Beyond Conservative Jacobians: Coderivative Calculus for the Lipschitz Stability of Implicit Networks

Guangyuan Wang*
McGill University
Montreal, Quebec, Canada

November 7, 2025

Abstract

We investigate the Lipschitz stability of solution mappings for nonsmooth parametric optimization problems, which is fundamental to understanding the robustness of solutions in machine learning. We first analyze stability using implicit differentiation based on conservative Jacobians, applying this framework to the LASSO. However, we demonstrate through a rigorous analytical counterexample that this approach can lead to intractable and loose bounds, revealing its limitations. This motivates the adoption of the more powerful framework of variational analysis and coderivative calculus. We show that the coderivative provides a precise characterization of the solution map's sensitivity. We re-derive a tight and correct Lipschitz bound for the LASSO, overcoming the issues of the conservative Jacobian approach.

The primary contribution of this work is the extension of this coderivative analysis to semialgebraic Deep Equilibrium Networks (DEQs) with nonsmooth activations. We establish verifiable conditions for the robust Lipschitzian stability of the DEQ's solution map with respect to its weights and biases. We prove that the Lipschitz modulus is finite and, by leveraging the semialgebraic structure, theoretically computable, providing a rigorous foundation for analyzing the stability and robustness of implicit auto-differentiation in neural networks and their training.

Keywords. Convex and variational analysis, coderivatives and graphical derivatives, nonsmooth semialgebraic geometry, deep equilibrium neural networks, machine learning theory, computational complexity.

1 Introduction

Deep Neural Networks (DNNs) have achieved state-of-the-art performance across a multitude of domains, from computer vision to natural language processing. However, their widespread adoption in safety-critical systems is hindered by a well-documented lack of robustness. Specifically, DNNs are notoriously vulnerable to adversarial perturbations: small, often imperceptible, changes to their input that can cause drastic and confident misclassifications.

This vulnerability is formally linked to the network's sensitivity to input changes, a property rigorously quantified by its Lipschitz constant. A network with a smaller Lipschitz constant is inherently less sensitive

*Corresponding author. Email: guangyuan.wang@mail.mcgill.ca.

to small perturbations, forming the basis for most modern certified defenses against adversarial attacks (Hein and Andriushchenko, 2017; Weng et al., 2018). The global Lipschitz constant, $\mathrm{Lip}(f)$, of a network f provides a deterministic guarantee: for any input x, the network's output cannot change more than $\mathrm{Lip}(f) \cdot \|\delta\|$ when x is perturbed by δ. Consequently, a robust classifier can be built by ensuring that this guaranteed output margin is larger than the gap between the correct class and any incorrect class (Raghunathan et al., 2018; Tsuzuku et al., 2018).

However, computing the exact Lipschitz constant for a multi-layer DNN is NP-hard (Virmaux and Scaman, 2018). This computational intractability has spurred a rich field of research focused on deriving efficiently computable *upper bounds* on the true Lipschitz constant. These methods range from simple products of layer-wise spectral norms (Miyato et al., 2018) to more sophisticated semidefinite programming (SDP) relaxations (Fazlyab et al., 2019) and layer-wise propagation techniques (Gouk et al., 2021).

A fundamental challenge complicating both the analysis and optimization of DNNs stems from their intrinsically *nonsmooth* nature. This nonsmoothness is not an academic edge case; it is a core design feature, introduced by the ubiquitous use of piecewise-linear activation functions such as the Rectified Linear Unit (ReLU) and its variants (e.g., Leaky ReLU), as well as max-pooling operations. While these functions are crucial for efficient training, they render the network's loss landscape non-differentiable, and the network mapping itself is only locally Lipschitz, not smooth. This nonsmooth structure demands a more sophisticated mathematical framework than classical differential calculus. Tools from variational analysis and nonsmooth optimization are essential for a rigorous understanding of these models. For instance, the very concept of a "gradient" for a ReLU network must be carefully defined. Standard backpropagation, while enormously successful in practice, relies on a chain rule that does not apply in the classical sense. Fortunately, recent work has shown that for the broad class of "path-differentiable" or "tame" functions, which includes virtually all modern DNNs, backpropagation computes a valid *conservative gradient* (Bolte and Pauwels, 2021; Davis et al., 2020). This gradient is a selector from the "Clarke subgradient" (Clarke, 1990), providing theoretical justification for applying gradient-based methods to these nonsmooth architectures.

The study of nonsmoothness is therefore critical for *both* optimization and robustness analysis. While many works focus on controlling the Lipschitz constant during training as a form of regularization (Latorre et al., 2020), they often rely on bounds (like spectral norm) that are loose for nonsmooth activations. Tighter bounds require explicitly modeling the nonsmooth structure, leveraging foundational concepts like proximal operators Combettes and Pesquet (2011) and the specific geometry of activation sets Lewis (2002). This direction is even more critical as "implicit layers" and nonsmooth operations become more common in modern ML (Bolte et al., 2023).

In this work, we bridge these domains by providing a rigorous analysis of Lipschitz stability for nonsmooth parametric optimization problems, with a particular focus on applications in machine learning.

Our contributions are as follows:

1. We first analyze the stability of solution mappings using implicit differentiation based on conservative Jacobians (Bolte and Pauwels, 2021), applying this framework to the LASSO problem.
2. We then demonstrate, through a rigorous analytical counterexample, that this conservative Jacobian approach can lead to loose and non-insightful bounds, highlighting its fundamental limitations for this class of problems.
3. Motivated by this, we adopt the more powerful framework of variational analysis and coderivative calculus (Mordukhovich, 2018). We show how the coderivative provides a precise characterization of

the solution map's sensitivity and use it to rederive a tighter Lipschitz bound for the LASSO.

4. Our primary contribution is the extension of this coderivative analysis to semialgebraic Deep Equilibrium Networks (DEQs) with nonsmooth activations.
5. We establish verifiable conditions for the robust Lipschitzian stability of the DEQ's solution map and prove that its Lipschitz modulus is finite and, by leveraging the semialgebraic structure, theoretically computable. This work provides a rigorous foundation for analyzing the stability and robustness of implicit auto-differentiation in modern neural networks.

Acknowledgements. The author would like to thank Professor Tim Hoheisel for useful discussions and comments during the development of this project. This work was partially supported by the Paul M. Pugh Faculty of Science Undergraduate Research Award at McGill University during Summer 2024. Any suggestions and corrections are welcome and appreciated.

2 Problem Set-up and Background

Nonsmooth optimization problems arise in various applications, notably in machine learning, where formulations like the LASSO and nuclear norm minimization are prevalent. The stability of solution mappings under parameter perturbations is critical for understanding the robustness of these solutions. This paper explores the Lipschitz stability of solution mappings in nonsmooth optimization, utilizing the framework of conservative fields introduced by Bolte and Pauwels (2021). Our approach hinges on implicit differentiation to derive explicit quantitative bounds on the Lipschitz constant of solution mappings, with applications to problems such as the LASSO and Square-root LASSO.

We consider a parametric optimization problem of the form

$$\min_{x\in\mathbb{R}^n} h(p,x)+\varphi(x), \tag{1}$$

where $h:\mathbb{R}^p\times\mathbb{R}^n\to\mathbb{R}$ is a C^1 (continuously differentiable) function such that $h(p,\cdot):\mathbb{R}^n\to\mathbb{R}$ is convex for every fixed $p\in\mathbb{R}^p$, and $\varphi:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ is a closed, proper, convex function (i.e., $\varphi\in\Gamma_0(\mathbb{R}^n)$). The solution map $S:\mathbb{R}^p\rightrightarrows\mathbb{R}^n$ is defined as

$$S(p):=\operatorname*{argmin}_{x\in\mathbb{R}^n}\{h(p,x)+\varphi(x)\}. \tag{2}$$

By Fermat's rule for convex optimization, the optimality condition for Eq. (1) is

$$0\in\nabla_x h(p,x)+\partial\varphi(x), \tag{3}$$

where $\partial\varphi(x)$ denotes the Moreau-Rockafellar subdifferential of φ at x. Our goal is to establish the Lipschitz continuity of S and compute explicit bounds on its Lipschitz constant (or modulus), which is defined as

$$\operatorname{Lip}(S):=\sup_{\substack{p_1,p_2\in\mathbb{R}^p\\ p_1\neq p_2}}\ \sup_{\substack{x_1\in S(p_1)\\ x_2\in S(p_2)}}\frac{\|x_1-x_2\|_2}{\|p_1-p_2\|_2}.$$

This bound can be estimated by reformulating the optimality condition Eq. (3) as a fixed-point equation

$F(p,x)=0$ whose solution set in x coincides with $S(p)$. The bound is then given by

$$\mathrm{Lip}(S) \leq \inf_F \inf_{J_F} \sup_{(p,x)\in \mathrm{gph}\, S} \sup_{[A\ B]\in J_F(p,x)} \left\|B^{-1}A\right\|_{\mathrm{op}}$$

where the infimum is taken over all such fixed-point formulations F and all valid conservative Jacobians J_F for F. The matrix $[A\ B]$ is the partition of an element of $J_F(p,x)$ with respect to the parameters p and variables x. Note that the inner infimum over convex-valued J_F is achieved by the Clarke generalized Jacobian $\partial^C F(p,x)$, as it is contained in any convex-valued conservative Jacobian $J_F(p,x)$ (Bolte and Pauwels, 2021, Theorem 4.1). To estimate quantities involving a conservative Jacobian J_F that is not necessarily convex-valued, we can always take its pointwise convex hull, $\tilde{J}_F(x) := \mathrm{conv}(J_F(x))$. This $\tilde{J}_F$ remains a conservative field and is convex-valued. For instance, using the chain rule for conservative Jacobians (Bolte and Pauwels, 2021, Theorem 3.2), the mapping $J_f(g(x))J_g(x)$ is a conservative Jacobian for the composition $f\circ g$, but it is not necessarily convex-valued even if J_f and J_g are. This is not an issue for bounding operator norms, because the operator norm $A \mapsto \|A\|_{\mathrm{op}}$ is a convex function (Horn and Johnson, 2012, Chapter 5). For example, if we have a bound

$$\|A\|_{\mathrm{op}} \leq C \quad \text{for all } A \in J_F(x_0), \text{ with } J_F(x_0) \text{ a nonconvex set,} \tag{4}$$

then it is also true that

$$\|B\|_{\mathrm{op}} \leq C \quad \text{for all } B \in \mathrm{conv}(J_F(x_0)). \tag{5}$$

This holds by the convexity of the norm: for any $A, A' \in J_F(x_0)$ and $\lambda \in [0,1]$,

$$\left\|\lambda A + (1-\lambda)A'\right\|_{\mathrm{op}} \leq \lambda \|A\|_{\mathrm{op}} + (1-\lambda)\left\|A'\right\|_{\mathrm{op}} \leq \lambda C + (1-\lambda)C = C. \tag{6}$$

Now, assume the solution mapping S is locally Lipschitz. By (Clarke, 1990, Proposition 2.6.6), its Lipschitz constant is bounded by the maximal operator norm of the elements in its Clarke subdifferential, $\mathrm{Lip}(S) = \sup_p \sup_{M\in\partial^C S(p)} \|M\|_{\mathrm{op}}$. If we assume that $\sup_{(p,x)\in\mathrm{gph}\, S} \sup_{[A\ B]\in J_F(p,x)} \left\|B^{-1}A\right\|_{\mathrm{op}}$ is bounded, then necessarily B is invertible for each $[A\ B] \in J_F(p,x)$ on the graph of S. By the Implicit Function Theorem for path-differentiable functions (Theorem 4), the solution mapping S is path differentiable with a conservative field $J_S(p)$ given by

$$J_S(p) \rightrightarrows \left\{-B^{-1}A : [A\ B] \in J_F(p, S(p))\right\}.$$

The convex hull of this conservative field must contain the Clarke subdifferential of the solution mapping, $\partial^C S(p) \subseteq \mathrm{conv}\, J_S(p)$. Therefore, any bound on the operator norms of elements in $J_S(p)$ also bounds the Lipschitz constant of S.

Our strategy for bounding $\mathrm{Lip}(S)$ then amounts to the following procedure: First, we construct a fixed-point equation $F(p,x)=0$ that models the optimality condition Eq. (3). We then compute a conservative Jacobian J_F for F. Finally, we derive an upper bound for $\sup \|B^{-1}A\|_{\mathrm{op}}$ over the relevant domain.

The conservative Jacobian approach offers a general framework for nonsmooth optimization, using path integration and conservative fields to derive a Lipschitz bound. However, as we will show, this bound is often loose, revealing factors like e^{λ} and $\lambda_{\min}(X_{\mathcal{E}}^{\mathsf{T}} X_{\mathcal{E}})^{-1}$ in the LASSO problem. On the other hand, the coderivative approach, rooted in variational analysis (Mordukhovich, 2018), provides a tighter bound (e.g.,

$\sqrt{|\mathcal{E}|}$ for LASSO) by exploiting the problem's specific geometric structure, such as normal cones, but it may require stronger assumptions like strict complementarity. The coderivative method is thus more precise for such problems, while the conservative Jacobian approach is broadly applicable but can be less sharp.

3 Preliminaries

We begin by defining the key concepts from nonsmooth analysis that are essential for our analysis, primarily following the framework of conservative differentiation from Bolte and Pauwels (2021).

Definition 1 (Conservative Jacobian). *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a locally Lipschitz function. A set-valued mapping $J_F : \mathbb{R}^n \rightrightarrows \mathbb{R}^{m\times n}$ is a* conservative Jacobian *(or conservative field) for F if:*

1. *$J_F(x)$ is a nonempty, compact set for all $x \in \mathbb{R}^n$.*
2. *The graph of J_F is closed.*
3. *For any absolutely continuous path $\gamma : [0,1] \to \mathbb{R}^n$, and for every measurable selection $M : [0,1] \to \mathbb{R}^{m\times n}$ such that $M(t) \in J_F(\gamma(t))$ for almost all $t \in [0,1]$, the function $t \mapsto M(t)\dot{\gamma}(t)$ is Lebesgue integrable and satisfies the path integral property:*

$$F(\gamma(1)) - F(\gamma(0)) = \int_0^1 M(t)\dot{\gamma}(t)\mathrm{d}t. \tag{7}$$

A function F admitting such a Jacobian is called path differentiable.

Theorem 2 (Properties of Conservative Jacobians). *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be path differentiable with a conservative Jacobian J_F.*

1. *At any point x where F is differentiable, $J_F(x) = \{\nabla F(x)\}$.*
2. *The Clarke generalized Jacobian $\partial^C F(x)$ is contained within the convex hull of $J_F(x)$:*

$$\partial^C F(x) \subseteq \operatorname{conv} J_F(x). \tag{8}$$

Theorem 2 provides the relevant properties for the vector-valued, path-differentiable functions used in this paper. We now state the implicit differentiation theorems that are central to our method.

Theorem 3 (Implicit differentiation (Bolte et al., 2023)). *Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be path differentiable on an open set $\mathcal{U} \times \mathcal{V} \subset \mathbb{R}^n \times \mathbb{R}^m$ with conservative Jacobian J_F. Let $G : \mathcal{U} \to \mathcal{V}$ be a locally Lipschitz function such that, for each $x \in \mathcal{U}$,*

$$F(x, G(x)) = 0.$$

Furthermore, assume that for each $x \in \mathcal{U}$, for each $[A\ B] \in J_F(x, G(x))$ (partitioned w.r.t. $\mathbb{R}^n$ and $\mathbb{R}^m$), the matrix B is invertible. Then, $G : \mathcal{U} \to \mathcal{V}$ is path differentiable with a conservative Jacobian given, for each $x \in \mathcal{U}$, by

$$J_G : x \rightrightarrows \left\{-B^{-1}A : \begin{bmatrix} A & B \end{bmatrix} \in J_F(x, G(x))\right\}.$$

Theorem 4 (Path differentiable implicit function theorem (Bolte et al., 2023)). *Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be path differentiable with conservative Jacobian J_F. Let $(\hat{x}, \hat{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ be such that $F(\hat{x}, \hat{y}) = 0$. Assume that $J_F(\hat{x}, \hat{y})$ is* *convex and that, for each $[A\ B] \in J_F(\hat{x}, \hat{y})$, the matrix B is invertible. Then, there exist open neighborhoods $\mathcal{U} \subset \mathbb{R}^n$ of $\hat{x}$ and $\mathcal{V} \subset \mathbb{R}^m$ of $\hat{y}$, and a path differentiable function $G : \mathcal{U} \to \mathcal{V}$ such that the conclusion of Theorem 3 holds.*

Theorem 5 (Path differentiable inverse function theorem Bolte et al. (2023))**.** *Let $\mathcal{U}$ and $\mathcal{V}$ be open neighborhoods of 0 in $\mathbb{R}^n$ and $\Phi : \mathcal{U} \to \mathcal{V}$ be path differentiable with $\Phi(0) = 0$. Assume that Φ has a conservative Jacobian J_Φ such that $J_\Phi(0)$ contains only invertible matrices. Then, locally, Φ has a path differentiable inverse Ψ with a conservative Jacobian given by*

$$J_\Psi(y) = \left\{A^{-1} : A \in J_\Phi(\Psi(y))\right\}.$$

4 Lipschitz Stability via Implicit Differentiation

Consider the fixed-point equation associated with the optimality condition Eq. (3). A common choice is the proximal-gradient fixed-point equation. Assuming $h(p,x)$ is L_h-smooth in x, we can define a fixed-point mapping

$$F(p,x) := x - \mathrm{prox}_{\gamma\varphi}(x - \gamma \nabla_x h(p,x)) = 0, \tag{9}$$

where $\gamma \in (0, 1/L_h]$ is a step size and $\mathrm{prox}_{\gamma\varphi}$ is the proximal operator of $\gamma\varphi$. The solution mapping $S(p)$ satisfies $x \in S(p)$ if and only if $F(p,x) = 0$. We aim to bound $\mathrm{Lip}(S)$ using the conservative Jacobian J_F of F.

Theorem 6 (Fixed-point Equation Lipschitz Bound via Conservative Jacobian)**.** *Let $F : \mathbb{R}^p \times \mathbb{R}^n \to \mathbb{R}^n$ be a fixed-point equation modeling the optimality condition Eq. (3), with $S(p) = \{x \in \mathbb{R}^n : F(p,x) = 0\}$. Suppose J_F is a convex-valued conservative field for F, and J_F^C is the Clarke generalized Jacobian. Then,*

$$\mathrm{Lip}(S) \le \inf_F \inf_{J_F} \sup_{[A\ \ B] \in J_F(p,x)} \|B^{-1}A\|_{\mathrm{op}}, \tag{10}$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm, and $[A\ \ B]$ represents the Jacobian matrix partitioned with respect to parameters p and variables x.

Proof. For $p_1, p_2 \in \mathbb{R}^p$, let $x_1 \in S(p_1)$, $x_2 \in S(p_2)$, so $F(p_1, x_1) = F(p_2, x_2) = 0$. Consider a path $\gamma : [0,1] \to \mathbb{R}^p \times \mathbb{R}^n$, $\gamma(t) = (p(t), x(t))$, with $p(0) = p_1$, $p(1) = p_2$, $x(0) = x_1$, $x(1) = x_2$, and $F(p(t), x(t)) = 0$. By the conservativity of J_F, for a measurable selection $[A(t)\, B(t)] \in J_F(p(t), x(t))$, it holds that $\frac{\mathrm{d}}{\mathrm{d}t} F(p(t), x(t)) = A(t)\dot{p}(t) + B(t)\dot{x}(t) = 0$. Thus, $\dot{x}(t) = -B(t)^{-1}A(t)\dot{p}(t)$. The operator norm satisfies

$$\|\dot{x}(t)\|_2 \le \|B(t)^{-1}A(t)\|_{\mathrm{op}} \|\dot{p}(t)\|_2. \tag{11}$$

Integrating over $[0,1]$ gives

$$\|x_2 - x_1\|_2 \le \int_{[0,1]} \|B(t)^{-1}A(t)\|_{\mathrm{op}} \|\dot{p}(t)\|_2 \,\mathrm{d}t. \tag{12}$$

Taking the infimum over all paths and selections, and supremum over all $[A\ B] \in J_F$, yields the bound. Since $J_F^C \subseteq J_F$, the result extends to the Clarke Jacobians. $\square$

5 A Conservative Jacobian Approach for Deriving the LASSO Lipschitz Bound for the Implicit Differentiation of the Solution Mapping

We now apply this framework to the LASSO problem, formulated with parameter $\lambda \in \mathbb{R}$ with $p = 1$ and variable $x \in \mathbb{R}^n$:

$$\min_{x \in \mathbb{R}^n} \left\{ \tfrac{1}{2}\|Ax - b\|^2 + \lambda \|x\|_1 \right\} \tag{13}$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. The optimality condition $0 \in A^\top(Ax - b) + \lambda\partial\|x\|_1$ can be characterized by the fixed-point equation for an optimal solution $\bar{x}$ (using the proximal gradient step with $\gamma = 1$, which assumes $\|A^\top A\|_{\mathrm{op}} \leq 1$ for convergence, but here serves to define the solution):

$$F(\lambda, \bar{x}) = \bar{x} - \operatorname{prox}_{\lambda\|\cdot\|_1}\left(\bar{x} - A^\top(A\bar{x} - b)\right) = 0. \tag{14}$$

To analyze this, we define two intermediate functions

$$\begin{aligned} T(\lambda, x) &:= \left(\lambda, x - A^\top(Ax - b)\right) \\ S(\lambda, x) &:= \operatorname{prox}_{\lambda\|\cdot\|_1}(x), \end{aligned}$$

where $T : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R} \times \mathbb{R}^n$ and $S : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$. Note the notational overload: the second argument of S is a generic vector, not necessarily the x from the LASSO problem. The fixed-point equation can then be written as a composition

$$F(\lambda, x) = x - S(T(\lambda, x)) = x - S(\lambda, x - A^\top(Ax - b)).$$

The conservative Jacobian J_F is partitioned as $J_F = [\partial_\lambda F \ \ \partial_x F]$.

$$J_F(\lambda, x) = \begin{bmatrix} 0 & \mathrm{Id}_n \end{bmatrix} - J_{S\circ T}(\lambda, x).$$

Then, we can compute $J_{S\circ T}$ using the chain rule (Bolte and Pauwels, 2021, Theorem 3.2):

$$J_{S\circ T}(\lambda, x) = J_S(T(\lambda, x)) \cdot J_T(\lambda, x).$$

Towards differentiating the fixed-point equation, one would need to compute the joint conservative Jacobian of F with respect to (λ, x). Since $F = x - (S \circ T)(\lambda, x)$, we can simply compute the joint conservative Jacobian of $S \circ T$ with respect to (λ, x) and our final conservative Jacobian will be

$$J_F(\lambda, x) = \left\{ \begin{bmatrix} A & \mathrm{Id} - B \end{bmatrix} : \begin{bmatrix} A & B \end{bmatrix} \in J_{S\circ T}(\lambda, x) \right\}.$$

Using the chain rule (lemma 5 Bolte/Pauwels), we have that $J_S(T(\lambda, x))J_T(\lambda, x)$ is a conservative Jacobian for $S \circ T$ whenever J_S and J_T are conservative Jacobians for S and T respectively.

Conservative Jacobian of T. Since $T(\lambda, x)$ is differentiable (indeed, it is even linear), we have

$$J_T(\lambda, x) = \begin{bmatrix} 1 & 0 \\ 0 & \mathrm{Id} - A^\star A \end{bmatrix}$$

Conservative Jacobian of S. The conservative Jacobian of S should be a set-valued mapping from $\mathbb{R}\times\mathbb{R}^n \rightrightarrows \mathbb{R}^{n\times(n+1)}$. Note that $S : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^n$ is separable in x so we may write

$$S(\lambda, x) = \begin{bmatrix} S_1(\lambda, x_1) \\ \vdots \\ S_n(\lambda, x_n) \end{bmatrix}$$

with $S_1 = \cdots = S_n$ the soft thresholding operator given by

$$S_i(\lambda, x_i) = \operatorname{prox}_{\lambda|\cdot|}(x_i) = \begin{cases} x + \lambda & x < -\lambda \\ 0 & x \in [-\lambda, \lambda] \\ x - \lambda & x > \lambda. \end{cases}$$

In Lemma 3 of (Bolte et al., 2023), it is shown that we can compute conservative Jacobians row-wise (i.e., according to coordinates of the output) but there is no corresponding result for computing things column-wise. So for the function S we can compute its conservative Jacobian by computing a conservative Jacobian with respect to (λ, x) associated to each function S_i. Then we can aggregate these results row-wise to form J_S. We note in addition that in the Bolte/Pauwels paper they use the transpose of the conservative gradient for each row but this is equivalent to the conservative Jacobian here.

Let us compute the Clarke Jacobian of each S_i for $i \in \{1,\dots,n\}$. Formally, $S_i : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}$, so the Clarke Jacobian is a set-valued mapping $\partial^c_{S_i} : \mathbb{R}\times\mathbb{R}^n \to \mathbb{R}^{1\times(1+n)}$ given by $\partial^c_{S_i}(\lambda, x) = \left(\partial_\lambda 0 \dots \partial_{x_i} \dots 0\right)$. For $\lambda > 0$, for each $i \in \{1,\dots,n\}$, the function S_i is differentiable on three open sets in the (λ, x)-plane:

$$\begin{aligned} R_1^i &= \{(\lambda, x) : x_i < -\lambda\} \\ R_2^i &= \{(\lambda, x) : x_i \in (-\lambda, \lambda)\} \\ R_3^i &= \{(\lambda, x) : x_i > \lambda\}. \end{aligned}$$

On R_1^i, the Jacobian of S_i with respect to (λ, x) is $J_{S_i}(\lambda, x)\big|_{R_1^i} = \begin{bmatrix} 1 & 0 & \cdots & 1 & \cdots & 0 \end{bmatrix}$. On R_2^i, the Jacobian of S_i with respect to (λ, x) is $J_{S_i}(\lambda, x)\big|_{R_2^i} = 0$. On R_3^i, the Jacobian of S_i with respect to (λ, x) is $J_{S_i}(\lambda, x)\big|_{R_3^i} = \begin{bmatrix} -1 & 0 & \dots & 1 & \dots & 0 \end{bmatrix}$. Define two more regions which are the boundaries of the previous regions where the function S_i is not smooth, $R_4^i = \{(\lambda, x) : x_i = -\lambda\}$ and $R_5^i = \{(\lambda, x) : x_i = \lambda\}$. On R_4^i, the Clarke Jacobian of S_i with respect to (λ, x) is

$$\operatorname{conv}\left(\left\{ J_{S_i}(\lambda, x)\big|_{R_1^i}, J_{S_i}(\lambda, x)\big|_{R_2^i} \right\}\right) = \left\{\begin{bmatrix} \gamma & 0 & \cdots & \gamma & \cdots & 0 \end{bmatrix} : \gamma \in [0,1]\right\},$$

and similarly on R_5^i it is given by

$$\operatorname{conv}\left(\left\{ J_{S_i}(\lambda, x)\big|_{R_2^i}, J_{S_i}(\lambda, x)\big|_{R_3^i} \right\}\right) = \left\{\begin{bmatrix} -\gamma & 0 & \cdots & \gamma & \cdots & 0 \end{bmatrix} : \gamma \in [0,1]\right\}.$$

Concatenating these Clarke Jacobians into a matrix we get back the conservative Jacobian associated to S

with respect to (λ, x):

$$
\begin{aligned}
&J_S(\lambda, x)\\
&= \left\{ \begin{bmatrix} q_1^\lambda & q_1^x & 0 & \dots & 0 \\ q_2^\lambda & 0 & q_2^x & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_n^\lambda & 0 & \dots & 0 & q_n^x \end{bmatrix} : q_i^\lambda = \begin{cases} 1 & (\lambda, x) \in R_1^i \\ [0,1] & (\lambda, x) \in R_4^i \\ 0 & (\lambda, x) \in R_2^i \\ [-1,0] & (\lambda, x) \in R_5^i \\ -1 & (\lambda, x) \in R_3^i \end{cases},\ q_i^x = \begin{cases} q_i^\lambda & (x, \lambda) \in R_1^i \cup R_2^i \cup R_4^i \\ -q_i^\lambda & (x, \lambda) \in R_3^i \cup R_5^i \end{cases} \right\}.
\end{aligned}
$$

Our main hurdle now is to compute the conservative Jacobian $J_S(T(\lambda, x))$ for $S \circ T$. The regions $R_1^i, \dots, R_5^i$ depend explicitly on the second argument of S, so we redefine them using $x - A^\star(Ax - b)$, which results in

$$
\begin{aligned}
R_1^i &= \left\{ (\lambda, x) : x_i - A_i^\star(Ax - b) < -\lambda \right\} \\
R_2^i &= \left\{ (\lambda, x) : x_i - A_i^\star(Ax - b) \in (-\lambda, \lambda) \right\} \\
R_3^i &= \left\{ (\lambda, x) : x_i - A_i^\star(Ax - b) > \lambda \right\} \\
R_4^i &= \left\{ (\lambda, x) : x_i - A_i^\star(Ax - b) = -\lambda \right\} \\
R_5^i &= \left\{ (\lambda, x) : x_i - A_i^\star(Ax - b) = \lambda \right\}.
\end{aligned}
$$

Now we can simply reuse the expression we already computed before with the new regions:

$$
J_S(T(\lambda, x)) = \left\{ \begin{bmatrix} q_1^\lambda & q_1^x & 0 & \dots & 0 \\ q_2^\lambda & 0 & q_2^x & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_n^\lambda & 0 & \dots & 0 & q_n^x \end{bmatrix} : q_i^\lambda = \begin{cases} 1 & (\lambda, x) \in R_1^i \\ [0,1] & (\lambda, x) \in R_4^i \\ 0 & (\lambda, x) \in R_2^i \\ [-1,0] & (\lambda, x) \in R_5^i \\ -1 & (\lambda, x) \in R_3^i. \end{cases} \right\}.
$$

Finally,

$$
\begin{aligned}
&J_S(T(\lambda, x)) J_T(\lambda, x)\\
&= \left\{ \begin{bmatrix} q_1^\lambda & q_1^x & 0 & \dots & 0 \\ q_2^\lambda & 0 & q_2^x & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_n^\lambda & 0 & \dots & 0 & q_n^x \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \mathrm{Id} - A^\star A \end{bmatrix} : \begin{array}{l} q_i^\lambda = \begin{cases} 1 & (\lambda, x) \in R_1^i \\ [0,1] & (\lambda, x) \in R_4^i \\ 0 & (\lambda, x) \in R_2^i \\ [-1,0] & (\lambda, x) \in R_5^i \\ -1 & (\lambda, x) \in R_3^i \end{cases} \\ q_i^x = \begin{cases} q_i^\lambda & (x, \lambda) \in R_1^i \cup R_2^i \cup R_4^i \\ -q_i^\lambda & (x, \lambda) \in R_3^i \cup R_5^i \end{cases} \end{array} \right\} \\
&= \left\{ \begin{bmatrix} q_1^\lambda & \\ q_2^\lambda & \\ \vdots & Q\left(\mathrm{Id} - A^\star A\right) \\ q_{n-1}^\lambda & \\ q_n^\lambda & \end{bmatrix} : q_i^\lambda = \begin{cases} 1 & (\lambda, x) \in R_1^i \\ [0,1] & (\lambda, x) \in R_4^i \\ 0 & (\lambda, x) \in R_2^i \\ [-1,0] & (\lambda, x) \in R_5^i \\ -1 & (\lambda, x) \in R_3^i \end{cases} \quad Q_{ii} = \begin{cases} q_i^\lambda & (x, \lambda) \in R_1^i \cup R_2^i \cup R_4^i \\ -q_i^\lambda & (x, \lambda) \in R_3^i \cup R_5^i \end{cases} \right\}
\end{aligned}
$$

where it is clear that Q is a diagonal matrix in the last expression. From this we can see that we recover the original claimed conservative Jacobian from $\partial_x F(\lambda, \bar{x}) = \mathrm{Id} - Q\left(\mathrm{Id} - A^\star A\right)$ with $Q = \mathrm{diag}(q)$ and

$$
q_i = \begin{cases} 0 & \left| \bar{x}_i - \left(A^\star\right)_i (A\bar{x} - b) \right| < \lambda \\ [0,1] & \left| \bar{x}_i - \left(A^\star\right)_i (A\bar{x} - b) \right| = \lambda \\ 1 & \left| \bar{x}_i - \left(A^\star\right)_i (A\bar{x} - b) \right| > \lambda \end{cases}
$$

where $\left(A^\star\right)_i$ is the i th row of $A^\star$. The matrices in $\partial_x F(\lambda,\bar{x})$ are similar to matrices of the form

$$\mathrm{Id} - Q^{1/2}\left(\mathrm{Id} - A^\star A\right)Q^{1/2} = (\mathrm{Id} - Q) + Q^{1/2}A^\star A Q^{1/2}.$$

Examining $Q^{1/2}A^\star A Q^{1/2}$ we can see that

$$\left(Q^{1/2}A^\star A Q^{1/2}\right)_{ij} = \begin{cases} 0 & \text{if } i \in \mathcal{E}^c \text{ or } j \in \mathcal{E}^c \\ \left(q_j q_i\right)^{1/2}\left(A^\star A\right)_{ij} & \text{if } (i,j) \in \mathcal{E}^2 \end{cases}$$

and

$$(\mathrm{Id} - Q)_{ij} = \begin{cases} 1 & i = j \text{ and } i \in \mathcal{E}^c \\ 1 - q_i & i = j \text{ and } i \in \mathcal{E} \\ 0 & i \neq j. \end{cases}$$

So, an entry of $(\mathrm{Id} - Q) + Q^{1/2}A^\star A Q^{1/2}$ has the form

$$\left((\mathrm{Id} - Q) - Q^{1/2}A^\star A Q^{1/2}\right)_{ij} = (1 - q_i)\,\delta_{ij} - \left(q_i q_j\right)^{1/2}\left(A^\star A\right)_{ij} \tag{15}$$

with δ_{ij} the kronecker delta.

$$M = \begin{bmatrix} A^\top A\big|_{\mathcal{S}} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & a \\ a & a \end{bmatrix}$$

This would result in a block submatrix of $A^\top A$ restricted to the support $\mathcal{S}$, a band around this corresponding to the indices in $\mathcal{E}\backslash\mathcal{S}$, and finally an identity block corresponding to $\mathcal{E}^c$. Further simplifying, the evaluation $J_S(T(\lambda,x))$ means we use the definitions of q_i^λ, q_i^x on the regions R_k^i defined by the output of $T(\lambda,x)$, which is (λ,z) where $z = x - A^\top(Ax - b)$. Hence

$$\begin{aligned} J_{S\circ T}(\lambda,x) &= \begin{bmatrix} v_\lambda & Q_x \end{bmatrix}\Big|_{(\lambda,z)} \cdot \begin{bmatrix} 1 & 0 \\ 0 & \mathrm{Id} - A^\top A \end{bmatrix} \\ &= \begin{bmatrix} (v_\lambda \cdot 1 + Q_x \cdot 0) & (v_\lambda \cdot 0 + Q_x(\mathrm{Id} - A^\top A)) \end{bmatrix} \\ &= \begin{bmatrix} v_\lambda & Q_x(\mathrm{Id} - A^\top A) \end{bmatrix} \end{aligned}$$

where v_λ and Q_x are evaluated at $(\lambda, x - A^\top(Ax - b))$. This gives the final conservative Jacobian J_F

$$\begin{aligned} J_F(\lambda,x) &= \begin{bmatrix} 0 & \mathrm{Id}_n \end{bmatrix} - J_{S\circ T}(\lambda,x) \\ &= \begin{bmatrix} 0 & \mathrm{Id}_n \end{bmatrix} - \begin{bmatrix} v_\lambda & Q_x(\mathrm{Id} - A^\top A) \end{bmatrix} \\ &= \left[\begin{array}{c|c} -v_\lambda & \mathrm{Id}_n - Q_x(\mathrm{Id}_n - A^\top A) \end{array}\right]. \end{aligned}$$

Applying Theorem 6, we identify the components $A = -v_\lambda \in \mathbb{R}^{n\times 1}$ and $B = \mathrm{Id}_n - Q_x(\mathrm{Id}_n - A^\top A) = (\mathrm{Id}_n - Q_x) + Q_x A^\top A \in \mathbb{R}^{n\times n}$. The Lipschitz constant is indeed bounded by $\sup\|B^{-1}A\|_{\mathrm{op}}$.

Analysis of the Jacobian at a Solution. Let $(\lambda,\bar{x})$ be a point on the graph of S. Let $\bar{z} = \bar{x} - A^\top(A\bar{x} - b)$. We partition the indices $\{1,\ldots,n\}$ based on the optimality conditions, which relate $\bar{x}$ to $\bar{z}$:

- Support $\mathcal{S} = \{i : \bar{x}_i \neq 0\}$: Optimality requires $A_i^\top(A\bar{x}-b) = -\lambda\,\mathrm{sgn}(\bar{x}_i)$. This implies $\bar{z}_i = \bar{x}_i - (-\lambda\,\mathrm{sgn}(\bar{x}_i)) = \bar{x}_i + \lambda\,\mathrm{sgn}(\bar{x}_i)$. If $\bar{x}_i > 0$, $\bar{z}_i > \lambda$ (R_3^i). Thus $q_i^\lambda = -1$, $q_i^x = 1$. If $\bar{x}_i < 0$, $\bar{z}_i < -\lambda$ (R_1^i). Thus $q_i^\lambda = 1$, $q_i^x = 1$. In both cases, $q_i^x = 1$ and $q_i^\lambda = -\mathrm{sgn}(\bar{x}_i)$.

- Inactive set $\mathcal{I} = \{i : \bar{x}_i = 0, |A_i^\top(A\bar{x}-b)| < \lambda\}$: Here, $\bar{z}_i = 0 - A_i^\top(A\bar{x}-b)$, so $|\bar{z}_i| < \lambda$ (R_2^i). Thus $q_i^\lambda = q_i^x = 0$.

- Active non-support $\mathcal{E} = \{i : \bar{x}_i = 0, |A_i^\top(A\bar{x}-b)| = \lambda\}$: Here, $\bar{z}_i = -A_i^\top(A\bar{x}-b)$, so $|\bar{z}_i| = \lambda$ (R_4^i or R_5^i). Thus $q_i^\lambda \in [-1,1]$ and $q_i^x \in [0,1]$ (with $q_i^x = \pm q_i^\lambda$ depending on the kink).

Let $\mathcal{A} = \mathcal{S} \cup \mathcal{E}$ be the full set of *active* indices where $\bar{x}_i \neq 0$ or the gradient constraint is binding. The matrix $Q_x = \mathrm{diag}(q_i^x)$ has $(Q_x)_{ii} = 1$ if $i \in \mathcal{S}$, $(Q_x)_{ii} \in [0,1]$ if $i \in \mathcal{E}$, and $(Q_x)_{ii} = 0$ if $i \in \mathcal{I}$. The vector $A = -v_\lambda$ has $A_i = \mathrm{sgn}(\bar{x}_i)$ if $i \in \mathcal{S}$, $A_i = 0$ if $i \in \mathcal{I}$, and $A_i \in [-1,1]$ if $i \in \mathcal{E}$.

Now, let's analyze the matrix $B = (\mathrm{Id}_n - Q_x) + Q_x A^\top A$. We permute the rows and columns to group indices by $\mathcal{A}$ (active) and $\mathcal{I}$ (inactive) to get

$$B = \begin{bmatrix} B_{\mathcal{A}\mathcal{A}} & B_{\mathcal{A}\mathcal{I}} \\ B_{\mathcal{I}\mathcal{A}} & B_{\mathcal{I}\mathcal{I}} \end{bmatrix}$$

For $i,j \in \mathcal{I}$, $(Q_x)_{ii} = 0$, $B_{ij} = (\mathrm{Id}-0)_{ij} + (0 \cdot A^\top A)_{ij} = \delta_{ij}$. Thus, $B_{\mathcal{I}\mathcal{I}} = \mathrm{Id}_{\mathcal{I}}$. If $i \in \mathcal{I}, j \in \mathcal{A}$, $(Q_x)_{ii} = 0$. $B_{ij} = (\mathrm{Id}-0)_{ij} + (0 \cdot A^\top A)_{ij} = 0$. Thus, $B_{\mathcal{I}\mathcal{A}} = 0$. The matrix B has a block upper-triangular structure:

$$B = \begin{bmatrix} B_{\mathcal{A}\mathcal{A}} & B_{\mathcal{A}\mathcal{I}} \\ 0 & \mathrm{Id}_{\mathcal{I}} \end{bmatrix}$$

where $B_{\mathcal{A}\mathcal{A}} = (\mathrm{Id}-Q_x)_{\mathcal{A}\mathcal{A}} + (Q_x A^\top A)_{\mathcal{A}\mathcal{A}}$ and $B_{\mathcal{A}\mathcal{I}} = (Q_x A^\top A)_{\mathcal{A}\mathcal{I}}$. This structure is very useful, as its inverse is

$$B^{-1} = \begin{bmatrix} B_{\mathcal{A}\mathcal{A}}^{-1} & -B_{\mathcal{A}\mathcal{A}}^{-1} B_{\mathcal{A}\mathcal{I}} \\ 0 & \mathrm{Id}_{\mathcal{I}} \end{bmatrix}$$

(assuming $B_{\mathcal{A}\mathcal{A}}$ is invertible). This block-triangular structure is the correct and rigorous decomposition of the Jacobian. This block decomposition provides a clear path forward to bound $\|B^{-1}A\|_{\mathrm{op}}$ by focusing on the invertibility of the active-set submatrix $B_{\mathcal{A}\mathcal{A}}$.

Lipschitz Bound via Block-Matrix Decomposition. We now analyze this Jacobian at a solution point $(\lambda, \bar{\beta}) \in \mathrm{gph}\, S$. At a solution, the LASSO optimality conditions provide a crucial link between $\bar{\beta}$, X, y, λ, and the arguments of Q_x and v_λ. Let $\bar{z} = \bar{\beta} - X^\top(X\bar{\beta} - y)$. The KKT conditions imply that for an inactive set $\mathcal{I}$. $i \in \mathcal{I} \iff \bar{\beta}_i = 0$ and $|X_i^\top(X\bar{\beta}-y)| < \lambda$. For $i \in \mathcal{I}$, $X_i^\top(X\bar{\beta}-y) \in (-\lambda, \lambda)$. This implies $\bar{z}_i = \bar{\beta}_i - X_i^\top(\ldots) = -X_i^\top(\ldots) \in (-\lambda,\lambda)$. Thus, for $i \in \mathcal{I}$, we are in the differentiable region where $q_i = 0$ and $v_i = 0$. On an active set $\mathcal{A}$: $i \in \mathcal{A} \iff i \notin \mathcal{I}$. This set contains both the support $\mathrm{supp}(\bar{\beta}) = \{i : \bar{\beta}_i \neq 0\}$ and the active non-support $\mathcal{E} \setminus \mathrm{supp}(\bar{\beta}) = \{i : \bar{\beta}_i = 0, |X_i^\top(X\bar{\beta}-y)| = \lambda\}$. For $i \in \mathcal{A}$, we have $q_i \in [0,1]$ and $v_i \in [-1,1]$.

This partitioning of the indices $\{1,\ldots,p\}$ into $\mathcal{A}$ and $\mathcal{I}$ reveals a powerful block structure. Let P be a permutation matrix that groups indices such that

$$P\beta = \begin{bmatrix} \beta_{\mathcal{A}} \\ \beta_{\mathcal{I}} \end{bmatrix}.$$

Applying this permutation to any $M \in J_F(\lambda, \bar{\beta})$, we get $M_P = PMP^\top$. The components $A = -v_\lambda$ and Q_x (and thus B) have a block structure based on this partition. For $i \in \mathcal{I}$, $(v_\lambda)_i = 0$ and $(Q_x)_{ii} = 0$, and so

$$A_P = PA = \begin{bmatrix} A_{\mathcal{A}} \\ A_{\mathcal{I}} \end{bmatrix} = \begin{bmatrix} A_{\mathcal{A}} \\ 0 \end{bmatrix}, \quad (Q_x)_P = PQ_xP^\top = \begin{bmatrix} (Q_x)_{\mathcal{A}\mathcal{A}} & 0 \\ 0 & (Q_x)_{\mathcal{I}\mathcal{I}} \end{bmatrix} = \begin{bmatrix} (Q_x)_{\mathcal{A}\mathcal{A}} & 0 \\ 0 & 0 \end{bmatrix}. \tag{16}$$

Now let us turn to $B_P = PBP^\top$. Let $G = \mathrm{Id}_p - X^\top X$. Then we necessarily have that

$$\begin{aligned} B_P &= P\left(\mathrm{Id}_p - Q_xG\right)P^\top = \mathrm{Id}_p - (PQ_xP^\top)(PGP^\top) \\ &= \mathrm{Id}_p - \begin{bmatrix} (Q_x)_{\mathcal{A}\mathcal{A}} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} G_{\mathcal{A}\mathcal{A}} & G_{\mathcal{A}\mathcal{I}} \\ G_{\mathcal{I}\mathcal{A}} & G_{\mathcal{I}\mathcal{I}} \end{bmatrix} \\ &= \mathrm{Id}_p - \begin{bmatrix} (Q_x)_{\mathcal{A}\mathcal{A}}G_{\mathcal{A}\mathcal{A}} & (Q_x)_{\mathcal{A}\mathcal{A}}G_{\mathcal{A}\mathcal{I}} \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{Id}_{\mathcal{A}} - (Q_x)_{\mathcal{A}\mathcal{A}}G_{\mathcal{A}\mathcal{A}} & -(Q_x)_{\mathcal{A}\mathcal{A}}G_{\mathcal{A}\mathcal{I}} \\ 0 & \mathrm{Id}_{\mathcal{I}} \end{bmatrix}. \end{aligned} \tag{17}$$

This reveals that B_P is block upper-triangular. Let $B_{\mathcal{A}\mathcal{A}} = \mathrm{Id}_{\mathcal{A}} - (Q_x)_{\mathcal{A}\mathcal{A}}G_{\mathcal{A}\mathcal{A}}$ and $B_{\mathcal{A}\mathcal{I}} = -(Q_x)_{\mathcal{A}\mathcal{A}}G_{\mathcal{A}\mathcal{I}}$.

$$B_P = \begin{bmatrix} B_{\mathcal{A}\mathcal{A}} & B_{\mathcal{A}\mathcal{I}} \\ 0 & \mathrm{Id}_{\mathcal{I}} \end{bmatrix}. \tag{18}$$

By Theorem 3, the conservative Jacobian for the solution map $S(\lambda)$ is $J_S(\lambda) \rightrightarrows \{-B^{-1}A\}$. So $C_P = (B_P)^{-1}A_P$. Using the standard formula for the inverse of a block upper-triangular matrix,

$$(B_P)^{-1} = \begin{bmatrix} (B_{\mathcal{A}\mathcal{A}})^{-1} & -(B_{\mathcal{A}\mathcal{A}})^{-1}B_{\mathcal{A}\mathcal{I}} \\ 0 & \mathrm{Id}_{\mathcal{I}} \end{bmatrix}, \tag{19}$$

assuming $B_{\mathcal{A}\mathcal{A}}$ is invertible. Now we compute C_P:

$$C_P = \begin{bmatrix} (B_{\mathcal{A}\mathcal{A}})^{-1} & -(B_{\mathcal{A}\mathcal{A}})^{-1}B_{\mathcal{A}\mathcal{I}} \\ 0 & \mathrm{Id}_{\mathcal{I}} \end{bmatrix} \begin{bmatrix} A_{\mathcal{A}} \\ 0 \end{bmatrix} = \begin{bmatrix} (B_{\mathcal{A}\mathcal{A}})^{-1}A_{\mathcal{A}} \\ 0 \end{bmatrix}. \tag{20}$$

Since the operator norm is invariant under permutation, $\|C\|_{\mathrm{op}} = \|C_P\|_{\mathrm{op}}$. The operator norm of this block matrix is

$$\|C_P\|_{\mathrm{op}} = \sup_{\|v\|_2=1} \|C_Pv\|_2 = \sup_{\|v_{\mathcal{A}}\|_2^2+\|v_{\mathcal{I}}\|_2^2=1} \left\| \begin{bmatrix} (B_{\mathcal{A}\mathcal{A}})^{-1}A_{\mathcal{A}}v_{\mathcal{A}} \\ 0 \end{bmatrix} \right\|_2. \tag{21}$$

This is maximized when $v_{\mathcal{I}} = 0$, giving:

$$\|C_P\|_{\mathrm{op}} = \sup_{\|v_{\mathcal{A}}\|_2=1} \|(B_{\mathcal{A}\mathcal{A}})^{-1}A_{\mathcal{A}}v_{\mathcal{A}}\|_2 = \|(B_{\mathcal{A}\mathcal{A}})^{-1}A_{\mathcal{A}}\|_{\mathrm{op}}. \tag{22}$$

Theorem 7 (LASSO Lipschitz Bound via Conservative Jacobian)**.** *Let S be the LASSO solution map and let F be the fixed-point mapping Eq. (14) with conservative Jacobian $J_F = [A\ B]$. Let $(\lambda, \bar{\beta}) \in \operatorname{gph} S$ be a solution point,* *and let $\mathcal{A} = \{i \mid \bar\beta_i \neq 0$ or $|X_i^\top(X\bar\beta - y)| = \lambda\}$ be the active set. Assume that for every matrix $B_{\mathcal{A}\mathcal{A}}$ in the set*

$$J_{B_{\mathcal{A}\mathcal{A}}} = \left\{\left(\mathrm{Id}_{\mathcal{A}} - Q_x(\mathrm{Id}_{\mathcal{A}} - (X^\top X)_{\mathcal{A}\mathcal{A}})\right) : \begin{bmatrix} A & B \end{bmatrix} \in J_F(\lambda, \bar\beta)\right\} \tag{23}$$

is invertible. Then the local Lipschitz constant of S at λ is bounded by

$$\mathrm{Lip}(S;\lambda) \leq \sup_{J_F(\lambda,\bar\beta)} \left\|(B_{\mathcal{A}\mathcal{A}})^{-1} A_{\mathcal{A}}\right\|_{\mathrm{op}}, \tag{24}$$

where $A_{\mathcal{A}}$ and $B_{\mathcal{A}\mathcal{A}}$ are the submatrices of A and B corresponding to the active set $\mathcal{A}$.

Proof. The proof follows directly from the application of Theorem 3 and the block-matrix decomposition performed on Eq. (22). The implicit function theorem provides that the conservative Jacobian of the solution map $J_S(\lambda)$ is given by the set $\{-B^{-1}A : [A\ B] \in J_F(\lambda, \bar\beta)\}$. The Lipschitz constant is the supremum of the operator norms of elements in this set. Our analysis shows that $\|B^{-1}A\|_{\mathrm{op}} = \|(B_{\mathcal{A}\mathcal{A}})^{-1}A_{\mathcal{A}}\|_{\mathrm{op}}$, as desired. □

Discussion and Limitations. This theorem provides a rigorous and correct bound on the Lipschitz constant of the LASSO solution map using the conservative Jacobian framework. However, this result also highlights a few limitations of the approach, which motivates moving to the coderivative framework:

1. Verifiability: The bound requires a non-degeneracy condition, i.e., the invertibility of $B_{\mathcal{A}\mathcal{A}}$ for *all* selections in the conservative Jacobian. This $B_{\mathcal{A}\mathcal{A}}$ depends on Q_x, which is set-valued on the active non-support $\mathcal{E} \setminus \mathrm{supp}(\bar\beta)$. Verifying this for all $q_i \in [0,1]$ on this set can be computationally intractable.
2. Computability: The final bound $\sup \|(B_{\mathcal{A}\mathcal{A}})^{-1}A_{\mathcal{A}}\|_{\mathrm{op}}$ is not an explicit, human-readable quantity. It is the solution to a complex optimization problem over the set-valued conservative Jacobian. It does not provide the clean, geometric insight that one would hope for.
3. Sharpness: Because the conservative Jacobian J_F is a "box" that must contain the true subdifferential $\partial^C F$, it is often a loose over-approximation. The bound derived from J_F may be significantly looser than a bound derived directly from the geometric properties of the optimality conditions.

Our analysis so far demonstrates that the conservative Jacobian implicit function theorem is a blunt instrument for this class of structured nonsmooth problems. The resulting bound is complex and difficult to compute. This motivates the need for a more refined toolset, which the coderivative calculus, explored in the next section, provides. The coderivative framework operates directly on the normal cone geometry of the optimality conditions, bypassing the need to construct an explicit (and often loose) conservative Jacobian for a fixed-point mapping.

6 A Variational Analysis of LASSO Stability: Primal and Dual Perspectives

In the previous section, we demonstrated the limitations of applying the conservative Jacobian implicit function theorem to the LASSO. The resulting bound was complex and not easily computable. We now pivot to the more powerful framework of variational analysis and coderivative calculus, following the foundational work of Mordukhovich (2009; 2018). This approach allows us to directly analyze the geometry of the optimality conditions, yielding a sharper and more insightful bound.

We first introduce the essential definitions from variational analysis.

Definition 8 (Fundamental Cones)**.** *Let X be a Banach space, and let $C \subset X$ be a nonempty set with $x \in C$. The* tangent cone *(Bouligand-Severi contingent cone) to C at x is defined as*

$$\begin{aligned} T_C(x) &:= \limsup_{\tau \searrow 0} \Big[\tau^{-1}(C - x)\Big] \\ &= \Big\{\Delta x \in X \mid \exists\, (x_k \in C, \tau_k \searrow 0) \text{ such that } x_k \to x, \Delta x = \lim_{k\to\infty} \Big[\tau_k^{-1}(x_k - x)\Big]\Big\}. \end{aligned}$$

The Fréchet normal cone *to C at x is*

$$\widehat{N}_C(x) := \left\{ x^\star \in X^\star \;\middle|\; \limsup_{C \ni \tilde{x} \to x, \tilde{x} \neq x} \frac{\langle x^\star, \tilde{x} - x\rangle_X}{\|\tilde{x} - x\|_X} \le 0 \right\}.$$

The limiting (Mordukhovich) normal cone *in a finite-dimensional space $\mathbb{R}^n$ is the outer limit of the Fréchet normal cones:*

$$\begin{aligned} N_C(x) &:= \limsup_{C \ni \tilde{x} \to x} \widehat{N}_C(\tilde{x}) \\ &= \Big\{x^\star \in \mathbb{R}^n : \exists\, (x_k \in C, x_k^\star \in \widehat{N}_C(x_k)) \text{ such that } x_k \to x, x_k^\star \to x^\star\Big\}. \end{aligned}$$

Lemma 9. *Let $C \subset \mathbb{R}^n$ be nonempty, closed, and convex, and let $x \in C$. Then the Fréchet and limiting normal cones coincide with the normal cone of convex analysis:*

$$\widehat{N}_C(x) = N_C(x) = \partial \delta_C(x) = \Big\{x^\star \in \mathbb{R}^n \mid \langle x^\star, \tilde{x} - x\rangle \le 0,\ \forall \tilde{x} \in C\Big\}, \tag{25}$$

where δ_C is the indicator function of C. For a general closed set C, $\widehat{N}_C(x) \subseteq N_C(x)$.

Proof. For a convex set C, the subdifferential of the indicator function $\partial\delta_C(x)$ is precisely the set of vectors $x^\star$ defining a supporting hyperplane to C at x. The Fréchet normal cone $\widehat{N}_C(x)$ consists of all $x^\star \in \mathbb{R}^n$ such that $\langle x^\star, \tilde{x} - x\rangle \le o(\|\tilde{x} - x\|_2)$ for all $\tilde{x} \in C$. By convexity, $\langle x^\star, \tilde{x} - x\rangle \le 0$ for all $\tilde{x} \in C$ implies $\langle x^\star, \tilde{x} - x\rangle \le o(\|\tilde{x} - x\|_2)$, and vice-versa. Thus $\widehat{N}_C(x) = \partial\delta_C(x)$. Since $N_C(x)$ is the outer limit of $\widehat{N}_C(\tilde{x})$, and for convex sets $\tilde{x} \mapsto \widehat{N}_C(\tilde{x})$ is an outer-semicontinuous multifunction, the limit $N_C(x)$ coincides with $\widehat{N}_C(x)$. The final inclusion follows from the definition of the limiting normal cone. □

Definition 10 (Graphical Derivatives and Coderivatives)**.** *Let $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ be a multifunction.*

1. *The* graphical derivative *of F at $(\bar{x}, \bar{y}) \in \operatorname{gph} F$ is a multifunction $DF(\bar{x}|\bar{y}) : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ defined by*

$$\operatorname{gph} DF(\bar{x}|\bar{y}) := T_{\operatorname{gph} F}(\bar{x}, \bar{y}), \tag{26}$$

   i.e., $\Delta y \in DF(\bar{x}|\bar{y})(\Delta x) \iff (\Delta x, \Delta y) \in T_{\operatorname{gph} F}(\bar{x}, \bar{y})$.

2. *The* limiting coderivative *of F at $(\bar{x}, \bar{y}) \in \operatorname{gph} F$ is a multifunction $D^\star F(\bar{x}|\bar{y}) : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ defined by*

$$D^\star F(\bar{x}|\bar{y})(y^\star) := \Big\{x^\star \in \mathbb{R}^n : (x^\star, -y^\star) \in N_{\operatorname{gph} F}(\bar{x}, \bar{y})\Big\}, \quad \forall y^\star \in \mathbb{R}^m. \tag{27}$$

The graphical derivative and coderivative are fundamental because they provide primal and dual characterizations of the solution map's sensitivity. The (local) Lipschitz constant of S, also known as its modulus of calmness, is equal to the norm of its graphical derivative and its coderivative (Rockafellar and Wets, 2009, Theorem 5.11), $\operatorname{Lip}(S;\bar{p}) = \|DS(\bar{p}|\bar{x})\| = \|D^\star S(\bar{p}|\bar{x})\|$. We will now compute this quantity for the LASSO.

Application to the LASSO Solution Map. We return to the LASSO optimality condition from Eq. (13), with parameter $p = \lambda$ and variable $x = \beta$:

$$0 \in \Psi(\lambda,\beta) := \underbrace{X^\mathsf{T}(X\beta - y)}_{f(\beta)} + \underbrace{\lambda\partial\|\cdot\|_1(\beta)}_{Q(\lambda,\beta)}. \tag{28}$$

The solution map is $S(\lambda) = \{\beta \mid 0 \in \Psi(\lambda,\beta)\}$. Let $(\bar{\lambda},\bar{\beta}) \in \operatorname{gph} S$ be a reference solution.

6.1 Primal Approach: Graphical Derivative

The most direct method is to compute the graphical derivative $DS(\bar{\lambda}|\bar{\beta})$.

$$\Delta\beta \in DS(\bar{\lambda}|\bar{\beta})(\Delta\lambda) \iff (\Delta\lambda,\Delta\beta) \in T_{\operatorname{gph} S}(\bar{\lambda},\bar{\beta}). \tag{29}$$

The graph $\operatorname{gph} S$ is the solution set to the generalized equation $0 \in f(\beta) + Q(\lambda,\beta)$. By definition, $(\Delta\lambda,\Delta\beta) \in T_{\operatorname{gph} S}(\bar{\lambda},\bar{\beta})$ if and only if they satisfy the linearized generalized equation:

$$0 \in D\Psi(\bar{\lambda},\bar{\beta}|0)(\Delta\lambda,\Delta\beta), \tag{30}$$

where $D\Psi$ is the graphical derivative of Ψ. Using the sum rule for graphical derivatives (Rockafellar and Wets, 2009, Exercise 10.43),

$$\begin{aligned} D\Psi(\bar{\lambda},\bar{\beta}|0)(\Delta\lambda,\Delta\beta) &= \nabla f(\bar{\beta})\Delta\beta + DQ(\bar{\lambda},\bar{\beta}|\bar{q})(\Delta\lambda,\Delta\beta) \\ &= X^\mathsf{T}X\Delta\beta + DQ(\bar{\lambda},\bar{\beta}|\bar{q})(\Delta\lambda,\Delta\beta), \end{aligned} \tag{31}$$

where $\bar{q} = -f(\bar{\beta}) = X^\mathsf{T}(y - X\bar{\beta})$ is the particular subgradient satisfying the optimality condition. The term $Q(\lambda,\beta) = \lambda\partial\|\beta\|_1$ is a product. Let $C = \operatorname{gph}(\partial\|\cdot\|_1)$, so $\operatorname{gph} Q = \{(\lambda,\beta,\lambda w) \mid (\beta,w) \in C\}$. The tangent cone $T_{\operatorname{gph} Q}(\bar{\lambda},\bar{\beta},\bar{\lambda}\bar{w})$ is

$$T_{\operatorname{gph} Q}(\bar{\lambda},\bar{\beta},\bar{\lambda}\bar{w}) = \Big\{(\Delta\lambda,\Delta\beta,\Delta q) \mid \Delta q = \bar{w}\Delta\lambda + \bar{\lambda}\Delta w,\ (\Delta\beta,\Delta w) \in T_C(\bar{\beta},\bar{w})\Big\}. \tag{32}$$

Thus, $DQ(\Delta\lambda,\Delta\beta) = \{\bar{w}\Delta\lambda + \bar{\lambda}\Delta w \mid (\Delta\beta,\Delta w) \in T_C(\bar{\beta},\bar{w})\}$. The linearized equation Eq. (30) becomes that of finding $(\Delta\lambda,\Delta\beta)$ such that

$$\exists(\Delta\beta,\Delta w) \in T_C(\bar{\beta},\bar{w}) \quad \text{s.t.} \quad 0 = X^\mathsf{T}X\Delta\beta + \bar{w}\Delta\lambda + \bar{\lambda}\Delta w. \tag{33}$$

We assume strict complementarity, so the active set $\mathcal{E} = \{i \mid |\bar{q}_i| = \bar{\lambda}\}$ is identical to the support $\operatorname{supp}(\bar{\beta}) = \{i \mid \bar{\beta}_i \neq 0\}$. Let $\mathcal{I} = \{1,\dots,p\} \setminus \mathcal{E}$ be the inactive set. For $i \in \mathcal{E}$: $\bar{\beta}_i \neq 0, \bar{w}_i = \operatorname{sgn}(\bar{\beta}_i)$. The graph $\operatorname{gph}(\partial|\cdot|)$ is locally smooth. $T_{\operatorname{gph}(\partial|\cdot|)}(\bar{\beta}_i,\bar{w}_i) = \{(\Delta\beta_i,\Delta w_i) \mid \Delta w_i = 0\}$. For $i \in \mathcal{I}$: $\bar{\beta}_i = 0, |\bar{w}_i| < \bar{\lambda}$. The graph is locally $\{0\} \times (-1,1)$. $T_{\operatorname{gph}(\partial|\cdot|)}(\bar{\beta}_i,\bar{w}_i) = \{(\Delta\beta_i,\Delta w_i) \mid \Delta\beta_i = 0, \Delta w_i \in \mathbb{R}\}$. So, the condition $(\Delta\beta,\Delta w) \in T_C(\bar{\beta},\bar{w})$ implies $\Delta w_\mathcal{E} = 0$ and $\Delta\beta_\mathcal{I} = 0$. We substitute these into Eq. (33) and split by components:

- <u>Case $i \in \mathcal{E}$</u>:

$$
\begin{aligned}
0 &= (X^\top X \Delta\beta)_i + \bar{w}_i \Delta\lambda + \bar{\lambda}\Delta w_i && (34)\\
0 &= (X^\top X)_{\mathcal{E},\mathcal{E}} \Delta\beta_{\mathcal{E}} + (X^\top X)_{\mathcal{E},\mathcal{I}} \underbrace{\Delta\beta_{\mathcal{I}}}_{=0} + \bar{w}_{\mathcal{E}}\Delta\lambda + \underbrace{\bar{\lambda}\Delta w_{\mathcal{E}}}_{=0} && (35)\\
\Longrightarrow (X_{\mathcal{E}}^\top X_{\mathcal{E}})\Delta\beta_{\mathcal{E}} &= -\bar{w}_{\mathcal{E}}\Delta\lambda. && (36)
\end{aligned}
$$

- For $i \in \mathcal{I}$:

$$
\begin{aligned}
0 &\in (X^\top X \Delta\beta)_i + \bar{w}_i \Delta\lambda + \bar{\lambda}\Delta w_i && (37)\\
0 &\in (X^\top X)_{\mathcal{I},\mathcal{E}} \Delta\beta_{\mathcal{E}} + (X^\top X)_{\mathcal{I},\mathcal{I}} \underbrace{\Delta\beta_{\mathcal{I}}}_{=0} + \bar{w}_{\mathcal{I}}\Delta\lambda + \bar{\lambda}\mathbb{R} && (38)\\
0 &\in (X^\top X)_{\mathcal{I},\mathcal{E}} \Delta\beta_{\mathcal{E}} + \bar{w}_{\mathcal{I}}\Delta\lambda + \mathbb{R}. && (39)
\end{aligned}
$$

The second case ($i \in \mathcal{I}$) is an inclusion that can always be satisfied by choosing an appropriate $\Delta w_i \in \mathbb{R}$, so it imposes no constraint on $\Delta\beta_{\mathcal{E}}$. The entire sensitivity is driven by the first case ($i \in \mathcal{E}$), which is a linear system. Assuming $X_{\mathcal{E}}^\top X_{\mathcal{E}}$ is invertible (a standard non-degeneracy condition), we can solve for $\Delta\beta_{\mathcal{E}}$:

$$
\Delta\beta_{\mathcal{E}} = -(X_{\mathcal{E}}^\top X_{\mathcal{E}})^{-1}\bar{w}_{\mathcal{E}}\Delta\lambda. \tag{40}
$$

Since $\Delta\beta_{\mathcal{I}} = 0$, we have $\|\Delta\beta\|_2 = \|\Delta\beta_{\mathcal{E}}\|_2$. The norm of the graphical derivative is

$$
\begin{aligned}
\mathrm{Lip}(S;\bar{\lambda}) = \|DS(\bar{\lambda}|\bar{\beta})\| &= \sup_{|\Delta\lambda|\le 1} \sup_{\Delta\beta \in DS(\bar{\lambda}|\bar{\beta})(\Delta\lambda)} \|\Delta\beta\|_2 \\
&= \sup_{|\Delta\lambda|\le 1} \left\| -(X_{\mathcal{E}}^\top X_{\mathcal{E}})^{-1}\bar{w}_{\mathcal{E}}\Delta\lambda \right\|_2 \\
&= \left\| (X_{\mathcal{E}}^\top X_{\mathcal{E}})^{-1}\bar{w}_{\mathcal{E}} \right\|_2 \\
&\le \left\| (X_{\mathcal{E}}^\top X_{\mathcal{E}})^{-1} \right\|_{\mathrm{op}} \cdot \|\bar{w}_{\mathcal{E}}\|_2 .
\end{aligned} \tag{41}
$$

We have $\|\bar{w}_{\mathcal{E}}\|_2 = \sqrt{\sum_{i\in\mathcal{E}} \bar{w}_i^2} = \sqrt{\sum_{i\in\mathcal{E}} \mathrm{sgn}(\bar{\beta}_i)^2} = \sqrt{|\mathcal{E}|}$. And $\left\|(X_{\mathcal{E}}^\top X_{\mathcal{E}})^{-1}\right\|_{\mathrm{op}} = 1/\lambda_{\min}(X_{\mathcal{E}}^\top X_{\mathcal{E}})$. This leads to the following result.

Theorem 11 (LASSO Lipschitz Bound via Graphical Derivative)**.** *Let S be the LASSO solution map. Let $(\bar{\lambda}, \bar{\beta}) \in \mathrm{gph}\, S$ be a solution that satisfies*

1. *<u>Strict complementarity</u>: The active set $\mathcal{E} = \{i \mid |(X^\top(X\bar{\beta} - y))_i| = \bar{\lambda}\}$ and the support $\mathrm{supp}(\bar{\beta})$ are identical.*
2. *<u>Non-degeneracy</u>: The restricted Gram matrix $X_{\mathcal{E}}^\top X_{\mathcal{E}}$ is invertible.*

Then, the solution map S is locally single-valued and Lipschitz continuous at $\bar{\lambda}$, with a modulus bounded by

$$
\mathrm{Lip}(S;\bar{\lambda}) = \|DS(\bar{\lambda}|\bar{\beta})\| \le \frac{\sqrt{|\mathcal{E}|}}{\lambda_{\min}(X_{\mathcal{E}}^\top X_{\mathcal{E}})}. \tag{42}
$$

Proof. The derivation above is the proof. The conditions ensure that $DS(\bar{\lambda}|\bar{\beta})(\Delta\lambda)$ is a singleton for each $\Delta\lambda$, so S is locally single-valued and its graph is locally a C^1-manifold, hence Lipschitz continuous. $\square$

6.2 Dual Approach: Coderivative

We can also compute the Lipschitz modulus via the coderivative norm $\mathrm{Lip}(S) = \|D^\star S(\bar{\lambda}|\bar{\beta})\|$. We need to compute

$$D^\star S(\bar{\lambda}|\bar{\beta})(\beta^\star) = \{\lambda^\star \in \mathbb{R} \mid (\lambda^\star, -\beta^\star) \in N_{\mathrm{gph}\, S}(\bar{\lambda}, \bar{\beta})\}. \tag{43}$$

We must find the normal cone $N_{\mathrm{gph}\, S}(\bar{\lambda}, \bar{\beta})$. Recall $\mathrm{gph}\, S = \{(\lambda, \beta) : \exists w, H(\lambda, \beta, w) = 0, (\beta, w) \in C\}$, where $H(\lambda, \beta, w) = X^\top (X\beta - y) + \lambda w$ and $C = \mathrm{gph}(\partial\|\cdot\|_1)$. Under a suitable qualification (MFCQ), the sum rule for normal cones (Rockafellar and Wets, 2009, Theorem 6.14) holds:

$$N_{\mathrm{gph}\, S}(\bar{\lambda}, \bar{\beta}) \subseteq \left\{ (\lambda^\star, \beta^\star) \mid \exists (\beta_C^\star, w_C^\star) \in N_C(\bar{\beta}, \bar{w}), \exists z^\star \in \mathbb{R}^p \text{ s.t. } \begin{bmatrix} \lambda^\star \\ \beta^\star \\ 0 \end{bmatrix} = \nabla H(\bar{\lambda}, \bar{\beta}, \bar{w})^\star z^\star + \begin{bmatrix} 0 \\ \beta_C^\star \\ w_C^\star \end{bmatrix} \right\}. \tag{44}$$

The adjoint of the Jacobian $\nabla H = [\, \underbrace{\bar{w}}_{\lambda} \;\; \underbrace{X^\top X}_{\beta} \;\; \underbrace{\bar{\lambda}\mathrm{Id}}_{w} \,]$ is

$$\nabla H(\bar{\lambda}, \bar{\beta}, \bar{w})^\star z^\star = \begin{bmatrix} \langle \bar{w}, z^\star \rangle \\ X^\top X z^\star \\ \bar{\lambda} z^\star \end{bmatrix}. \tag{45}$$

Plugging this in, $(\lambda^\star, -\beta^\star) \in N_{\mathrm{gph}\, S}(\bar{\lambda}, \bar{\beta})$ if $\exists z^\star \in \mathbb{R}^p, (\beta_C^\star, w_C^\star) \in N_C(\bar{\beta}, \bar{w})$ s.t.

1. $\lambda^\star = \langle \bar{w}, z^\star \rangle$
2. $-\beta^\star = X^\top X z^\star + \beta_C^\star$
3. $0 = \bar{\lambda} z^\star + w_C^\star \implies w_C^\star = -\bar{\lambda} z^\star$

As before, we use the structure of $N_C(\bar{\beta}, \bar{w})$ under strict complementarity $\mathcal{E} = \mathrm{supp}(\bar{\beta})$:

$$(\beta_C^\star, w_C^\star) \in N_C(\bar{\beta}, \bar{w}) \iff \beta_{C,\mathcal{I}}^\star = 0 \text{ and } w_{C,\mathcal{E}}^\star = 0. \tag{46}$$

From (3), $w_C^\star = -\bar{\lambda} z^\star$. Since $w_{C,\mathcal{E}}^\star = 0$ and $\bar{\lambda} > 0$, we must have $z_{\mathcal{E}}^\star = 0$. Now we use this to characterize $\lambda^\star$ in terms of $\beta^\star$. From (2), for $i \in \mathcal{I}$: $-\beta_{\mathcal{I}}^\star = (X^\top X z^\star)_{\mathcal{I}} + \beta_{C,\mathcal{I}}^\star = (X^\top X)_{\mathcal{I},\mathcal{I}} z_{\mathcal{I}}^\star + (X^\top X)_{\mathcal{I},\mathcal{E}} z_{\mathcal{E}}^\star + 0$. Since $z_{\mathcal{E}}^\star = 0$, this simplifies to $-\beta_{\mathcal{I}}^\star = (X_{\mathcal{I}}^\top X_{\mathcal{I}}) z_{\mathcal{I}}^\star$. If $X_{\mathcal{I}}^\top X_{\mathcal{I}}$ is invertible, $z_{\mathcal{I}}^\star = -(X_{\mathcal{I}}^\top X_{\mathcal{I}})^{-1} \beta_{\mathcal{I}}^\star$. From (1), $\lambda^\star = \langle \bar{w}, z^\star \rangle = \langle \bar{w}_{\mathcal{E}}, z_{\mathcal{E}}^\star \rangle + \langle \bar{w}_{\mathcal{I}}, z_{\mathcal{I}}^\star \rangle = 0 + \langle \bar{w}_{\mathcal{I}}, z_{\mathcal{I}}^\star \rangle$. Substituting $z_{\mathcal{I}}^\star$: $\lambda^\star = -\langle \bar{w}_{\mathcal{I}}, (X_{\mathcal{I}}^\top X_{\mathcal{I}})^{-1} \beta_{\mathcal{I}}^\star \rangle$. This derivation shows that $D^\star S(\bar{\lambda}|\bar{\beta})(\beta^\star)$ is a singleton (a linear functional of $\beta^\star$) only if $\beta_{\mathcal{E}}^\star$ is constrained by the remaining equation. From (2), for $i \in \mathcal{E}$: $-\beta_{\mathcal{E}}^\star = (X^\top X z^\star)_{\mathcal{E}} + \beta_{C,\mathcal{E}}^\star = (X^\top X)_{\mathcal{E},\mathcal{I}} z_{\mathcal{I}}^\star + \beta_{C,\mathcal{E}}^\star$. This dual derivation is far more complex and shows that the coderivative $D^\star S$ maps a general $\beta^\star$ to $\lambda^\star$ only if $\beta^\star$ satisfies this constraint. This complexity reinforces that for this specific problem, the primal graphical derivative approach is more direct for computing the Lipschitz modulus. Both, however, are derived from the same geometric principles.

7 Robust Lipschitzian Stability of Semialgebraic Deep Equilibrium Networks via Coderivative Calculus

Having demonstrated the precision of coderivative calculus on a canonical *convex* nonsmooth problem, we now escalate the complexity. We will apply this same advanced machinery to the *non-convex* and highly-structured setting of implicit neural networks, showing the framework is robust enough to handle modern deep learning architectures.

We now leverage the advanced coderivative calculus and robust stability frameworks, particularly from the seminal work on parametric variational systems (Mordukhovich, 2009), to establish verifiable conditions for the Lipschitz-like stability (Aubin property) of the solution map. We focus on the solution map $z : \mathbb{R}^{m\times m} \times \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ associated with the semialgebraic deep equilibrium net (DEQ) fixed-point equation

$$0 \in \Phi(W,b,z) := z - \sigma(Wz+b), \tag{47}$$

where $\sigma : \mathbb{R}^m \to \mathbb{R}^m$ is a semialgebraic function (e.g., ReLU, Leaky ReLU, or even a proximal operator). This semialgebraicity ensures that Φ is a semialgebraic mapping and $\operatorname{gph}\Phi$ is a semialgebraic set.

We recall that a set $S \subset \mathbb{R}^n$ is *semialgebraic* if it is a finite union of sets defined by a finite number of polynomial equalities and inequalities. This class of sets is stable under finite unions, intersections, complements, and, crucially, under linear projections (or via the Tarski-Seidenberg Theorem). This final property ensures that if a multifunction is semialgebraic, its domain, range, and graph are all semialgebraic sets. We will use this powerful geometric property as the main tool of our subsequent analysis.

Problem Formulation. We frame the DEQ stability problem as a parametric generalized equation (GE). The perturbation parameters are $x := (W,b) \in X := \mathbb{R}^{m\times m} \times \mathbb{R}^m$. The decision variable is $y := z \in Y := \mathbb{R}^m$. The auxiliary space is $Z := \mathbb{R}^m$. All spaces are finite-dimensional Euclidean, and thus are Asplund.

A common approach is to formulate the GE as $0 \in f(x,y)+Q(y)$ by setting $f(x,y) = \Phi(x,y)$ and $Q(y) = \{0\}$. However, this fails to separate the parameters x from the decision variable y inside the mapping f, as Φ contains the coupled term Wz.

A more rigorous and powerful formulation, which correctly fits the standard form of parametric GEs analyzed in Mordukhovich (2009), is to set the base mapping $f : X \times Y \to Z$ to be trivial and embed all complexity into the set-valued field Q. Let:

$$f(x,y) := 0 \quad \text{(the zero map)} \tag{48}$$

$$Q(x,y) := \Phi(W,b,z) = z - \sigma(Wz+b) \tag{49}$$

The solution map for the equilibrium z as a function of the parameters (W,b) is then

$$S(x) := \{y \in Y \mid 0 \in f(x,y) + Q(x,y)\} \equiv \{z \in \mathbb{R}^m \mid 0 \in \Phi(W,b,z)\}. \tag{50}$$

This formulation $S(x) = Q(x,\cdot)^{-1}(0)$ fits precisely into the "full" parametric framework of (Mordukhovich, 2009, Theorem 4.4). The base mapping $f \equiv 0$ is trivially C^∞ (and thus strictly differentiable). The multifunction $Q(x,y)$ is semialgebraic since σ is, and its graph is locally closed. Furthermore, as a semialgebraic mapping, Q is *sequentially normally compact (SNC)* at every point of its graph, a key hypothesis of the stability theorems.

Lipschitzian Stability via Coderivative Calculus. Our primary objective is to find the local Lipschitz modulus of S around a reference equilibrium $(\bar{x},\bar{y}) = ((\bar{W},\bar{b}),\bar{z}) \in \operatorname{gph} S$. This modulus, $\operatorname{Lip} S(\bar{x},\bar{y})$, is defined by the existence of neighborhoods U of $\bar{x}$ and V of $\bar{y}$ and a constant $\ell \geq 0$ such that

$$S(x) \cap V \subset S(u) + \ell \|x-u\| \cdot \mathbb{B}_Y \quad \forall x,u \in U. \tag{51}$$

By (Mordukhovich, 2005, Theorem 2.1), since X, Y are finite-dimensional (and thus Asplund and strongly coderivatively normal), this stability is equivalent to the coderivative criterion $D^\star S(\bar{x}|\bar{y})(0) = \{0\}$, and the exact Lipschitz modulus is given by the norm of the coderivative:

$$\operatorname{Lip} S(\bar{x},\bar{y}) = \|D^\star S(\bar{x}|\bar{y})\| := \sup_{\|y^\star\|\leq 1} \sup\{\|x^\star\| \mid x^\star \in D^\star S(\bar{x}|\bar{y})(y^\star)\}. \tag{52}$$

To compute this, we first recall a few necessary tools.

Theorem 12 (Characterizations of Lipschitzian stability, (Mordukhovich, 2009, Theorem 4.4))**.** *Let $S(x) = \{y \in Y \mid 0 \in f(x,y) + Q(x,y)\}$ be the solution map, where $f : X \times Y \to Z$ is strictly Lipschitzian at $(\bar{x},\bar{y}) \in \operatorname{gph} S$, where $Q : X \times Y \rightrightarrows Z$ is locally closed-graph around $(\bar{x},\bar{y},\bar{v})$ with $\bar{v} := -f(\bar{x},\bar{y})$ and SNC at this point, and where X, Y, Z are Asplund.*

(i) The map S is Lipschitz-like around $(\bar{x},\bar{y})$ if the following qualification condition holds:

$$\left[(x^*,0) \in \partial\langle z^*, f\rangle(\bar{x},\bar{y}) + D^*_N Q(\bar{x},\bar{y},\bar{v})(z^*)\right] \implies x^* = 0, z^* = 0. \tag{53}$$

(ii) If S is Lipschitz-like and $\dim X < \infty$*, its Lipschitz modulus is bounded by:*

$$\begin{aligned} \operatorname{Lip} S(\bar{x},\bar{y}) \leq \sup\{\|x^*\| \mid \exists z^* \in Z^* \quad \textit{with } (x^*,-y^*) \in \partial\langle z^*, f\rangle(\bar{x},\bar{y}) \\ +D^*_N Q(\bar{x},\bar{y},\bar{v})(z^*), \quad \left\|y^*\right\| \leq 1\Big\}. \end{aligned} \tag{54}$$

We apply this theorem to our formulation where $f \equiv 0$ and $Q = \Phi$.

- $f \equiv 0$ is C^∞, so it is strictly Lipschitzian and $\partial\langle z^*, f\rangle(\bar{x},\bar{y}) = \{0\}$.
- $\bar{v} = -f(\bar{x},\bar{y}) = 0$.
- $Q = \Phi$ is semialgebraic, locally closed-graph, and SNC.

The qualification condition Eq. (53) becomes

$$\left[(x^*,0) \in \{0\} + D^*_N \Phi(\bar{x},\bar{y},0)(z^*)\right] \implies x^* = 0, z^* = 0. \tag{55}$$

And the Lipschitz bound Eq. (54) simplifies to

$$\operatorname{Lip} S(\bar{x},\bar{y}) \leq \sup\Big\{\|x^*\| \mid \exists z^* \in Z^*, (x^*,-y^*) \in D^*_N \Phi(\bar{x},\bar{y},0)(z^*), \left\|y^*\right\| \leq 1\Big\}. \tag{56}$$

By definition of the coderivative (Def. 10), $(x^*,-y^*) \in D_N^*\Phi(\bar{x},\bar{y},0)(z^*)$ means $(x^*,y^*,-z^*) \in N_{\mathrm{gph}\,\Phi}(\bar{x},\bar{y},0)$. This is exactly the coderivative of the solution map $S = \Phi(\cdot,\cdot)^{-1}(0)$.

Thus, the entire problem reduces to analyzing $D_N^*\Phi(\bar{W},\bar{b},\bar{z}|0)(z^\star)$. Let $\psi(W,b,z) := \langle z^\star, \Phi(W,b,z)\rangle$. Since Φ is semialgebraic and locally Lipschitz (as σ is), it is strictly Lipschitzian, and the scalarization lemma (Lemma 20) holds:

$$D_N^\star\Phi(\bar{W},\bar{b},\bar{z}|0)(z^\star) = \partial\psi(\bar{W},\bar{b},\bar{z}), \tag{57}$$

where ∂ is the limiting subdifferential. We have $\psi(W,b,z) = \langle z^\star, z - \sigma(g(W,b,z))\rangle$, where $g(W,b,z) = Wz + b$. This is a composition of $\psi_1(z) = \langle z^\star, z\rangle$ and $\psi_2(u) = \langle z^\star, \sigma(u)\rangle$ with the smooth map g. Using the subdifferential sum and chain rules (Rockafellar and Wets, 2009, Thm 10.6, Thm 10.49),

$$\partial\psi(\bar{W},\bar{b},\bar{z}) \subseteq \partial\psi_1(\bar{W},\bar{b},\bar{z}) - \partial(\psi_2 \circ g)(\bar{W},\bar{b},\bar{z}). \tag{58}$$

Since g is smooth and ψ_2 is Lipschitz, the chain rule holds with equality:

$$\partial(\psi_2 \circ g)(\bar{W},\bar{b},\bar{z}) = \nabla g(\bar{W},\bar{b},\bar{z})^\star \partial\psi_2(\bar{u}), \tag{59}$$

where $\bar{u} = g(\bar{W},\bar{b},\bar{z}) = \bar{W}\bar{z} + \bar{b}$. We have:

- $\partial\psi_1(\bar{W},\bar{b},\bar{z}) = (0,0,z^\star)$ (as a partial subdifferential w.r.t (W,b,z)).
- $\partial\psi_2(\bar{u}) = \partial\langle z^\star,\sigma\rangle(\bar{u})$. Let $u^\star \in \partial\langle z^\star,\sigma\rangle(\bar{u})$. By scalarization, $u^\star \in D^\star\sigma(\bar{u}|\sigma(\bar{u}))(z^\star)$.
- $\nabla g(\bar{W},\bar{b},\bar{z})$ is the Jacobian of g w.r.t (W,b,z). Its adjoint $\nabla g^\star u^\star$ maps $u^\star \in \mathbb{R}^m$ to $(W^\star, b^\star, z_g^\star) \in X \times Y$:
  - $\langle (W^\star,b^\star,z_g^\star),(\Delta W,\Delta b,\Delta z)\rangle = \langle u^\star, \nabla g \cdot (\Delta W,\Delta b,\Delta z)\rangle$
  - $\nabla g \cdot (\Delta W,\Delta b,\Delta z) = (\Delta W)\bar{z} + \bar{W}(\Delta z) + \Delta b$
  - $\langle u^\star, (\Delta W)\bar{z} + \bar{W}(\Delta z) + \Delta b\rangle = \langle u^\star,(\Delta W)\bar{z}\rangle + \langle u^\star, \bar{W}\Delta z\rangle + \langle u^\star, \Delta b\rangle$
  - $= \mathrm{Tr}((u^\star\bar{z}^\top)^\top \Delta W) + \langle \bar{W}^\top u^\star, \Delta z\rangle + \langle u^\star, \Delta b\rangle$
- By identification, $\nabla g(\bar{W},\bar{b},\bar{z})^\star u^\star = \begin{bmatrix} u^\star\bar{z}^\top \\ u^\star \\ \bar{W}^\top u^\star \end{bmatrix}$.

Combining these parts, any element $(W^\star, b^\star, z^\star_{part}) \in \partial\psi(\bar{W},\bar{b},\bar{z})$ is of the form:

$$\begin{bmatrix} W^\star \\ b^\star \\ z^\star_{part} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ z^\star \end{bmatrix} - \begin{bmatrix} u^\star\bar{z}^\top \\ u^\star \\ \bar{W}^\top u^\star \end{bmatrix} = \begin{bmatrix} -u^\star\bar{z}^\top \\ -u^\star \\ z^\star - \bar{W}^\top u^\star \end{bmatrix}, \quad \text{for some } u^\star \in \partial\langle z^\star,\sigma\rangle(\bar{u}). \tag{60}$$

Qualification Condition. We check the qualification Eq. (53) using our $f = 0, Q = \Phi$ formulation. We need to show:

$$\left[(W^\star,b^\star,0) \in D_N^*\Phi(\bar{W},\bar{b},\bar{z}|0)(z^*)\right] \implies (W^\star,b^\star) = (0,0), z^* = 0.$$

From Eq. (60), this means we set $z^\star_{part} = 0$:

$$z^\star - \bar{W}^\top u^\star = 0 \implies z^\star = \bar{W}^\top u^\star, \quad \text{where } u^\star \in \partial\langle z^\star,\sigma\rangle(\bar{u}). \tag{61}$$

This implies $b^\star = -u^\star$ and $W^\star = -u^\star \bar{z}^\top$. To show $W^\star = 0, b^\star = 0, z^\star = 0$, it suffices to show $u^\star = 0$. Since σ is a proximal operator, it is 1-Lipschitz. The subdifferential $\partial\langle \cdot, \sigma\rangle(\bar{u})$ is norm-bounded by 1. That is, for any $u^\star \in \partial\langle z^\star, \sigma\rangle(\bar{u})$, we have $\|u^\star\| \le \|z^\star\|$. Substituting $z^\star = \bar{W}^\top u^\star$:

$$\|u^\star\| \le \|\bar{W}^\top u^\star\| \le \|\bar{W}^\top\|_{\mathrm{op}} \|u^\star\| = \|\bar{W}\|_{\mathrm{op}} \|u^\star\|. \tag{62}$$

This can be written as $\|u^\star\|(1 - \|\bar{W}\|_{\mathrm{op}}) \le 0$. If we assume the standard DEQ condition that $\bar{W}$ is a contraction, $\|\bar{W}\|_{\mathrm{op}} < 1$, then $(1 - \|\bar{W}\|_{\mathrm{op}}) > 0$. This implies $\|u^\star\| \le 0$, so $u^\star = 0$. If $u^\star = 0$, then $b^\star = -u^\star = 0$, $W^\star = -u^\star \bar{z}^\top = 0$, and $z^\star = \bar{W}^\top u^\star = 0$. The qualification condition holds if $\|\bar{W}\|_{\mathrm{op}} < 1$.

Theorem 13 (Lipschitz Stability of DEQs)**.** *Let S be the solution map for the DEQ Eq. (54) with a semialgebraic, 1-Lipschitz activation σ. At an equilibrium point $((\bar{W}, \bar{b}), \bar{z})$ where $\bar{W}$ is a contraction, i.e., $\|\bar{W}\|_{\mathrm{op}} = \rho < 1$, the solution map S is locally Lipschitz-like. Its Lipschitz modulus is bounded by:*

$$\operatorname{Lip} S(\bar{W}, \bar{b} | \bar{z}) \le \frac{\sqrt{1 + \|\bar{z}\|_2^2}}{1 - \|\bar{W}\|_{\mathrm{op}}}. \tag{63}$$

Proof. The qualification condition Eq. (53) holds as shown above. We apply the Lipschitz bound formula Eq. (54) with $f \equiv 0$, $Q = \Phi$, $x = (W, b)$, $y = z$, and $\bar{v} = 0$,

$$\operatorname{Lip} S(\bar{x}, \bar{y}) \le \sup\left\{ \|x^*\| \mid \exists z^*, (x^*, -y^*) \in D_N^* \Phi(\bar{x}, \bar{y}, 0)(z^*), \left\| y^* \right\| \le 1 \right\}. \tag{64}$$

From Eq. (60), this is

$$\operatorname{Lip} S \le \sup_{\substack{\|z_{in}^\star\| \le 1, z^\star \in \mathbb{R}^m \\ u^\star \in \partial\langle z^\star, \sigma\rangle(\bar{u})}} \left\{ \left\| \begin{bmatrix} -u^\star \bar{z}^\top \\ -u^\star \end{bmatrix} \right\|_{\mathrm{op}} \,\middle|\, z^\star = \bar{W}^\top u^\star - z_{in}^\star \right\}. \tag{65}$$

We use the operator norm $\|(M, v)\|_{\mathrm{op}} := \sqrt{\|M\|_F^2 + \|v\|_2^2}$.

$$\left\| \begin{bmatrix} -u^\star \bar{z}^\top \\ -u^\star \end{bmatrix} \right\|_{\mathrm{op}}^2 = \| - u^\star \bar{z}^\top \|_F^2 + \| - u^\star \|_2^2 = (\|u^\star\|_2 \|\bar{z}\|_2)^2 + \|u^\star\|_2^2 = \|u^\star\|_2^2 (1 + \|\bar{z}\|_2^2). \tag{66}$$

As shown in the qualification step, the 1-Lipschitz property of σ implies $\|u^\star\| \le \|z^\star\|$. Substituting $z^\star = \bar{W}^\top u^\star - z_{in}^\star$:

$$\|u^\star\| \le \|\bar{W}^\top u^\star - z_{in}^\star\| \le \|\bar{W}^\top\|_{\mathrm{op}} \|u^\star\| + \|z_{in}^\star\|. \tag{67}$$

Letting $\rho = \|\bar{W}\|_{\mathrm{op}} < 1$, we have $\|u^\star\|(1 - \rho) \le \|z_{in}^\star\|$. Since we supremum over $\|z_{in}^\star\| \le 1$, we have $\|u^\star\| \le \frac{1}{1-\rho}$. Substituting this into the norm calculation:

$$\left\| \begin{bmatrix} W^\star \\ b^\star \end{bmatrix} \right\|_{\mathrm{op}} = \|u^\star\|_2 \sqrt{1 + \|\bar{z}\|_2^2} \le \frac{\sqrt{1 + \|\bar{z}\|_2^2}}{1 - \rho}. \tag{68}$$

Taking the supremum over all such elements gives the bound. □

Theorem 14 (Finiteness and Computability)**.** *For a semialgebraic DEQ, the Lipschitz modulus* $\operatorname{Lip} S(\bar{W}, \bar{b}|\bar{z})$ *is the solution to a semialgebraic optimization problem. It is therefore a finite, semialgebraic function of the equilibrium point and (in theory) computable via Cylindrical Algebraic Decomposition (CAD).*

Proof. The graph of S is semialgebraic by Tarski-Seidenberg, as $\operatorname{gph} S = \Phi^{-1}(0)$ and $\operatorname{gph}\Phi$ is semialgebraic. The limiting normal cone $N_{\operatorname{gph} S}(\bar{W}, \bar{b}, \bar{z})$ is a semialgebraic set. The coderivative $D_N^\star S(\bar{W}, \bar{b}|\bar{z})$ has a semialgebraic graph, as it is a linear projection of the normal cone. The operator norm $\|(W^\star, b^\star)\|_{\mathrm{op}}$ is a semialgebraic function (its square is a polynomial). The expression for the Lipschitz modulus,

$$\operatorname{Lip} S = \sup_{\|y^\star\|_2^2=1} \inf\Big\{r \ge 0 \mid \exists (W^\star, b^\star) \in D_N^\star S(\bar{W}, \bar{b}|\bar{z})(y^\star) \text{ s.t. } \|(W^\star, b^\star)\|_{\mathrm{op}}^2 \le r^2\Big\}, \tag{69}$$

is a first-order formula over the real field. By the Tarski-Seidenberg theorem, the set of achievable $\ell = \operatorname{Lip} S$ is semialgebraic (a point or interval). Since Theorem 13 proves the modulus is bounded (finite), this value is well-defined and computable via CAD. □

Example (ReLU DEQ)**.** *For* $\sigma = \mathrm{ReLU}$, $\operatorname{gph}\sigma$ *is a finite union of polyhedral cones, which is semialgebraic. As* ReLU *is 1-Lipschitz, Theorem 13 applies directly, giving the explicit bound* $\operatorname{Lip} S \le (1 - \|\bar{W}\|_{\mathrm{op}})^{-1}\sqrt{1 + \|\bar{z}\|_2^2}$ *at any equilibrium where* $\|\bar{W}\|_{\mathrm{op}} < 1$. ♣

This result provides a rigorous, computable, and non-vacuous bound for the stability of nonsmooth DEQs, a critical step for guaranteeing robustness in their training and deployment.

8 Robust Lipschitzian Stability and Exact Coderivative Modulus of Semialgebraic Deep Equilibrium Networks

We now employ the full power of Mordukhovich's coderivative calculus and the associated stability framework for parametric variational systems (Mordukhovich, 2005) to analyze the robust stability of semialgebraic Deep Equilibrium Networks (DEQs). Our analysis provides an *exact* characterization of the solution map's coderivative, from which we derive a verifiable, explicit bound on its Lipschitz modulus.

We consider the DEQ's fixed-point equation, where the parameters are $x := (W, b) \in X := \mathbb{R}^{m\times m} \times \mathbb{R}^m$ and the solution variable is $y := z \in Y := \mathbb{R}^m$. The equilibrium is defined by the generalized equation (GE):

$$0 \in z - \sigma(Wz + b) \tag{70}$$

Here, $\sigma : \mathbb{R}^m \to \mathbb{R}^m$ is a locally Lipschitz and semialgebraic function. This semialgebraic property is the key technical foundation, ensuring all constituent sets and mappings are "tame" and that their generalized derivatives are computable and semialgebraic.

8.1 Parametric Generalized Equation Formulation

To apply the coderivative stability theorems, we formulate the DEQ as a parametric GE $0 \in f(x, y) + Q(x, y)$, where f is a trivial base mapping and Q encapsulates the full parametric dependence. Let $Z := \mathbb{R}^m$ be the image space. We define the base mapping $f(x, y) := 0 \quad (\forall (x, y) \in X \times Y)$ and the set-valued field $Q(x, y) :=$

$z - \sigma(Wz + b) = \Phi(W, b, z)$. The solution map $S : X \rightrightarrows Y$ is the set of equilibria z for a given parameter $x = (W, b)$, namely

$$S(x) := \{y \in Y \mid 0 \in f(x, y) + Q(x, y)\} = Q(x, \cdot)^{-1}(0). \tag{71}$$

This $S(x)$ is precisely the graph of the DEQ's implicit layer. The mapping f is C^∞, and $Q(x, y) = \Phi(x, y)$ is a single-valued, locally Lipschitz, and semialgebraic mapping. Consequently, Q is sequentially normally compact (SNC) at every point of its graph, fulfilling the technical requirements of the stability framework in (Mordukhovich, 2005, Theorem 4.4).

8.2 Exact Coderivative Modulus and Stability Bound

The local Lipschitz-like property (Aubin property) of S at an equilibrium point $(\bar{x}, \bar{y}) = ((\bar{W}, \bar{b}), \bar{z}) \in \operatorname{gph} S$ is equivalent to the coderivative criterion $D^\star S(\bar{x}|\bar{y})(0) = \{0\}$. The exact Lipschitz modulus is the norm of this coderivative multifunction (Rockafellar and Wets, 2009, Theorem 9.40):

$$\operatorname{Lip} S(\bar{x}, \bar{y}) = \|D^\star S(\bar{x}|\bar{y})\| := \sup_{\|y^\star\| \le 1} \sup\left\{\|x^\star\| \mid x^\star \in D^\star S(\bar{x}|\bar{y})(y^\star)\right\}. \tag{72}$$

Our main result provides the exact formula for $D^\star S$ and a computable bound on its norm.

Theorem 15 (Exact Coderivative and Lipschitz Modulus of Semialgebraic DEQs)**.** *Let S be the solution map for the DEQ Eq. (70) with a semialgebraic, L_σ-Lipschitz activation σ. Let $(\bar{x}, \bar{y}) = ((\bar{W}, \bar{b}), \bar{z})$ be an equilibrium point in* $\operatorname{gph} S$*. Let $\bar{u} = \bar{W}\bar{z} + \bar{b}$.*

(i) The limiting coderivative $D^\star S(\bar{x}|\bar{y}) : \mathbb{R}^m \rightrightarrows \mathbb{R}^{m\times m} \times \mathbb{R}^m$ is given by

$$D^\star S(\bar{x}|\bar{y})(y^\star) = \left\{\left(-u^\star \bar{z}^\top, -u^\star\right) \mid \exists z^\star \in \mathbb{R}^m \text{ s.t. } z^\star = \bar{W}^\top u^\star - y^\star, u^\star \in \partial\langle z^\star, \sigma\rangle(\bar{u})\right\} \tag{73}$$

where $\partial\langle z^\star, \sigma\rangle(\bar{u})$ is the limiting subdifferential of the scalarized function $u \mapsto \langle z^\star, \sigma(u)\rangle$ at $\bar{u}$.

(ii) If the contraction condition $L_\sigma\|\bar{W}\|_{\mathrm{op}} = \rho < 1$ holds, the map S is locally Lipschitz-like at $(\bar{x}, \bar{y})$ and its Lipschitz modulus is bounded by

$$\operatorname{Lip} S(\bar{W}, \bar{b}|\bar{z}) \le \frac{L_\sigma\sqrt{1 + \|\bar{z}\|_2^2}}{1 - L_\sigma\|\bar{W}\|_{\mathrm{op}}}. \tag{74}$$

Proof. We apply Theorem 12 to our formulation $f = 0$, $Q = \Phi$. The stability analysis hinges on computing the coderivative $D_N^* Q(\bar{x}, \bar{y}, \bar{v})(z^*)$ where $\bar{v} = -f(\bar{x}, \bar{y}) = 0$. Since $Q = \Phi$ is locally Lipschitz, its coderivative is given by the scalarization (Rockafellar and Wets, 2009, Theorem 9.13):

$$D_N^\star Q(\bar{x}, \bar{y}|0)(z^\star) = \partial\psi(\bar{x}, \bar{y}) \tag{75}$$

where ∂ is the limiting subdifferential and $\psi(x, y) = \langle z^\star, Q(x, y)\rangle$.

Step 1: Compute the Subdifferential $\partial\psi(\bar{x}, \bar{y})$. The scalarized function is $\psi(x, y) = \psi(W, b, z) = \langle z^\star, z - \sigma(g(W, b, z))\rangle$, where $g(W, b, z) = Wz + b$ is a smooth (polynomial) map. We decompose ψ and apply the sum and chain rules for limiting subdifferentials (Rockafellar and Wets, 2009, Theorem 10.6, Theorem 10.49):

$$\psi(W,b,z) = \underbrace{\langle z^\star, z\rangle}_{\psi_1(z)} - \underbrace{\langle z^\star, \sigma(g(W,b,z))\rangle}_{\psi_2(W,b,z)}$$

$$\partial\psi(\bar{x},\bar{y}) = \nabla\psi_1(\bar{y}) + \partial(-\psi_2)(\bar{x},\bar{y}) = \nabla\psi_1(\bar{y}) - \partial\psi_2(\bar{x},\bar{y}).$$

The chain rule for $\psi_2 = (\langle z^\star, \sigma\rangle \circ g)$ holds with equality as g is smooth and $\langle z^\star, \sigma\rangle$ is locally Lipschitz:

$$\partial\psi_2(\bar{x},\bar{y}) = \nabla g(\bar{x},\bar{y})^\star \partial\langle z^\star, \sigma\rangle(\bar{u}) \tag{76}$$

where $\bar{u} = g(\bar{W},\bar{b},\bar{z}) = \bar{W}\bar{z} + \bar{b}$. First, $\nabla\psi_1(\bar{y}) = (0,0,z^\star)$ as a partial gradient w.r.t. $x = (W,b)$ and $y = z$. Second, we compute the adjoint operator $\nabla g(\bar{x},\bar{y})^\star : Z^\star \to X^\star \times Y^\star$ (i.e., $\mathbb{R}^m \to (\mathbb{R}^{m\times m} \times \mathbb{R}^m) \times \mathbb{R}^m$). For any $u^\star \in \mathbb{R}^m$ and any tangent vector $(\Delta W, \Delta b, \Delta z) \in X \times Y$:

$$\begin{aligned}
\langle \nabla g(\bar{x},\bar{y})^\star u^\star, (\Delta W, \Delta b, \Delta z)\rangle_{X\times Y} &= \langle u^\star, \nabla g(\bar{x},\bar{y})[\Delta W, \Delta b, \Delta z]\rangle_Z \\
&= \langle u^\star, (\Delta W)\bar{z} + \bar{W}(\Delta z) + \Delta b\rangle \\
&= \langle u^\star, (\Delta W)\bar{z}\rangle + \langle u^\star, \bar{W}(\Delta z)\rangle + \langle u^\star, \Delta b\rangle \\
&= \mathrm{Tr}(\bar{z}^\top (\Delta W)^\top u^\star) + \langle (\bar{W})^\top u^\star, \Delta z\rangle + \langle u^\star, \Delta b\rangle \\
&= \mathrm{Tr}((u^\star \bar{z}^\top)^\top \Delta W) + \langle u^\star, \Delta b\rangle + \langle \bar{W}^\top u^\star, \Delta z\rangle \\
&= \left\langle \begin{bmatrix} u^\star \bar{z}^\top \\ u^\star \\ \bar{W}^\top u^\star \end{bmatrix}, \begin{bmatrix} \Delta W \\ \Delta b \\ \Delta z \end{bmatrix} \right\rangle.
\end{aligned}$$

By identification of the components, $\nabla g(\bar{x},\bar{y})^\star u^\star = (u^\star \bar{z}^\top, u^\star, \bar{W}^\top u^\star)$. Substituting these components back, any element $(x^\star, y^\star_{part}) = ((W^\star, b^\star), z^\star_{part}) \in \partial\psi(\bar{x},\bar{y})$ is of the form:

$$\begin{bmatrix} W^\star \\ b^\star \\ z^\star_{part} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ z^\star \end{bmatrix} - \begin{bmatrix} u^\star \bar{z}^\top \\ u^\star \\ \bar{W}^\top u^\star \end{bmatrix} = \begin{bmatrix} -u^\star \bar{z}^\top \\ -u^\star \\ z^\star - \bar{W}^\top u^\star \end{bmatrix}, \quad \text{for some } u^\star \in \partial\langle z^\star, \sigma\rangle(\bar{u}). \tag{77}$$

Step 2: Characterize Coderivative and Bound Modulus. From Theorem 12, $x^\star \in D^\star S(\bar{x}|\bar{y})(y^\star)$ if and only if $(x^\star, -y^\star) \in \partial\langle z^\star, f\rangle + D^*_N Q(z^\star)$ for some $z^\star$. With $f = 0$ and $Q = \Phi$, this is

$$((W^\star, b^\star), -y^\star) \in \partial\psi(\bar{W},\bar{b},\bar{z}).$$

Using our result from Eq. (77), we set $x^\star = (W^\star, b^\star)$ and $y^\star_{part} = -y^\star$. This yields the system consisting of $W^\star = -u^\star \bar{z}^\top$, $b^\star = -u^\star$, $-y^\star = z^\star - \bar{W}^\top u^\star \implies z^\star = \bar{W}^\top u^\star - y^\star$, and $u^\star \in \partial\langle z^\star, \sigma\rangle(\bar{u})$. This proves part (i) of the theorem, giving the exact characterization of the coderivative fiber.

For part (ii), we first check that $(x^\star, 0) \in \partial\psi \implies x^\star = 0, z^\star = 0$. This corresponds to setting $y^\star = 0$ in our system, $z^\star = \bar{W}^\top u^\star$ and $u^\star \in \partial\langle z^\star, \sigma\rangle(\bar{u})$. Since σ is L_σ-Lipschitz, $\partial\langle z^\star, \sigma\rangle(\bar{u})$ is norm-bounded (Rockafellar and Wets, 2009, Prop. 9.12), so $\|u^\star\| \le L_\sigma \|z^\star\|$. Substituting $z^\star$ in, we have that

$$\|u^\star\| \le L_\sigma \|\bar{W}^\top u^\star\| \le L_\sigma \|\bar{W}^\top\|_{\mathrm{op}} \|u^\star\| = (L_\sigma \|\bar{W}\|_{\mathrm{op}}) \|u^\star\|.$$

Put $\rho = L_\sigma \|\bar{W}\|_{\text{op}}$. This gives $\|u^\star\|(1-\rho) \le 0$. By hypothesis, $\rho < 1$, so $1-\rho > 0$. This forces $\|u^\star\| \le 0$, which implies $u^\star = 0$. If $u^\star = 0$, then $W^\star = 0$, $b^\star = 0$, and $z^\star = 0$. The qualification condition holds, so S is Lipschitz-like. Finally, we compute the Lipschitz modulus bound. We must bound $\|x^\star\| = \|(W^\star, b^\star)\|$ over all $u^\star, z^\star$ generated by $\|y^\star\| \le 1$. We use the norm $\|(W,b)\|_{\text{op}}^2 := \|W\|_F^2 + \|b\|_2^2$ to arrive at

$$\|(W^\star, b^\star)\|_{\text{op}}^2 = \|-u^\star \bar{z}^\top\|_F^2 + \|-u^\star\|_2^2 = (\|u^\star\|_2 \|\bar{z}\|_2)^2 + \|u^\star\|_2^2 = \|u^\star\|_2^2(1+\|\bar{z}\|_2^2).$$

We then need to bound $\|u^\star\|_2$. From the system $z^\star = \bar{W}^\top u^\star - y^\star$ and $\|u^\star\| \le L_\sigma \|z^\star\|$,

$$\begin{aligned}
\|u^\star\|_2 &\le L_\sigma \|z^\star\|_2 = L_\sigma \|\bar{W}^\top u^\star - y^\star\|_2 \\
&\le L_\sigma \left(\|\bar{W}^\top u^\star\|_2 + \|y^\star\|_2\right) \quad \text{(by triangle inequality)} \\
&\le L_\sigma \left(\|\bar{W}^\top\|_{\text{op}} \|u^\star\|_2 + \|y^\star\|_2\right).
\end{aligned}$$

Rearranging the inequality,

$$\|u^\star\|_2 (1 - L_\sigma \|\bar{W}\|_{\text{op}}) \le L_\sigma \|y^\star\|_2.$$

Since $\rho = L_\sigma \|\bar{W}\|_{\text{op}} < 1$,

$$\|u^\star\|_2 \le \frac{L_\sigma}{1-\rho} \|y^\star\|_2.$$

Substitute this into the norm expression for $(W^\star, b^\star)$,

$$\|(W^\star, b^\star)\|_{\text{op}} = \|u^\star\|_2 \sqrt{1+\|\bar{z}\|_2^2} \le \left(\frac{L_\sigma \|y^\star\|_2}{1-\rho}\right) \sqrt{1+\|\bar{z}\|_2^2}.$$

The Lipschitz modulus is the supremum of this quantity over the set $\|y^\star\|_2 \le 1$:

$$\operatorname{Lip} S(\bar{x}, \bar{y}) = \sup_{\|y^\star\| \le 1} \|(W^\star, b^\star)\|_{\text{op}} \le \frac{L_\sigma \sqrt{1+\|\bar{z}\|_2^2}}{1-\rho}.$$

This completes the proof. □

Theorem 16 (Finiteness and Computability of the Lipschitz Modulus). *Let $\sigma : \mathbb{R}^m \to \mathbb{R}^m$ be a semialgebraic function, and let S be the DEQ solution map. At any equilibrium point $(\bar{W}, \bar{b}, \bar{z})$ where S is Lipschitz-like (e.g., under the conditions of Theorem 15), the exact Lipschitz modulus* $\operatorname{Lip} S(\bar{W}, \bar{b} \mid \bar{z})$ *is a finite, semialgebraic function of the equilibrium point and is (in theory) computable via Cylindrical Algebraic Decomposition (CAD).*

Proof. This proof relies on the Tarski-Seidenberg theorem, which states that any set defined by a first-order formula over the real numbers (using polynomial equalities/inequalities, logical operators, and quantifiers $\exists, \forall$) is a semialgebraic set.

1. Semialgebraicity of the Coderivative Graph: The graph of the solution map, $\operatorname{gph} S = \{(W, b, z) \mid z - \sigma(Wz+b) = 0\}$, is semialgebraic because σ is semialgebraic and $Wz+b$ is polynomial. The limiting normal cone $N((\bar{W}, \bar{b}, \bar{z}); \operatorname{gph} S)$ to a semialgebraic set is also a semialgebraic set. The graph of the coderivative, $\operatorname{gph} D^\star S = \{(y^\star, W^\star, b^\star) \mid (W^\star, b^\star, -y^\star) \in N((\bar{W}, \bar{b}, \bar{z}); \operatorname{gph} S)\}$, is a linear projection of the normal cone's graph and is therefore semialgebraic by Tarski-Seidenberg.

2. Semialgebraicity of the Norm Function: Let $K = \{y^\star \in \mathbb{R}^m \mid \|y^\star\|_2^2 - 1 = 0\}$, a compact semialgebraic set. Define the function $f : K \to \mathbb{R}$ as

$$f(y^\star) := \sup\Big\{\|(W^\star, b^\star)\|_{\mathrm{op}} \mid (W^\star, b^\star) \in D^\star S(\bar{W}, \bar{b} \mid \bar{z})(y^\star)\Big\}.$$

(Note: The sup and inf in the definition of the modulus norm are equivalent here because the set $D^\star S(\cdot)(y^\star)$ is a compact set, as shown in the proof of Theorem 15). The operator norm $\|(W^\star, b^\star)\|_{\mathrm{op}}$ is a semialgebraic function. The supremum of a semialgebraic function over the fibers of a semialgebraic multifunction ($D^\star S$) is a semialgebraic function. Thus, $f(y^\star)$ is semialgebraic.

3. Finiteness: The Lipschitz modulus is $\operatorname{Lip} S = \sup_{y^\star \in K} f(y^\star)$. This is the supremum of the semialgebraic function f over the compact semialgebraic set K. A continuous semialgebraic function on a compact semialgebraic set must achieve its maximum. The function f is continuous (upper-semicontinuous by definition of the outer limit, and lower-semicontinuous by the compactness of fibers), so its maximum is achieved and finite. Theorem 15 provides an explicit upper bound, confirming finiteness.

4. Computability: The value $\ell = \operatorname{Lip} S$ is the solution to an optimization problem that can be expressed as a first-order formula:

$$\ell = \sup\Big\{r \mid \exists y^\star, W^\star, b^\star \text{ s.t. } (\|y^\star\|_2^2 = 1) \wedge ((y^\star, W^\star, b^\star) \in \operatorname{gph} D^\star S) \wedge (\|(W^\star, b^\star)\|_{\mathrm{op}} = r)\Big\}.$$

All sets and functions in this formula ($\operatorname{gph} D^\star S$, norms) are semialgebraic, defined by a finite number of polynomial equalities/inequalities. The Tarski-Seidenberg theorem guarantees that this quantified formula can be eliminated, resulting in a quantifier-free semialgebraic description of ℓ (e.g., ℓ as the root of a specific polynomial). Algorithms such as Cylindrical Algebraic Decomposition (CAD) provide a constructive, albeit doubly-exponential, method for this elimination, proving that ℓ is, in principle, computable.

□

Applications to loss landscape analysis. This framework provides a rigorous tool for analyzing the stability of implicit differentiation. For a composite loss function $g(W, b) = \ell(z(W, b))$ where ℓ is a smooth semialgebraic function (e.g., squared error), the chain rule for coderivatives holds. The (transposed) gradient of the loss is related to the coderivative of the solution map via $D^\star g(\bar{W}, \bar{b}|\bar{g})(w^\star) = D^\star z(\bar{W}, \bar{b}|\bar{z})(\nabla \ell(\bar{z})^\star w^\star)$. The Lipschitz constant of the gradient ∇g (which bounds the Hessian) can be directly related to $\operatorname{Lip} S$. A stable S (small $\operatorname{Lip} S$) is crucial for ensuring a well-behaved loss landscape with bounded curvature, which is essential for the stability and convergence of gradient-based training algorithms.

9 Conclusion and Outlook

In this work, we provided a rigorous analysis of the Lipschitz stability of solution mappings for parametric nonsmooth optimization problems. We first highlighted the limitations of the conservative Jacobian framework, which can produce loose or intractable bounds for structured problems like the LASSO. This motivated the adoption of the more powerful and geometrically precise tools of variational analysis and coderivative calculus. By applying this framework, we derived a sharp stability bound for the LASSO and, as our primary contribution, extended this analysis to nonsmooth, semialgebraic Deep Equilibrium Networks (DEQs). Our key result establishes verifiable, contraction-based conditions under which the DEQ's solution map is locally Lipschitz-like. Crucially, by leveraging the semialgebraic structure of the network, we proved that the exact Lipschitz modulus is finite and theoretically computable. This provides a rigorous mathematical foundation for the stability of implicit auto-differentiation, a cornerstone of training modern implicit neural networks.

For future work, a significant direction is to bridge the gap between theoretical computability via CAD and practical estimation. The semialgebraic formulation of the Lipschitz modulus opens the door to developing tractable algorithms, such as sum-of-squares (SOS) programming or semidefinite relaxations, to efficiently compute tight upper bounds on this modulus. These bounds could then be used not only for certification but also as novel, powerful regularizers during training to enforce stability. Finally, the methodology presented here is broadly applicable and can be extended to analyze the robustness of other complex implicit layers, such as those defined by optimization problems, variational inequalities, or differential inclusions, further solidifying the foundations of robust implicit deep learning.

References

J. Bolte and E. Pauwels. Conservative gradients for nonsmooth optimization. *Mathematical Programming*, 188:1–39, 2021.

J. Bolte, T. Le, E. Pauwels, and A. Silveti-Falls. Nonsmooth implicit differentiation for machine learning and optimization. *Advances in Neural Information Processing Systems*, 2021.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

N. Cesa-Bianchi, O. Dekel, and O. Shamir. Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems 26*, 2013.

F. R. K. Chung and P. D. Seymour. Graphs with small bandwidth and cutwidth. *Discrete Mathematics*, 75(1–3):113–119, 1989.

F. H. Clarke. *Optimization and Nonsmooth Analysis*. Society for Industrial and Applied Mathematics (SIAM), 1990.

P. L. Combettes and J. C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, 2011.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2006.

D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20:119–144, 2020.

O. Dekel and E. Hazan. Better rates for any adversarial deterministic MDP. In *Proceedings of the Thirtieth International Conference on Machine Learning*, 2013.

M. Fazlyab, A. Robey, H. Hassani, M. Morari, and G. J. Pappas. Efficient and accurate estimation of Lipschitz constants for deep neural networks. *Advances in Neural Information Processing Systems*, 2019.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

S. Geulen, B. Vöcking, and M. Winkler. Regret minimization for online buffering problems using the weighted majority algorithm. In *Proceedings of the 23rd International Conference on Learning Theory*, pages 132–143, 2010.

H. Gouk, E. Frank, B. Pfahringer, and M. J. Cree. A new method for estimating the Lipschitz constant of a deep neural network. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10):5510–5521, 2021.

A. Gyorgy and G. Neu. Near-optimal rates for limited-delay universal lossy source coding. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2218–2222, 2011.

M. Hein and M. Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. *Advances in Neural Information Processing Systems*, 2017.

R. A. Horn and C. R. Johnson. *Matrix Analysis (2nd ed.)*. Cambridge University Press, 2012.

A. Kalai and S. Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71:291–307, 2005.

F. Latorre, P. M. G. Rüue, and A. Bar-Noy. Lipschitz-constrained neural networks: a new approach. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

A. S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2002.

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *International Conference on Learning Representations (ICLR)*, 2018.

B. S. Mordukhovich. Coderivative calculus and robust Lipschitzian stability for variational systems. *Journal of Convex Analysis*, 12(1):79–116, 2005.

B. S. Mordukhovich. New applications of variational analysis to optimization and control. In *System Modeling and Optimization*, pages 100–128. Springer, Berlin, Heidelberg, 2009.

B. S. Mordukhovich. *Variational Analysis and Applications*. Springer, Berlin, 2018.

A. Raghunathan, J. Steinhardt, and P. Liang. Certified defenses against adversarial examples. *International Conference on Learning Representations (ICLR)*, 2018.

R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer, Berlin, 2009.

Y. Tsuzuku, I. Sato, and M. Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Advances in Neural Information Processing Systems*, 2018.

A. Virmaux and K. Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 2018.

L. Weng, H. Zhang, H. Chen, C. Hsieh, D. Boning, and L. Daniel. Towards fast computation of certified robustness for neural networks. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

A. Yao. Probabilistic computations: Toward a unified measure of complexity. In *Proceedings of the 18th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 222–227, 1977.

J. Y. Yu, S. Mannor, and N. Shimkin. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009.