives $V \in J_{x^\star_{\min}}(\theta)$. Finally, $M = HW$ and $HS = 0$, so $HV + M = -HH^\dagger M - HSW + M = 0$. ■

D.2 Proof of the Pseudoinverse Formula Theorem 12

Proof (Theorem 12) Fix $(H\ M) \in \mathcal{R}(\theta)$, and choose $\varepsilon_k \downarrow 0$ and $(H_k\ M_k) \in J_F(x^\star_{\varepsilon_k}(\theta), \theta)$ realizing this limit. Discarding finitely many terms, we may assume that $0 < \varepsilon_k \leq \bar{\varepsilon}$. Define $R_k := (H_k + \varepsilon_k I_n)^{-1} H_k, W_k := H_k^\dagger M_k$, and $V_k := -(H_k + \varepsilon_k I_n)^{-1} M_k = -R_k W_k$. Assumption **(R)**, applied with $K = \{\theta\}$, gives $C_\theta > 0$ and matrices $\widehat{W}_k$ such that $M_k = H_k\widehat{W}_k$ and $\|\widehat{W}_k\| \leq C_\theta$. Since $H_k^\dagger H_k$ is an orthogonal projector, $W_k = H_k^\dagger H_k \widehat{W}_k$ is bounded and $H_kW_k = M_k$. Passing to a subsequence, we may assume that $W_k \to W$, and hence $M = HW$.

We claim that $R_k \to H^\dagger H$. Let $H_k = \sum_{i=1}^n \lambda_i(H_k) u_{i,k} u_{i,k}^\mathsf{T}$ be a spectral decomposition with $\lambda_1(H_k) \geq \cdots \geq \lambda_n(H_k) \geq 0$. Then $R_k = \sum_{i=1}^n \frac{\lambda_i(H_k)}{\lambda_i(H_k)+\varepsilon_k} u_{i,k} u_{i,k}^\mathsf{T}$. Suppose first that **(G)** holds. For large k, we also have $\varepsilon_k \leq \varepsilon_G$. With $P_k := \Pi_{\mathrm{range}(H_k)}$, $\|R_k - P_k\| = \max_{i:\lambda_i(H_k)>0} \frac{\varepsilon_k}{\lambda_i(H_k)+\varepsilon_k} \leq \frac{\varepsilon_k}{\lambda^*} \to 0$, where the maximum is taken to be 0 if $H_k = 0$. The fixed spectral gap and $H_k \to H$ also give, using that the number of eigenvalues of H_k above λ^* is eventually identical to $\mathrm{rank}(H)$, $P_k \to \Pi_{\mathrm{range}(H)} = H^\dagger H$.

Suppose instead that **(V)** holds, and set $r := \mathrm{rank}(H)$. The first r eigenvalues of H_k converge to the positive eigenvalues of H, while the remaining eigenvalues converge to zero. If P_k^r denotes the spectral projector onto the first r eigenspaces, with $P_k^0 := 0$, then $\|R_k - P_k^r\| = \max\{\max_{i\leq r} \frac{\varepsilon_k}{\lambda_i(H_k)+\varepsilon_k},\ \max_{i>r} \frac{\lambda_i(H_k)}{\lambda_i(H_k)+\varepsilon_k}\} \to 0$. We take for convention the maximum over an empty index set to be 0. The first maximum tends to 0 because the limiting eigenvalues are positive, while the second tends to 0 by **(V)**. If $r = 0$ or $r = n$, then $P_k^r \to \Pi_{\mathrm{range}(H)}$ is immediate because $P_k^0 = 0$ and $P_k^n = I_n$. If $0 < r < n$, the positive and null spectra of H are separated, so the continuity of the corresponding spectral projector gives the same conclusion. Thus $P_k^r \to \Pi_{\mathrm{range}(H)} = H^\dagger H$, which proves the claim in both cases.

It follows that $V_k = -R_kW_k \to -H^\dagger HW = -H^\dagger M$. Since $V_k \in J_{x^\star}(\theta, \varepsilon_k)$ and $x^\star_{\varepsilon_k}(\theta) \to x^\star_{\min}(\theta)$ by Theorem 36, the definition (2) gives $-H^\dagger M \in J_{x^\star_{\min}}(\theta)$. The reachable block $(H\ M)$ was arbitrary, which proves the result. ■

D.3 Uniform Bounds for Weighted Gram Systems

The proof of Corollary 14 requires a general estimate for Gram systems regularized by a positive diagonal matrix. We state and prove this estimate first in Lemma 43, because it will also be used for analyzing the LASSO problem, then we prove Corollary 14.

Lemma 43 (Uniform bound for diagonally regularized Gram systems)**.** *Let $r, s \in \mathbb{N}$ with $r, s \geq 1$ and $B \in \mathbb{R}^{r\times s}$, and let $\chi(B)$ denote the quantity defined in* (5) *with A replaced by B. Then, for every positive definite diagonal matrix $D \in \mathbb{R}^{s\times s}$, $\|(D + B^\mathsf{T}B)^{-1}B^\mathsf{T}\| \leq \chi(B)$.*

Proof (Lemma 43) Fix $v \in \mathbb{R}^r$ and write $u = (D + B^\mathsf{T}B)^{-1}B^\mathsf{T}v$. We first represent u as part of a weighted minimal-norm solution of a linear system. Note the identity

$$(D + B^\mathsf{T}B)^{-1}B^\mathsf{T} = D^{-1}B^\mathsf{T}(I_r + BD^{-1}B^\mathsf{T})^{-1} \tag{24}$$

which follows by multiplying both sides by $D+B^\mathsf{T}B$ and rearranging terms. Set $A_0 := (B\ I_r)$ and $\Omega := \operatorname{diag}(D^{-1}, I_r)$. Then $I_r + BD^{-1}B^\mathsf{T} = A_0\Omega A_0^\mathsf{T}$, and (24) gives $u = (I_s\ 0)z^\star$ and $z^\star := \Omega A_0^\mathsf{T}(A_0\Omega A_0^\mathsf{T})^{-1}v$. Equivalently, $z^\star$ is the unique solution of the *weighted minimal-norm problem* $\min_{z\in\mathbb{R}^{s+r}} \ \frac{1}{2}z^\mathsf{T}\Omega^{-1}z$ subject to $A_0z = v$.

We next use the fact that $z^\star$ is a convex combination of the basic solutions of $A_0z = v$. To see this, let the columns of $N \in \mathbb{R}^{(s+r)\times s}$ form a basis of $\ker(A_0)$, and take $z_p = (0; v)$ as a particular solution. Every feasible point has the form $z_p + N\beta$, so the weighted minimal-norm problem is a diagonally weighted Least-Squares problem in β. The geometric description of weighted Least-Squares solutions in Ben-Tal and Teboulle (1990) gives, for $\beta^\star$ such that $z^\star = z_p + N\beta^\star$, $\beta^\star \in \operatorname{conv}\{-(N_{\mathcal{I},:})^{-1}(z_p)_\mathcal{I} : |\mathcal{I}| = s,\ N_{\mathcal{I},:} \text{ nonsingular}\}$. For every such $\mathcal{I}$, the corresponding point $\zeta^\mathcal{I} := z_p - N(N_{\mathcal{I},:})^{-1}(z_p)_\mathcal{I}$ is feasible and vanishes on $\mathcal{I}$. Moreover, $A_{0,:,\mathcal{I}^c}$ is nonsingular. Indeed, otherwise a nonzero vector supported on $\mathcal{I}^c$ would lie in $\ker(A_0)$, and its representation $N\beta$ would satisfy $N_{\mathcal{I},:}\beta = 0$, contradicting the nonsingularity of $N_{\mathcal{I},:}$. Thus $\zeta^\mathcal{I}$ is the basic solution supported on $\mathcal{I}^c$, which is a basis of A_0. Since $\beta \mapsto z_p + N\beta$ is affine, we have proved that

$$z^\star \in \operatorname{conv}\{z^\mathcal{J} : \mathcal{J} \text{ is a basis of } A_0\}. \tag{25}$$

It remains to bound the first block of each basic solution. Any basis of $A_0 = (B\ I_r)$ can be written as $\mathcal{J} = \mathcal{I}_1 \cup \{s+k : k \in K\}$, where $\mathcal{I}_1 \subset [1:s]$, $K \subset [1:r]$, and $|\mathcal{I}_1| + |K| = r$. Expanding along the selected identity columns shows that $B_{K^c,\mathcal{I}_1}$ is nonsingular. If $z^\mathcal{J} = (u'; w)$ is the corresponding basic solution, then u' is supported on $\mathcal{I}_1$, $\begin{pmatrix} B_{K^c,\mathcal{I}_1} & I_{K^c,K} \\ B_{K,\mathcal{I}_1} & I_{K,K} \end{pmatrix}\begin{pmatrix} u' \\ w \end{pmatrix} = \begin{pmatrix} v_{K^c} \\ v_K \end{pmatrix}$ and the rows indexed by K^c give $(u')_{\mathcal{I}_1} = (B_{K^c,\mathcal{I}_1})^{-1}v_{K^c}$ (since $I_{K^c,K} = 0$). Consequently, $\|(I_s\ 0)z^\mathcal{J}\| \le \chi(B)\|v\|$. This also covers $\mathcal{I}_1 = \emptyset$, when the first block is zero. Combining this bound with (25) yields $\|(D + B^\mathsf{T}B)^{-1}B^\mathsf{T}v\| = \|(I_s\ 0)z^\star\| \le \chi(B)\|v\|$, which proves the lemma. ∎

Proof (Corollary 14) Fix $u \in \mathbb{R}^m$, and set $B := \sqrt{Q}A$ and $y := \sqrt{Q}u$. Then $(\delta I_n + A^\mathsf{T}QA)^{-1}A^\mathsf{T}Qu = (\delta I_n + B^\mathsf{T}B)^{-1}B^\mathsf{T}y$. The proof of Lemma 43, with $D = \delta I_n$, represents the right-hand side as the first block of a convex combination of basic solutions of $(B\ I_m)z = y$. Consider one such basic solution, corresponding to $\mathcal{J} = \mathcal{I}_1 \cup \{n + k : k \in K\}$. Its first block is supported on $\mathcal{I}_1$, and the rows indexed by K^c give $(z^\mathcal{J})_{\mathcal{I}_1} = (B_{K^c,\mathcal{I}_1})^{-1}y_{K^c}$. Here $|\mathcal{I}_1| = |K^c|$. Nonsingularity of $B_{K^c,\mathcal{I}_1} = \sqrt{Q_{K^c}}A_{K^c,\mathcal{I}_1}$ implies that Q_{K^c} is positive definite and $A_{K^c,\mathcal{I}_1}$ is nonsingular. Since $y_{K^c} = \sqrt{Q_{K^c}}u_{K^c}$, the matching row weights cancel: $(z^\mathcal{J})_{\mathcal{I}_1} = (\sqrt{Q_{K^c}}A_{K^c,\mathcal{I}_1})^{-1}\sqrt{Q_{K^c}}u_{K^c} = (A_{K^c,\mathcal{I}_1})^{-1}u_{K^c}$. Thus the first block of every basic solution has norm at most $\chi(A)\|u\|$; when $\mathcal{I}_1 = \emptyset$, it is zero. The same bound holds for their convex combination, proving $\|(\delta I_n + A^\mathsf{T}QA)^{-1}A^\mathsf{T}Q\| \le \chi(A)$.

For fixed Q, an SVD of $B = \sqrt{Q}A$ gives $(\delta I_n + B^\mathsf{T}B)^{-1}B^\mathsf{T}\sqrt{Q} \to B^\dagger\sqrt{Q}$ as $\delta \downarrow 0$. Passing to the limit in the first estimate proves the second one. ∎

Remark 44 (Scaled projections and pseudoinverses)**.** The estimate in Lemma 43 uses arguments common to results on scaled projections and pseudoinverses. For a full row rank

matrix A_0, the weighted pseudoinverses $\Omega A_0^\mathsf{T}(A_0\Omega A_0^\mathsf{T})^{-1}$ are bounded uniformly over positive definite diagonal matrices Ω (Stewart, 1989; O'Leary, 1990; Forsgren, 1996). This is used in the condition measure $\bar{\chi}_{A_0}$ often employed for complexity analysis of interior-point methods (Vavasis and Ye, 1996). Such condition measures can be large but we only need a uniform bound. To the best of our knowledge, Lemma 43 does not already appear somewhere in this literature, although the underlying arguments are not unknown.

D.4 Path-differentiability for Huber Regression Corollary 15

Proof (Corollary 15) Let $f(x,\theta) := \phi(Ax-\theta)$. The Huber loss is finite, convex, semialgebraic, and $\mathcal{C}^{1,1}$. Moreover, $h(t) \geq |t| - 1/2$, so ϕ is coercive. Minimizing ϕ over the closed affine space $\mathrm{range}(A) - \theta$ therefore gives a solution of (Hub) for every $\theta \in \mathbb{R}^m$. Since $\nabla\phi$ is 1-Lipschitz, $\|F(x,\theta) - F(x',\theta')\| \leq \|A\|\big(\|A\|\|x-x'\| + \|\theta-\theta'\|\big)$, and Assumptions **(Reg)** and **(D)** hold with $\Theta = \mathbb{R}^m$. The chain-rule construction and calculation in Section 3.4 verify Assumption **(R)** globally for any $\bar{\varepsilon} > 0$, with the uniform bound $\|W_\varepsilon\| \leq \chi(A)$. The conclusion follows from Theorem 7. ■

D.5 Proof of the Huber Limiting-Derivative Formula Proposition 16

Proof (Proposition 16) Choose $\varepsilon_k \downarrow 0$ and $Q_{\varepsilon_k} \to Q$ as in the definition of $\mathcal{R}^{\mathrm{Hub}}(\theta)$, and write $y_k := Ax^\star_{\varepsilon_k}(\theta) - \theta$. The coordinatewise description of $\partial(\nabla\phi)$ shows that, for all large k, the constant matrix Q itself belongs to $\partial(\nabla\phi)(y_k)$. If $(Q)_{ii} \in]0,1[$, then eventually $(Q_{\varepsilon_k})_{ii} \in]0,1[$ and hence $|(y_k)_i| = 1$. If $(Q)_{ii} = 0$, then eventually $(Q_{\varepsilon_k})_{ii} < 1$ and hence $|(y_k)_i| \geq 1$. Finally, if $(Q)_{ii} = 1$, then eventually $(Q_{\varepsilon_k})_{ii} > 0$ and hence $|(y_k)_i| \leq 1$.

Set $B := \sqrt{Q}A$. For all large k, we may therefore select $V_k := (\varepsilon_k I_n + B^\mathsf{T}B)^{-1}B^\mathsf{T}\sqrt{Q} \in J_{x^\star}(\theta,\varepsilon_k)$. An SVD of B gives $V_k \to B^\dagger\sqrt{Q}$. Together with $x^\star_{\varepsilon_k}(\theta) \to x^\star_{\min}(\theta)$, the definition (2) yields $B^\dagger\sqrt{Q} \in J_{x^\star_{\min}}(\theta)$. Finally, since $B^\mathsf{T}B = A^\mathsf{T}QA$, $B^\mathsf{T}\sqrt{Q} = A^\mathsf{T}Q$, and $(B^\mathsf{T}B)^\dagger B^\mathsf{T} = B^\dagger$ by defining properties of the pseudoinverse, we have $B^\dagger\sqrt{Q} = (A^\mathsf{T}QA)^\dagger A^\mathsf{T}Q$, which proves the claim. ■

Appendix E. Results from Section 4

We recall problem (P-Comp), its Tikhonov regularization (P-Comp$_\varepsilon$), and its fixed-point formulation given in Section 4. For $D = (Q,P,H^f,M^f) \in \mathcal{D}_\varepsilon(x,\theta)$, the blocks of the conservative Jacobian (8) of the residual Φ^γ_ε and reduced operators $\Pi_Q, H_\varepsilon, M_\varepsilon, \Sigma, V$, and T_ε are denoted using the notation given in Section 4.2.

The subsections below follow the order of the results in Section 4: the regularized conservative Jacobian, the LASSO uniform derivative bound, projected inversion, composite path-differentiability and cluster-point formulas, and the reachable binary LASSO derivatives.

E.1 Path-differentiability of Regularized Composite Problems Proposition 18

Proof (Proposition 18) Fix $0 < \varepsilon \leq \bar{\varepsilon}$. The smooth function $x \to f(x,\theta) + \frac{\varepsilon}{2}\|x\|_2^2$ is ε-strongly convex and has an $(L+\varepsilon)$-Lipschitz gradient. The Tikhonov term and its unique minimizer are definable, and the joint Clarke Jacobians are definable by Lemma 41; hence all the objects in the fixed-point construction and in (9) are definable. Moreover, the fixed stepsize satisfies $0 < \gamma < \frac{2}{L+2\bar{\varepsilon}} < \frac{2}{L+\bar{\varepsilon}} \leq \frac{2}{L+\varepsilon}$. The gradient and the parameterized proximal map are therefore jointly path-differentiable. The stepsize argument of Bolte et al. (2024, Remark 3.6) applies in the same way when strong convexity lies in the smooth term. Thus Bolte et al. (2024,

Theorem 3.5), applied componentwise on O, gives the path-differentiability of $\theta \mapsto x_\varepsilon^\star(\theta)$ and exactly the definable conservative Jacobian in (9). ■

E.2 Proof of the LASSO Uniform Derivative Bound Proposition 20
We use the problem notation introduced in Section 4.1.

Proof (Proposition 20) If $Q = 0$ then the bound is trivial. Otherwise, permute the coordinates of x so that S_+ comes first; this multiplies the matrices below on the left by a permutation matrix and permutes the columns of A, which changes neither operator norms nor $\chi(A)$. Then $Q = \operatorname{diag}(Q_{S_+,S_+}, 0)$ with $Q_{S_+,S_+} \succ 0$, $\gamma Q A^\top = \begin{pmatrix} \gamma Q_{S_+,S_+} A_{:,S_+}^\top \\ 0 \end{pmatrix}$ and $I_n - Q + \gamma Q(A^\top A + \varepsilon I_n) = \begin{pmatrix} I_{|S_+|} - Q_{S_+,S_+} + \gamma\varepsilon Q_{S_+,S_+} + \gamma Q_{S_+,S_+} A_{:,S_+}^\top A_{:,S_+} & \gamma Q_{S_+,S_+} A_{:,S_+}^\top A_{:,S_+^c} \\ 0 & I_{n-|S_+|} \end{pmatrix}$.

Inverting blockwise the upper triangular matrix shows that $\big(I_n - Q(I_n - \gamma(A^\top A + \varepsilon I_n))\big)^{-1}\gamma Q A^\top$ vanishes on the rows outside S_+ and, on the rows indexed by S_+, is equal to $\big(I_{|S_+|} - Q_{S_+,S_+} + \gamma\varepsilon Q_{S_+,S_+} + \gamma Q_{S_+,S_+} A_{:,S_+}^\top A_{:,S_+}\big)^{-1}\gamma Q_{S_+,S_+} A_{:,S_+}^\top$. Factoring $\gamma Q_{S_+,S_+} \succ 0$ on the left shows this to be the matrix in (11). For $\varepsilon > 0$, the diagonal matrix $\gamma^{-1}(Q_{S_+,S_+}^{-1} - I_{|S_+|}) + \varepsilon I_{|S_+|}$ is positive definite since entries of Q_{S_+,S_+} lie in $]0,1]$, and Lemma 43 with $B = A_{:,S_+}$ and this matrix as D gives (11); the last inequality in (11) obviously holds since every square submatrix of $A_{:,S_+}$ is a square submatrix of A. For $\varepsilon = 0$, $\operatorname{range}(A_{:,S_+}^\top) = \operatorname{range}(A_{:,S_+}^\top A_{:,S_+})$ is contained in the range of the positive semidefinite matrix $\gamma^{-1}(Q_{S_+,S_+}^{-1} - I_{|S_+|}) + A_{:,S_+}^\top A_{:,S_+}$, since the kernel of a sum of positive semidefinite matrices is the intersection of their kernels; a spectral decomposition of that matrix then shows that the matrix in (11) at $\varepsilon = 0$ is the limit of its counterparts as $\varepsilon \downarrow 0$, and the bound survives the limit. Finally, by Corollary 19 and the description of $\partial \operatorname{prox}_{\gamma e^\lambda \|\cdot\|_1}$ in Section 4.1, every element of $J_{x_\varepsilon^\star}(\theta)$ is such a matrix for some $Q \in \operatorname{diag}([0,1]^n)$, hence has norm at most $\chi(A)$, whence $\sup\{\|J\| : J \in J_{x_\varepsilon^\star}(\theta),\ 0 < \varepsilon \le \varepsilon_0,\ \theta \in K\} \le \chi(A)$ for every compact K and $0 < \varepsilon_0 \le \bar{\varepsilon}$, which is condition (iv) of Theorem 3. ■

E.3 Proof of Projected-Inversion Lemma 21

Proof (Lemma 21) Since Q is symmetric, $\Pi = QQ^\dagger$ is the orthogonal projector onto $\operatorname{range}(Q)$ with $Q^\dagger = \Pi Q^\dagger \Pi$ and $Q^\dagger Q = QQ^\dagger = \Pi$. The nonzero eigenvalues of Q belong to $]0,1]$, so the eigenvalues of $Q^\dagger - \Pi$ on $\operatorname{range}(Q)$ are of the form $q^{-1} - 1 \ge 0$. Together with $H^f \succeq 0$, this proves that Σ is symmetric positive semidefinite. It also gives $\Pi\Sigma = \Sigma\Pi = \Sigma$. The matrix $T_\varepsilon = \Sigma + \gamma\varepsilon\Pi$ is symmetric and vanishes on $\ker(Q)$. For every $u \in \operatorname{range}(Q)$, $u^\top T_\varepsilon u = u^\top \Sigma u + \gamma\varepsilon\|u\|^2 \ge \gamma\varepsilon\|u\|^2$. It is therefore positive definite on $\operatorname{range}(Q)$, with $\operatorname{range}(T_\varepsilon) = \operatorname{range}(Q)$ and $T_\varepsilon^\dagger T_\varepsilon = T_\varepsilon T_\varepsilon^\dagger = \Pi$.

We next recall that $(I_n - \Pi)Q = 0$ and use the following identities: $(I_n - \Pi)H = I_n - \Pi$ and $Q^\dagger H = T_\varepsilon + \gamma\Pi H^f(I_n - \Pi)$. If $HX = 0$, the first identity gives $(I_n - \Pi)X = 0$, while the second gives $T_\varepsilon \Pi X = 0$. The positive definiteness of T_ε on $\operatorname{range}(Q)$ implies $\Pi X = 0$, and hence $X = 0$. Thus H is invertible. For $X = H^{-1}M$, the same identities yield $(I_n - \Pi)X = (I_n - \Pi)M$ and $T_\varepsilon \Pi X = Q^\dagger M - \gamma\Pi H^f(I_n - \Pi)M = V$. Multiplying the second equality by $T_\varepsilon^\dagger$ proves the claimed decomposition of $H^{-1}M$.

Finally, on an eigenvector of $\Sigma|_{\operatorname{range}(Q)}$ with eigenvalue $\lambda \ge 0$, the matrix $T_\varepsilon^\dagger \Sigma$ acts by $\frac{\lambda}{\lambda + \gamma\varepsilon} \in [0,1[$. It vanishes on $\ker(Q)$, so $\|T_\varepsilon^\dagger \Sigma\| \le 1$. If $\operatorname{range}(V) \subseteq \operatorname{range}(\Sigma)$, then $V = \Sigma\Sigma^\dagger V$ since $\Sigma\Sigma^\dagger$ is a projector. Therefore $\|T_\varepsilon^\dagger V\| = \|T_\varepsilon^\dagger \Sigma\Sigma^\dagger V\| \le \|\Sigma^\dagger V\|$. The decomposition

of $H^{-1}M$ and the triangle inequality give the second bound in (12). If $V = \Sigma W$, then $\Sigma^\dagger V = \Sigma^\dagger \Sigma W$ is the orthogonal projection of W onto range(Σ), so $\|\Sigma^\dagger V\| \leq \|W\|$. ■

E.4 Proof of Composite Path-Differentiability Theorem 23

Proof (Theorem 23) Set $h(x,\theta) := f(x,\theta) + g(x,\theta)$. Under **(Reg')**, each $h(\cdot,\theta)$ is proper, closed, convex, and has a nonempty set of minimizers. Moreover, h is definable. The graph of its Tikhonov minimizers is definable because it can be described by the absence of a point with strictly smaller objective value. For $\varepsilon > 0$, strong convexity of $h + \varepsilon\|\cdot\|^2/2$ makes this minimizer unique. The graph of the minimal-norm selection is likewise definable by minimizing $\|x\|$ over $\operatorname{argmin} h(\cdot,\theta)$. Consequently, the joint selection $x^\star(\theta,\varepsilon) = x^\star_\varepsilon(\theta)$ if $\varepsilon > 0$ and $x^\star_{\min}(\theta)$ if $\varepsilon = 0$ is definable on $O \times [0,\bar\varepsilon]$.

For every fixed $0 < \varepsilon \leq \bar\varepsilon$, Proposition 18 supplies the definable conservative Jacobian (9). Thus the assembled mapping $(\theta,\varepsilon) \mapsto J_{x^\star_\varepsilon}(\theta)$ is definable on $O \times]0,\bar\varepsilon]$.

It remains to establish a bound uniform in ε. Fix a compact set $K \subset O$, let $\theta \in K$, $\varepsilon \in]0,\bar\varepsilon]$, and take $D \in \mathcal{D}_\varepsilon(x^\star_\varepsilon(\theta),\theta)$. By **(R')**, the columns of $V(D)$ belong to range$(\Sigma(D))$ and $\|\Sigma(D)^\dagger V(D)\| \leq C_K$. The estimate (12) therefore gives $\|H_\varepsilon(D)^{-1}M_\varepsilon(D)\| \leq \|(I_n - \Pi_Q)M_\varepsilon(D)\| + \|\Sigma(D)^\dagger V(D)\|$. Using **(M')**, we therefore obtain $\sup_{\substack{\theta\in K,\ 0<\varepsilon\leq\bar\varepsilon \\ J\in J_{x^\star_\varepsilon}(\theta)}} \|J\| \leq B_K + C_K$.

For each fixed θ, Theorem 36 applied to $h(\cdot,\theta) \in \Gamma_0(\mathbb{R}^n)$ gives $x^\star_\varepsilon(\theta) \to x^\star_{\min}(\theta)$ as $\varepsilon \downarrow 0$. Once the uniform bound above and this pointwise convergence are available, the part of the proof of Theorem 5 following the bound on $J_{x^\star_\varepsilon}$ uses only these two facts. It therefore applies here with the bound $B_K + C_K$, showing that $x^\star_{\min}$ is locally Lipschitz and that the joint selection $x^\star$ is continuous at $O \times \{0\}$. Compactness then gives $x^\star_\varepsilon \to x^\star_{\min}$ uniformly on compact subsets of O.

We have now verified the four hypotheses of Theorem 3. Therefore $J_{x^\star_{\min}}$ is a definable conservative Jacobian of $x^\star_{\min}$, which is consequently path-differentiable on O. ■

E.5 Rank-changing Composite Cluster Points Proposition 24

Proof (Proposition 24) By Theorem 36, $x^\star_{\varepsilon_k}(\theta) \to x^\star_{\min}(\theta)$. The definition of z_{ε_k} and the closed graph property of the two Clarke Jacobians in $\mathcal{D}_{\varepsilon_k}$ therefore give $D_0 \in \mathcal{D}_0(x^\star_{\min}(\theta),\theta)$ and $\mathcal{M}_k \to \mathcal{M}_0$. For the selected blocks, **(R')** with $K = \{\theta\}$ gives a constant C_θ such that $V_k = \Sigma_k W_k$, $W_k := \Sigma_k^\dagger V_k$, and $\|W_k\| \leq C_\theta$. The matrices Σ_k and Π_k commute, and both vanish on $\ker(Q_k)$. On range(Q_k), the eigenvalues of $R_k := T_k^\dagger \Sigma_k$ are $\lambda/(\lambda + \gamma\varepsilon_k)$, where λ ranges over the eigenvalues of Σ_k on this subspace. Hence $0 \preceq R_k \preceq \Pi_k$, $T_k^\dagger V_k = R_k W_k$, and $\|T_k^\dagger V_k\| \leq \|W_k\|$. Together with **(M')** and (13), this proves that $(J_k)_k$ is bounded. Every cluster point of $(J_k)_k$ belongs to $J_{x^\star_{\min}}(\theta)$ by (2).

Fix a subsequence along which $J_k \to J$. The sequences $(\Pi_k)_k$ and $(T_k^\dagger V_k)_k$ are bounded, so a further subsequence satisfies $\Pi_k \to \bar\Pi$ and $T_k^\dagger V_k \to \bar Y$. A limit of orthogonal projectors is an orthogonal projector. Since $\Pi_k Q_k = Q_k \Pi_k = Q_k$ and $\Pi_k T_k^\dagger V_k = T_k^\dagger V_k$, passing to the limit yields $\bar\Pi Q_0 = Q_0 \bar\Pi = Q_0$ and $\bar\Pi \bar Y = \bar Y$. The first two identities imply range$(Q_0) \subseteq$ range$(\bar\Pi)$. Passing to the limit in (13) gives $J = -(I_n - \bar\Pi)\mathcal{M}_0 - \bar Y$, which proves the claim. ■

E.6 Stable-Q Residual Formula Corollary 25

Proof (Corollary 25) Fix a subsequence along which $J_k \to J$. Since $\operatorname{rank}(Q_k) = \operatorname{rank}(Q_0)$ eventually, continuity of the pseudoinverse on matrices of fixed rank gives $Q_k^\dagger \to Q_0^\dagger$, $\Pi_k \to \Pi_0$, $\Sigma_k \to \Sigma_0$, and $V_k \to V_0$. Assumption **(R')** makes $W_k := \Sigma_k^\dagger V_k$ bounded.

Set $R_k := T_k^\dagger \Sigma_k$. Its eigenvalues on range(Q_k) are $\lambda/(\lambda+\gamma\varepsilon_k)$, so $0 \preceq R_k \preceq \Pi_k$. After taking a further subsequence, there are matrices W and R such that $W_k \to W$, and $R_k \to R$. The identities $V_k = \Sigma_k W_k$ and $T_k^\dagger V_k = R_k W_k$, together with (13), imply $V_0 = \Sigma_0 W$ and $J = -(I_n - \Pi_0)\mathcal{M}_0 - RW$. Moreover, $0 \preceq R \preceq \Pi_0$. Since R_k commutes with Σ_k and $\|\Sigma_k - \Sigma_k R_k\| = \max_{\lambda \in \mathrm{spec}(\Sigma_k)} \frac{\gamma\varepsilon_k \lambda}{\lambda+\gamma\varepsilon_k} \le \gamma\varepsilon_k$, we obtain $\Sigma_0 R = R\Sigma_0 = \Sigma_0$. Multiplying these identities by $\Sigma_0^\dagger$ shows that $\Sigma_0^\dagger\Sigma_0 R = R\Sigma_0^\dagger\Sigma_0 = \Sigma_0^\dagger\Sigma_0$. Therefore $S := R - \Sigma_0^\dagger\Sigma_0$ satisfies $0 \preceq S \preceq \Pi_0 - \Sigma_0^\dagger\Sigma_0$ and $\Sigma_0 S = S\Sigma_0 = 0$. Finally, $RW = \Sigma_0^\dagger\Sigma_0 W + SW = \Sigma_0^\dagger V_0 + SW$, which proves that $J = -(I_n - \Pi_0)\mathcal{M}_0 - (\Sigma_0^\dagger V_0 + SW)$ as claimed. ■

E.7 Pseudoinverse Cluster Formula Corollary 26

Proof (Corollary 26) Fix a cluster point J and the subsequence and limits supplied by Corollary 25. We have $S = R - \Sigma_0^\dagger\Sigma_0$, so it is enough to prove that $R_k \to \Sigma_0^\dagger\Sigma_0$.

Under (G_Σ), let $P_{\Sigma,k} := \Sigma_k^\dagger\Sigma_k$. The spectral description in the proof of Corollary 25 gives $\|R_k - P_{\Sigma,k}\| \le \frac{\gamma\varepsilon_k}{\lambda^*+\gamma\varepsilon_k} \to 0$. The spectral gap and $\Sigma_k \to \Sigma_0$ also give $P_{\Sigma,k} \to \Sigma_0^\dagger\Sigma_0$, hence $R_k \to \Sigma_0^\dagger\Sigma_0$.

Under (V_Σ), let $r := \mathrm{rank}(\Sigma_0)$. The spectral projector associated with the r largest eigenvalues of Σ_k converges to $\Sigma_0^\dagger\Sigma_0$, and the corresponding multipliers $\lambda_i(\Sigma_k)/(\lambda_i(\Sigma_k)+\gamma\varepsilon_k)$ converge to one. Every remaining eigenvalue tends to zero, while (V_Σ) makes its multiplier converge to zero. Consequently $R_k \to \Sigma_0^\dagger\Sigma_0$ in this case as well. Thus $S = 0$ under either condition, which proves $J = -(I_n - \Pi_0)\mathcal{M}_0 - (\Sigma_0^\dagger V_0)$. ■

E.8 Reachable Binary Limits Corollary 28

Proof (Corollary 28) Fix $\theta \in \mathbb{R}^m$. The map $\varepsilon \mapsto x_\varepsilon^\star(\theta)$ is definable because its graph is the graph of the unique minimizer of the semialgebraic, strongly convex elastic-net objective. Hence every $\rho_i(\cdot;\theta)$ in the corollary is definable. The sets on which $\rho_i(\cdot;\theta)$ is positive, negative, or zero are therefore definable subsets of $\mathbb{R}$. Since each has finitely many connected components, exactly one of them contains all sufficiently small positive ε.

By the soft-threshold formula in Section 4.1, the ith diagonal entry of a matrix in $\mathcal{Q}_\varepsilon(\theta)$ belongs to either $\{1\}$, $\{0\}$, or $[0,1]$. It follows that every binary diagonal matrix satisfying the coordinatewise condition of the corollary belongs to $\mathcal{Q}_\varepsilon(\theta)$ for all sufficiently small ε. Moreover, Theorem 36 gives $x_\varepsilon^\star(\theta) \to x_{\min}^\star(\theta)$, and therefore $z_\varepsilon(x_\varepsilon^\star(\theta),\theta) \to z_0(x_{\min}^\star(\theta),\theta)$. The closed graph of the Clarke Jacobian then places such a binary matrix in $\mathcal{Q}_0(\theta)$, hence in $\mathcal{R}_{\mathrm{bin}}^{\mathrm{LASSO}}(\theta)$. Conversely, if $Q_k \in \mathcal{Q}_{\varepsilon_k}(\theta)$ converges to a binary diagonal matrix Q, the eventual sign of each $\rho_i(\cdot;\theta)$ forces $(Q)_{ii}$ to have the value stated in the corollary. This proves the characterization of $\mathcal{R}_{\mathrm{bin}}^{\mathrm{LASSO}}(\theta)$ and, since each of its coordinate sets is nonempty, the nonemptiness.

Fix $Q \in \mathcal{R}_{\mathrm{bin}}^{\mathrm{LASSO}}(\theta)$. The above shows that $Q \in \mathcal{Q}_\varepsilon(\theta)$ for ε sufficiently small. Take a sequence $\varepsilon_k \downarrow 0$ in this range. By Corollary 19, $J_k := \big(I_n - Q(I_n - \gamma(A^\mathsf{T}A + \varepsilon_k I_n))\big)^{-1}\gamma Q A^\mathsf{T} \in J_{x^\star}(\theta,\varepsilon_k)$. If $Q = 0$, then $J_k = 0 = (AQ)^\dagger$ for every k. Otherwise $S_+ = \{i : (Q)_{ii} = 1\}$ and $Q_{S_+,S_+} = I_{|S_+|}$, so Proposition 20 gives $[J_k]_{S_+,:} = (\varepsilon_k I_{|S_+|} + A_{:,S_+}^\top A_{:,S_+})^{-1} A_{:,S_+}^\top$ and $[J_k]_{S_+^c,:} = 0$. An SVD of $A_{:,S_+}$ yields $[J_k]_{S_+,:} \to A_{:,S_+}^\dagger$, and the limit is $(AQ)^\dagger$. Finally, $x^\star(\theta,\varepsilon_k) \to x_{\min}^\star(\theta)$ by Theorem 36, so $(AQ)^\dagger \in J_{x_{\min}^\star}(\theta)$, as claimed. ■

References

Samir Adly and R. Tyrrell Rockafellar. Sensitivity analysis of maximally monotone inclusions via the proto-differentiability of the resolvent operator. *Mathematical Programming*, 189: 37–54, 05 2021. doi: 10.1007/s10107-020-01515-z.

Michael Arbel and Julien Mairal. Non-convex bilevel games with critical point selection maps. In *NeurIPS 2022 - Proceedings of the 36th Conference on Neural Information Processing Systems*, pages 1–34, 11 2022. doi: 10.48550/arXiv.2207.04888.

Hedy Attouch. Viscosity solutions of minimization problems. *Siam Journal on Optimization - SIAM J OPTIMIZATION*, 6, 08 1996. doi: 10.1137/S1052623493259616.

Jean-François Aujol, Gilles Aubert, Laure Blanc-Féraud, and Antonin Chambolle. Image decomposition into a bounded variation component and an oscillating component. *Journal of Mathematical Imaging and Vision*, 22(1):71–88, 2005. doi: 10.1007/s10851-005-4783-8.

Antoine Bambade, Fabian Schramm, Adrien Taylor, and Justin Carpentier. Leveraging augmented-lagrangian techniques for differentiating over infeasible quadratic programs in machine learning. In *Proceedings of the 12th International Conference on Learning Representations*, 2024.

Heinz Bauschke and Patrick Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, 01 2017. ISBN 978-3-319-48310-8. doi: 10.1007/978-3-319-48311-5.

Aharon Ben-Tal and Marc Teboulle. A geometric property of the least squares solution of linear equations. *Linear algebra and its applications*, 139:165–170, 1990.

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, 02 2020. doi: 10.48550/arXiv.2002.08943.

Quentin Bertrand, Quentin Klopfenstein, Mathurin Massias, Mathieu Blondel, Samuel Vaiter, Alexandre Gramfort, and Joseph Salmon. Implicit differentiation for fast hyperparameter selection in non-smooth convex learning. *Journal of Machine Learning Research*, 2022. doi: 10.48550/arXiv.2105.01637.

Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-Lopez, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. In *Proceedings of the 36th Conference on Neural Information Processing Systems*, 05 2021. doi: 10.48550/arXiv.2105.15183.

Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188:19–51, 04 2021. doi: 10.1007/s10107-020-01501-5.

Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18:556–572, 05 2007. doi: 10.1137/060670080.

Jérôme Bolte, Tam Le, Edouard Pauwels, and Antonio Silveti-Falls. Nonsmooth implicit differentiation for machine learning and optimization. In *Advances in Neural Information Processing Systems*, volume 34, pages 13537–13549, 2021. doi: 10.48550/arXiv.2106.04350.

Jérôme Bolte, Edouard Pauwels, and Samuel Vaiter. Automatic differentiation of nonsmooth iterative algorithms. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 05 2022. doi: 10.48550/arXiv.2206.00457.

Jérôme Bolte, Edouard Pauwels, and Antonio Silveti-Falls. Differentiating nonsmooth solutions to parametric monotone inclusion problems. *SIAM Journal on Optimization*, 34: 71–97, 01 2024. doi: 10.1137/22M1541630.

Jérôme Bolte, Quốc-Tùng Lê, Edouard Pauwels, and Samuel Vaiter. Geometric and computational hardness of bilevel programming. *Mathematical Programming*, 215(1):539–574, 2026a.

Jérôme Bolte, Quốc-Tùng Lê, Edouard Pauwels, and Samuel Vaiter. Bilevel gradient methods and the morse parametric qualification condition. *Mathematics of Operations Research*, 04 2026b. doi: 10.1287/moor.2025.0914.

J. Frédéric Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Verlag, New York, 01 2000. ISBN 978-1-4612-7129-1. doi: 10.1007/978-1-4612-1394-9.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. Jax: composable transformations of python+numpy programs, 2018. URL `http://github.com/jax-ml/jax`.

Felix E. Browder. Existence and approximation of solutions of nonlinear variational inequalities. *Proceedings of the National Academy of Sciences*, 56(4):1080–1086, 1966. doi: 10.1073/pnas.56.4.1080.

Felix E. Browder. Convergence of approximants to fixed points of nonexpansive nonlinear mappings in banach spaces. *Archive for Rational Mechanics and Analysis*, 24:82–90, 1967.

Francis H. Clarke. *Optimization and Nonsmooth Analysis*, volume 1 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, USA, second edition, 1990. ISBN 0-89871-256-4.

Michel Coste. An introduction to o-minimal geometry, 1999.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20, 04 2020. doi: 10.1007/s10208-018-09409-5.

Jean Dieudonné. *Foundations of Modern Analysis, enlarged and corrected printing*. Academic Press, 1969.

Asen L Dontchev and R. Tyrrell Rockafellar. *Implicit Functions and Solution Mappings: A View from Variational Analysis*. Springer Monographs in Mathematics. Springer, 01 2009. ISBN 978-0-387-87820-1. doi: 10.1007/978-0-387-87821-8.

Logan Engstrom, Andrew Ilyas, Benjamin Chen, Axel Feldmann, William Moses, and Aleksander Madry. Optimizing ML training with metagradient descent. *arXiv preprint arXiv:2503.13751*, 2025.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression – Models, Methods and Applications*. Springer, 2013.

Anders Forsgren. On linear least-squares problems with diagonally dominant weight matrices. *SIAM Journal on Matrix Analysis and Applications*, 17(4):763–788, 1996.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, 07 2018.

Riccardo Grazzi, Massimiliano Pontil, and Saverio Salzo. Nonsmooth implicit differentiation: deterministic and stochastic convergence rates. In *41st International Conference on Machine Learning*, pages 16250 – 16274, 07 2024. doi: 10.48550/arXiv.2403.11687.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. doi: 10.1214/aoms/1177703732.

Alexander Ioffe. An invitation to tame optimization. *SIAM Journal on Optimization*, 19: 1894–1917, 02 2009. doi: 10.1137/080722059.

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, 11 2020. doi: 10.48550/arXiv.1911.02590.

Julien Mairal, Francis Bach, and Jean Ponce. Task-driven dictionary learning. *IEEE transactions on pattern analysis and machine intelligence*, 34:791–804, 07 2012. doi: 10.1109/TPAMI.2011.156.

Saeed Masiha, Zebang Shen, Negar Kiyavash, and Niao He. Select-then-differentiate: Solving bilevel optimization with manifold lower-level solution sets. *arXiv preprint arXiv:2605.09209*, 05 2026.

Rodrigo Maulen-Soto, Jalal Fadili, and Hedy Attouch. Tikhonov regularization for stochastic non-smooth convex optimization in hilbert spaces. *Open Journal of Mathematical Optimization*, 6:1–30, 2025. doi: 10.21203/rs.3.rs-4076301/v1.

Sheheryar Mehmood and Peter Ochs. Automatic differentiation of optimization algorithms with time-varying updates. *arXiv preprint arXiv:2410.15923*, 2024.

Yves Meyer. *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations*, volume 22 of *University Lecture Series*. American Mathematical Society, Providence, Rhode Island, 2001. doi: 10.1090/ulect/022.

E. Hastings Moore. On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society*, 26:394–395, 1919.

Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

NYC Open Data. 311 service requests from 2020 to present. NYC Open Data, data set `erm2-nwe9`, 2026. Accessed August 24, 2026.

Dianne P. O'Leary. On bounds for scaled projections and pseudoinverses. *Linear Algebra and its Applications*, 132:115–117, 1990.

Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20(3):389–403, 07 2000. ISSN 0272-4979. doi: 10.1093/imanum/20.3.389.

Panagiotis Patrinos, Lorenzo Stella, and Alberto Bemporad. Forward-backward truncated newton methods for large-scale convex composite optimization. *Optimization Online*, 02 2014.

Edouard Pauwels and Samuel Vaiter. The derivatives of Sinkhorn–Knopp converge. *SIAM Journal on Optimization*, 33(3):1494–1517, 2023.

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 737–746, 02 2016. doi: 10.48550/arXiv.1602.02355.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Roger Penrose. A generalized inverse for matrices. *Proceedings Cambridge Philos. Soc.*, 51: 406–413, 1955.

Gabriel Peyré and Jalal Fadili. Learning analysis sparsity priors. In *International Conference on Sampling Theory and Applications (SampTA)*, Singapour, 2011.

Hans Rademacher. Über partielle und totale differenzierbarkeit von funktionen mehrerer variablen und über die transformation der doppelintegrale. *Mathematische Annalen*, 79, January 1919. doi: 10.1007/BF01498415.

Stephen Robinson. An implicit-function theorem for a class of nonsmooth functions. *Mathematics of Operations Research - MOR*, 16:292–309, 05 1991. doi: 10.1287/moor.16.2.292.

Ferdinando S. Samaria and Andy C. Harter. Parameterisation of a stochastic model for human face identification. In *Proceedings of the 2nd IEEE Workshop on Applications of Computer Vision*, pages 138–142, Sarasota, Florida, 1994.

Sholom Schechtman. The gradient's limit of a definable family of functions admits a variational stratification. *SIAM Journal on Optimization*, 36(2):1075–1099, 2026.

Gilbert W Stewart. On scaled projections and pseudoinverses. *Linear Algebra and its Applications*, 112:189–193, 1989.

Ensio Suonperä and Tuomo Valkonen. Single-loop approaches to nonsmooth bilevel optimisation. *arXiv preprint arXiv:2606.19143*, 2026.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 58:267–288, 1996.

Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011. doi: https://doi.org/10.1111/j.1467-9868.2011.00771.x.

Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7, 01 2013. doi: 10.1214/13-EJS815.

Andrei N. Tikhonov. Solution of ill-posed problems and the regularization method. *Soviet Mathematics Doklady*, 4:1035–1038, 1963.

Samuel Vaiter, Mohammad Golbabaee, Jalal Fadili, and Gabriel Peyré. Model selection with low complexity priors. *Information and Inference: A Journal of the IMA*, 4(3):230–287, 2015.

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.

Lou van den Dries and Chris Miller. Geometric categories and o-minimal structures. *Duke Mathematical Journal*, 84, 08 1996. doi: 10.1215/S0012-7094-96-08416-1.

Stephen A Vavasis and Yinyu Ye. A primal-dual interior point method whose running time depends only on the constraint matrix. *Mathematical Programming*, 74(1):79–120, 1996.

Paul Vicol, Jonathan Lorraine, Fabian Pedregosa, David Duvenaud, and Roger Grosse. On implicit bias in overparameterized bilevel optimization. In *39th International Conference on Machine Learning PMLR*, pages 22234–22259, 12 2022. doi: 10.48550/arXiv.2212.14032.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17 (3):261–272, 2020.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei Efros. Dataset distillation, 09 2018.

Ezra Winston and J. Zico Kolter. Monotone operator equilibrium networks. In *Advances in Neural Information Processing Systems*, volume 33, 06 2020. doi: 10.48550/arXiv.2006.08591.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:301 – 320, 04 2005. doi: 10.1111/j.1467-9868.2005.00503.x.