Differentiating Minimal-Norm Solutions to Parametric Problems with Many Solutions

Baptiste Plaquevent-Jourdain,
Jalal Fadili, Antonio Silveti-Falls

October, 29th 2026

Outline

1. Introduction
2. Implicit Framework
3. At the Limit?
4. Applications

Problem Statement

Parametric / bilevel optimization

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

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

- hyperparameter tuning [Ped16; LVD19; Ber+20]
- dictionary learning [PF11; MBP11],
- meta-learning [Fra+18; Eng+25],
- data cleaning/poisoning/... [Wan+18; Blo+21],
- model selection in statistics [Vai+15; Vai+17],
- deep learning [WK20].

First-order Methods Ideally

$$\min_{\theta \in \Theta} \mathcal{L}(x^\star(\theta))$$
$$\text{s.t.} \quad x^\star(\theta) \in \mathcal{S}(\theta) := \arg\min_{x \in \mathbb{R}^n} f(x, \theta).$$

- well-defined: $\mathcal{S}(\theta) = \{x^\star(\theta)\}$;
- for $x^\star(\theta)$: Optimality Conditions (**OC**) $\nabla_x f(x^\star(\theta), \theta) = 0$;
- $\theta^+ = \theta - \delta \nabla(\mathcal{L} \circ x^\star)(\theta)$; chain rule $\nabla_\theta(\mathcal{L} \circ x^\star) = \nabla_\theta x^\star \nabla_x \mathcal{L}$;
- IFT on the **OC**: system to solve
  $\nabla_\theta x^\star(\theta) \nabla^2_{xx} f(x^\star(\theta), \theta) + \nabla^2_{x\theta} f(x^\star(\theta), \theta) = 0$,

First-order Methods Ideally

$$\min_{\theta \in \Theta} \mathcal{L}(x^\star(\theta))$$
$$\text{s.t.} \quad x^\star(\theta) \in \mathcal{S}(\theta) := \arg\min_{x \in \mathbb{R}^n} f(x, \theta).$$

- well-defined: $\mathcal{S}(\theta) = \{x^\star(\theta)\}$;
- for $x^\star(\theta)$: Optimality Conditions (**OC**) $\nabla_x f(x^\star(\theta), \theta) = 0$;
- $\theta^+ = \theta - \delta \nabla(\mathcal{L} \circ x^\star)(\theta)$; chain rule $\nabla_\theta(\mathcal{L} \circ x^\star) = \nabla_\theta x^\star \nabla_x \mathcal{L}$;
- IFT on the **OC**: system to solve
  $\nabla_\theta x^\star(\theta) \nabla^2_{xx} f(x^\star(\theta), \theta) + \nabla^2_{x\theta} f(x^\star(\theta), \theta) = 0$,

First-order Methods Ideally

$$\min_{\theta \in \Theta} \mathcal{L}(x^\star(\theta))$$
$$\text{s.t.} \quad x^\star(\theta) \in \mathcal{S}(\theta) := \arg\min_{x \in \mathbb{R}^n} f(x, \theta).$$

- well-defined: $\mathcal{S}(\theta) = \{x^\star(\theta)\}$; Why unique?
- for $x^\star(\theta)$: Optimality Conditions (**OC**) $\nabla_x f(x^\star(\theta), \theta) = 0$;
- $\theta^+ = \theta - \delta \nabla(\mathcal{L} \circ x^\star)(\theta)$; chain rule $\nabla_\theta(\mathcal{L} \circ x^\star) = \nabla_\theta x^\star \nabla_x \mathcal{L}$;
- IFT on the **OC**: system to solve Why smooth?
  $\nabla_\theta x^\star(\theta) \nabla^2_{xx} f(x^\star(\theta), \theta) + \nabla^2_{x\theta} f(x^\star(\theta), \theta) = 0$,

Outline

1 Introduction

2 Implicit Framework

3 At the Limit?

4 Applications

Smooth Implicit Function Theorem

$$\begin{cases} F(x, y) = x^2 + y^2 - 1, \\ F(\hat{x}, \hat{y}) = 0. \end{cases}$$

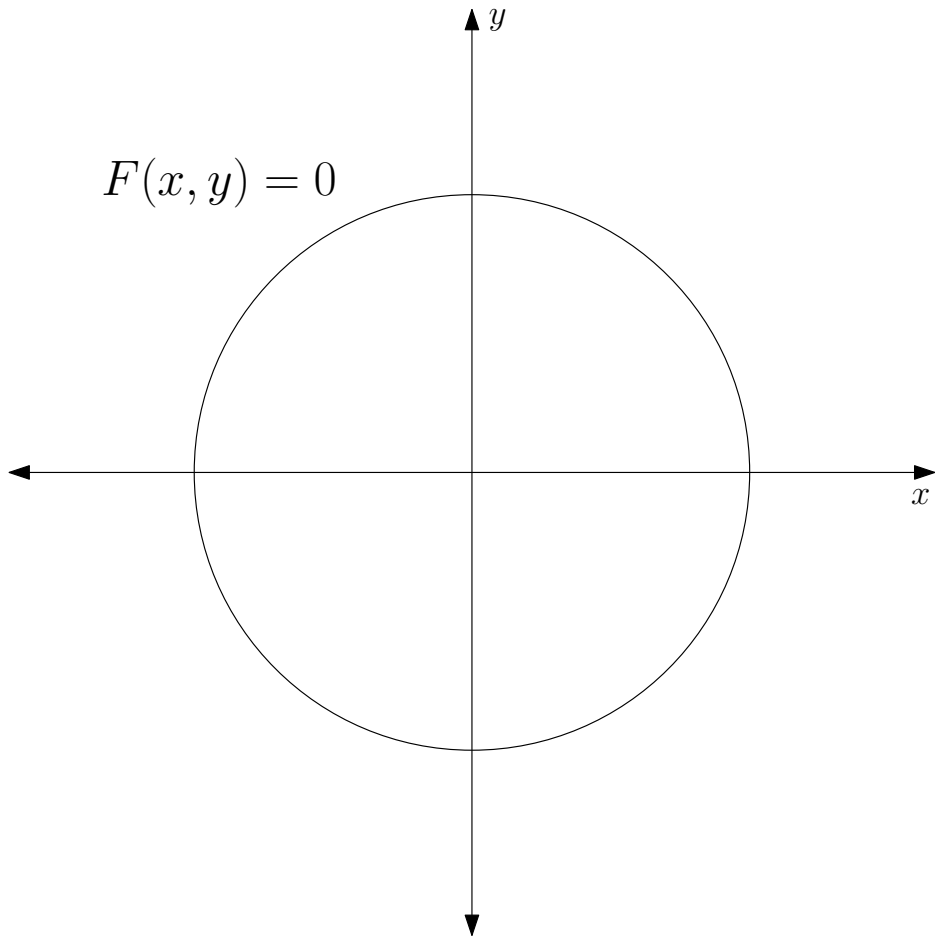
- Can we find $y = G(x)$ so that $F(x, G(x)) = 0$?

Existence

- Can we compute the gradient of G ?

Calculus

If $\frac{\partial F}{\partial y}(\hat{x}, \hat{y})$ nonsingular.



Smooth Implicit Function Theorem

$$\begin{cases} F(x,y) = x^2 + y^2 - 1, \\ F(\hat{x},\hat{y}) = 0. \end{cases}$$

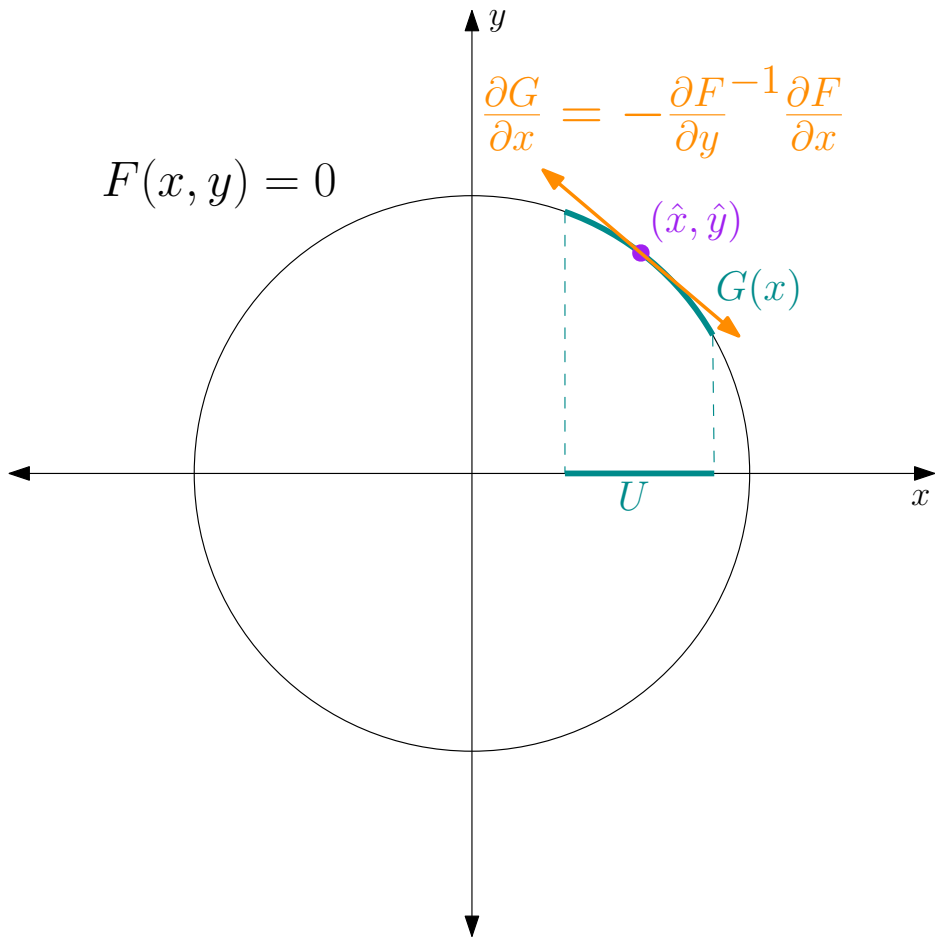
- Can we find $y = G(x)$ so that $F(x, G(x)) = 0$?

Existence

- Can we compute the gradient of G ?

Calculus

If $\frac{\partial F}{\partial y}(\hat{x},\hat{y})$ nonsingular.



Simple Example

Easy Least-Squares

(lower pb) $\min_x \|Ax - \theta\|_2^2/2, \quad \mathcal{N}(A) = \{0\} \Leftrightarrow A^\mathsf{T} A \succ 0$

OC: $F(x, \theta) = A^\mathsf{T}(Ax - \theta) = 0.$

Simple Example

Easy Least-Squares

(lower pb) $\min_x \|Ax - \theta\|_2^2/2, \quad \mathcal{N}(A) = \{0\} \Leftrightarrow A^\mathsf{T}A \succ 0$

OC: $F(x, \theta) = A^\mathsf{T}(Ax - \theta) = 0.$

direct $x^\star(\theta) = (A^\mathsf{T}A)^{-1}A^\mathsf{T}\theta \overset{\text{SVD}}{=} A^\dagger\theta, \quad \frac{\partial x^\star}{\partial \theta} = A^\dagger$

Simple Example

Easy Least-Squares

(lower pb) $\min_x \|Ax - \theta\|_2^2/2, \quad \mathcal{N}(A) = \{0\} \Leftrightarrow A^\mathsf{T}A \succ 0$

OC: $F(x, \theta) = A^\mathsf{T}(Ax - \theta) = 0.$

direct $x^\star(\theta) = (A^\mathsf{T}A)^{-1}A^\mathsf{T}\theta \overset{\text{SVD}}{=} A^\dagger\theta, \quad \frac{\partial x^\star}{\partial \theta} = A^\dagger$

IFT $\frac{\partial F}{\partial x} = A^\mathsf{T}A, \frac{\partial F}{\partial \theta} = -A^\mathsf{T}, \quad \frac{\partial x^\star}{\partial \theta} = -(A^\mathsf{T}A)^{-1}(-A^\mathsf{T}) = A^\dagger.$

$\frac{\partial F}{\partial x} = \frac{\partial^2 f}{\partial^2 x}$ **H**essian, $\frac{\partial F}{\partial \theta} = \frac{\partial^2 f}{\partial x \partial \theta}$ **M**ixed derivative.

Simple Example

Easy Least-Squares

(lower pb) $\min_x \|Ax-\theta\|_2^2/2, \quad \mathcal{N}(A)=\{0\} \Leftrightarrow A^\mathsf{T}A \succ 0$

OC: $F(x,\theta) = A^\mathsf{T}(Ax-\theta) = 0.$

direct $x^\star(\theta) = (A^\mathsf{T}A)^{-1}A^\mathsf{T}\theta \overset{\text{SVD}}{=} A^\dagger\theta, \quad \frac{\partial x^\star}{\partial \theta} = A^\dagger$

IFT $\frac{\partial F}{\partial x} = A^\mathsf{T}A, \frac{\partial F}{\partial \theta} = -A^\mathsf{T}, \quad \frac{\partial x^\star}{\partial \theta} = -(A^\mathsf{T}A)^{-1}(-A^\mathsf{T}) = A^\dagger.$

$\frac{\partial F}{\partial x} = \frac{\partial^2 f}{\partial^2 x}$ **H**essian, $\frac{\partial F}{\partial \theta} = \frac{\partial^2 f}{\partial x \partial \theta}$ **M**ixed derivative.

- Well-posedness: in general, $\mathcal{N}(A) \neq \{0\}$ so $A^\mathsf{T}A$ singular.
- Nonsmooth: some IFTs exist but no ideal derivative formula.

Regularization of Convex Problems

$$x_\varepsilon^\star(\theta) := \arg\min_x f(x, \theta) + \frac{\varepsilon}{2}\|x\|_2^2,$$
$$x_\varepsilon^\star(\theta) : \nabla_x f(x_\varepsilon^\star(\theta), \theta) + \varepsilon x_\varepsilon^\star(\theta) = 0.$$

Regularization of Convex Problems

$$x_\varepsilon^\star(\theta) := \arg\min_x f(x,\theta) + \frac{\varepsilon}{2}\|x\|_2^2,$$
$$x_\varepsilon^\star(\theta) : \nabla_x f(x_\varepsilon^\star(\theta),\theta) + \varepsilon x_\varepsilon^\star(\theta) = 0.$$
$$x_\varepsilon^\star(\theta) \underset{\varepsilon\to 0}{\to} x_{\min}^\star(\theta) := \arg\min\{\|x\|_2 : x \in \mathcal{S}(\theta)\}$$

$x_{\min}^\star$ is a computable element of $\mathcal{S}(\theta)$.

Regularization of Convex Problems

$$x_\varepsilon^\star(\theta) := \arg\min_x f(x,\theta) + \frac{\varepsilon}{2}\|x\|_2^2,$$
$$x_\varepsilon^\star(\theta) : \nabla_x f(x_\varepsilon^\star(\theta),\theta) + \varepsilon x_\varepsilon^\star(\theta) = 0.$$
$$x_\varepsilon^\star(\theta) \underset{\varepsilon\to 0}{\to} x_{\min}^\star(\theta) := \arg\min\{\|x\|_2 : x \in \mathcal{S}(\theta)\}$$

$x_{\min}^\star$ is a computable element of $\mathcal{S}(\theta)$.

Presentation on Least-Squares

(lower pb)	$\min_x \|Ax-\theta\|_2^2/2 + \frac{\varepsilon}{2}\|x\|_2^2$
solution	$x_\varepsilon^\star(\theta) = (A^\mathsf{T}A + \varepsilon I)^{-1}A^\mathsf{T}\theta,$
Jacobian	$J_{x_\varepsilon^\star}(\theta) = (A^\mathsf{T}A + \varepsilon I)^{-1}A^\mathsf{T},$
unregularized solutions	$\mathcal{S}(\theta) = A^\dagger\theta + \mathcal{N}(A), x_{\min}^\star(\theta) = A^\dagger\theta$
unregularized Jacobian	$J_{x_{\min}^\star}(\theta) = A^\dagger$

Problematic

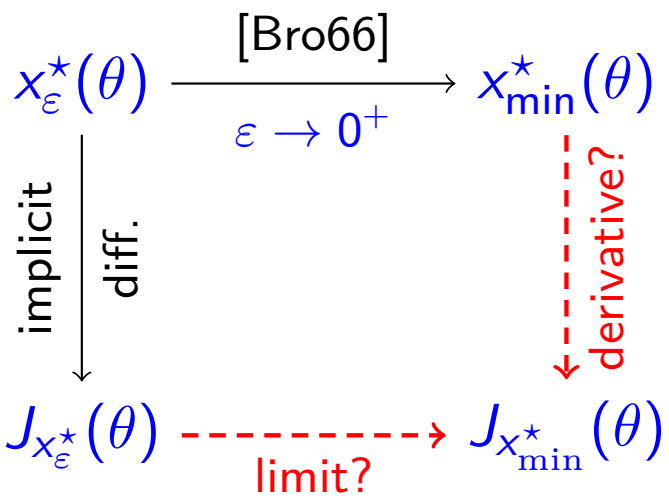
$$x_\varepsilon^\star(\theta) = (A^\mathsf{T}A + \varepsilon I)^{-1}A^\mathsf{T}\theta, \quad J_{x_\varepsilon^\star}(\theta) = (A^\mathsf{T}A + \varepsilon I)^{-1}A^\mathsf{T},$$
$$\mathcal{S}(\theta) = x_{\min}^\star(\theta) + \mathcal{N}(A), \qquad J_{x_{\min}^\star}(\theta) = A^\dagger,$$

$(A^\mathsf{T}A + \varepsilon I)^{-1}$ blows up, BUT $(A^\mathsf{T}A + \varepsilon I)^{-1}A^\mathsf{T} \to A^\dagger$.

Problematic

$$x_\varepsilon^\star(\theta) = (A^\mathsf{T}A + \varepsilon I)^{-1}A^\mathsf{T}\theta, \quad J_{x_\varepsilon^\star}(\theta) = (A^\mathsf{T}A + \varepsilon I)^{-1}A^\mathsf{T},$$
$$\mathcal{S}(\theta) = x_{\min}^\star(\theta) + \mathcal{N}(A), \qquad J_{x_{\min}^\star}(\theta) = A^\dagger,$$

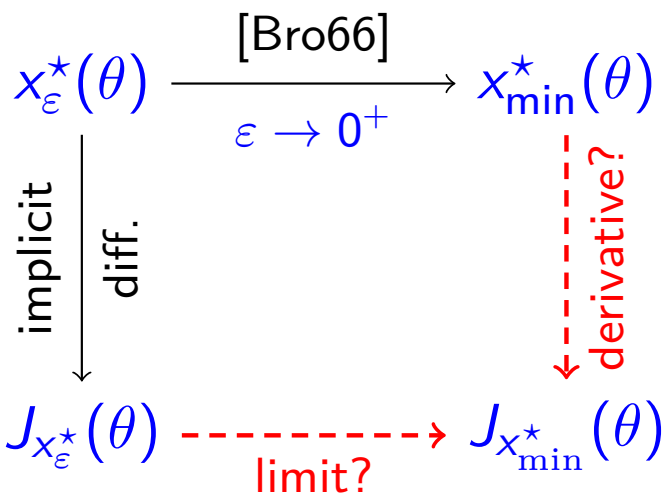
$(A^\mathsf{T}A + \varepsilon I)^{-1}$ blows up, BUT $(A^\mathsf{T}A + \varepsilon I)^{-1}A^\mathsf{T} \to A^\dagger$.



Problematic

$$x_\varepsilon^\star(\theta) = (A^\mathsf{T}A + \varepsilon I)^{-1}A^\mathsf{T}\theta, \quad J_{x_\varepsilon^\star}(\theta) = (A^\mathsf{T}A + \varepsilon I)^{-1}A^\mathsf{T},$$
$$\mathcal{S}(\theta) = x_{\min}^\star(\theta) + \mathcal{N}(A), \qquad J_{x_{\min}^\star}(\theta) = A^\dagger,$$

$(A^\mathsf{T}A + \varepsilon I)^{-1}$ blows up, BUT $(A^\mathsf{T}A + \varepsilon I)^{-1}A^\mathsf{T} \to A^\dagger$.



Huber/LASSO/... not $\mathcal{C}^2$. $\exists$ nonsmooth IFTs, but no formula.

Conservativeness

Answer to nonsmoothness: conservative mappings [BP20]

- a new sort of generalized derivatives;
- calculus rules like sum, product, composition: suited for autodiff / backpropagation;
- subgradient-type methods/convergence;
- IFT with formula [Bol+21], strongly convex case [BPS24];
- issue: not uniquely defined, $\exists \infty$ conservative mappings.
- see also [DD22; LT21; BHS20; BPV22; GPS24].

Conservativeness

Answer to nonsmoothness: conservative mappings [BP20]

- a new sort of generalized derivatives;
- calculus rules like sum, product, composition: suited for autodiff / backpropagation;
- subgradient-type methods/convergence;
- IFT with formula [Bol+21], strongly convex case [BPS24];
- issue: not uniquely defined, $\exists \infty$ conservative mappings.
- see also [DD22; LT21; BHS20; BPV22; GPS24].

Conservativeness

Answer to nonsmoothness: conservative mappings [BP20]

- a new sort of generalized derivatives;
- calculus rules like sum, product, composition: suited for autodiff / backpropagation;
- subgradient-type methods/convergence;
- IFT with formula [Bol+21], strongly convex case [BPS24];
- issue: not uniquely defined, $\exists \infty$ conservative mappings.
- see also [DD22; LT21; BHS20; BPV22; GPS24].

Definability subtleties are skipped.

Outline

1 Introduction

2 Implicit Framework

3 At the Limit?

4 Applications

Pseudoinverse Option

$$x_\varepsilon^\star(\theta) = (A^\mathsf{T}A + \varepsilon I)^{-1}A^\mathsf{T}\theta, \quad J_{x_\varepsilon^\star}(\theta) = (A^\mathsf{T}A + \varepsilon I)^{-1}A^\mathsf{T},$$
$$\mathcal{S}(\theta) = x_{\min}^\star(\theta) + \mathcal{N}(A), \qquad J_{x_{\min}^\star}(\theta) = A^\dagger,$$

Observe: $A^\dagger = -(A^\mathsf{T}A)^\dagger(-A^\mathsf{T}) = -(\text{Hessian})^\dagger\text{Mixed}$, $(-H^{-1}M)$.

Pseudoinverse Option

$$x_\varepsilon^\star(\theta) = (A^\mathsf{T}A + \varepsilon I)^{-1}A^\mathsf{T}\theta, \quad J_{x_\varepsilon^\star}(\theta) = (A^\mathsf{T}A + \varepsilon I)^{-1}A^\mathsf{T},$$
$$\mathcal{S}(\theta) = x_{\min}^\star(\theta) + \mathcal{N}(A), \qquad J_{x_{\min}^\star}(\theta) = A^\dagger,$$

Observe: $A^\dagger = -(A^\mathsf{T}A)^\dagger(-A^\mathsf{T}) = -(\text{Hessian})^\dagger\text{Mixed}$, $(-H^{-1}M)$.

- **pseudoinverse for quadratic costs** [Vic+22],
- ~~convex~~ case: parametric Morse-Bott [AM22; Bol+26],
- ~~convex~~ case: Polyak-Łojasiewicz assumption, select then differentiate [Mas+26],
- value function [SV26].

Framework

$$x_\varepsilon^\star(\theta) : 0 = \nabla_x f(x_\varepsilon^\star(\theta), \theta) + \varepsilon x_\varepsilon^\star(\theta) = F_\varepsilon(x_\varepsilon^\star(\theta), \theta).$$

Conservative (Nonsmooth) IFT Step

$$J_{F_\varepsilon}(x, \theta) = \left\{ (H + \varepsilon I \;\; M) : (H \;\; M) \in J_{\nabla_x f}(x, \theta) \right\}, \quad \text{at } (x_\varepsilon^\star(\theta), \theta)$$

$$J_{x_\varepsilon^\star}(\theta) = \left\{ -(H_\varepsilon + \varepsilon I)^{-1} M_\varepsilon : (H_\varepsilon \;\; M_\varepsilon) \in J_{\nabla_x f}(x_\varepsilon^\star(\theta), \theta) \right\}.$$

Framework

$$x_\varepsilon^\star(\theta) : 0 = \nabla_x f(x_\varepsilon^\star(\theta), \theta) + \varepsilon x_\varepsilon^\star(\theta) = F_\varepsilon(x_\varepsilon^\star(\theta), \theta).$$

Conservative (Nonsmooth) IFT Step

$$J_{F_\varepsilon}(x, \theta) = \left\{ (H + \varepsilon I \;\; M) : (H \;\; M) \in J_{\nabla_x f}(x, \theta) \right\}, \quad \text{at } (x_\varepsilon^\star(\theta), \theta)$$

$$J_{x_\varepsilon^\star}(\theta) = \left\{ -(H_\varepsilon + \varepsilon I)^{-1} M_\varepsilon : (H_\varepsilon \;\; M_\varepsilon) \in J_{\nabla_x f}(x_\varepsilon^\star(\theta), \theta) \right\}.$$

Convergence?

$$\lim_{\varepsilon \to 0} \underbrace{-(H_\varepsilon + \varepsilon I)^{-1} M_\varepsilon}_{\forall (H_\varepsilon \;\; M_\varepsilon) \in J_{\nabla_x f}(x_\varepsilon^\star(\theta), \theta)} \overset{?}{=} \underbrace{-H_0^\dagger M_0}_{\forall (H_0 \;\; M_0) \in J_{\nabla_x f}(x_{\min}^\star(\theta), \theta)} \overset{?}{\in} J^?_{x_{\min}^\star}(\theta)$$

Schechtman's Results (in our Setting) [Sch26]

Limits of parametrized conservative mappings: ✓ for $J_{x^\star_\varepsilon}$.

$$J_{x^\star_{\min}}(\theta) := \left\{ V : \begin{array}{l} \exists\ V_k \to V, V_k \in J_{x^\star_{\varepsilon_k}}(\theta_k),\ \varepsilon_k \downarrow 0, \\ \theta_k \in O,\ \theta_k \to \theta, x^\star_{\varepsilon_k}(\theta_k) \to x^\star_{\min}(\theta) \end{array} \right\} \quad (1)$$

Schechtman's Results (in our Setting) [Sch26]

Limits of parametrized conservative mappings: ✓ for $J_{x^\star_\varepsilon}$.

$$J_{x^\star_{\min}}(\theta) := \left\{ V : \begin{array}{l} \exists\ V_k \to V, V_k \in J_{x^\star_{\varepsilon_k}}(\theta_k),\ \varepsilon_k \downarrow 0, \\ \theta_k \in O,\ \theta_k \to \theta, x^\star_{\varepsilon_k}(\theta_k) \to x^\star_{\min}(\theta) \end{array} \right\} \quad (1)$$

Theorem ([Sch26, Lemma 4.2, Theorem 4.3 and Corollary 4.5])

If: (i) definability, (ii) $J_{x^\star_\varepsilon}$ cons. map. of $x^\star_\varepsilon$, $\forall\ K \subseteq O$ compact,

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

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

Then $J_{x^\star_{\min}}$ is a conservative mapping of $x^\star_{\min}$.

Requirements for Schechtman's Theorem

- $(i), (ii)$ ✓;
- (iii): uniform continuity of $x^\star(\varepsilon, \theta)$;
- (iv): uniform boundedness on $J_{x_\varepsilon^\star}$; actually $(iv) \Rightarrow (iii)$.

Requirements for Schechtman's Theorem

- $(i), (ii)$ ✓;
- (iii): uniform continuity of $x^\star(\varepsilon, \theta)$;
- (iv): uniform boundedness on $J_{x_\varepsilon^\star}$; actually $(iv) \Rightarrow (iii)$.

Conservative (Nonsmooth) IFT Step

$$J_{x_\varepsilon^\star}(\theta) = \left\{ -\left(H_\varepsilon + \varepsilon I\right)^{-1} M_\varepsilon : \begin{pmatrix} H_\varepsilon & M_\varepsilon \end{pmatrix} \in J_{\nabla_x f}(x_\varepsilon^\star(\theta), \theta) \right\}.$$

We need that: $\forall\ K$ compact, $\exists\ C_K > 0, \varepsilon_K > 0$ s.t.

$$\sup \left\{ \left\| \left(H_\varepsilon + \varepsilon I_n\right)^{-1} M_\varepsilon \right\| : \begin{array}{l} \theta \in K, \quad 0 < \varepsilon \leqslant \varepsilon_K, \\ \begin{pmatrix} H_\varepsilon & M_\varepsilon \end{pmatrix} \in J_F(x_\varepsilon^\star(\theta), \theta) \end{array} \right\} \leqslant C_K.$$

Relaxing Nonsingularity

Goal: bounded derivatives without assuming nonsingularity.

Idea: uniform range assumption (R)

$$(M_\varepsilon \in \operatorname{range}(H_\varepsilon)) \quad M_\varepsilon = H_\varepsilon W_\varepsilon, \quad \|W_\varepsilon\| \leqslant C$$

$$-(H_\varepsilon + \varepsilon I_n)^{-1} M_\varepsilon = -\underbrace{(H_\varepsilon + \varepsilon I_n)^{-1} H_\varepsilon}_{0 \preceq \cdot \preceq I} \underbrace{W_\varepsilon}_{\|\cdot\| \leqslant C}.$$

Least-Squares $M_\varepsilon \in \operatorname{range}(H_\varepsilon)$: $\underbrace{-A^\mathsf{T}}_{=M_\varepsilon} = -\underbrace{(A^\mathsf{T}A)}_{H_\varepsilon} \underbrace{A^\dagger}_{W_\varepsilon \equiv W}$ (SVD)

Nonsingularity: $\forall\ (H_0\ M_0) \in J_F(x^\star_{\min}(\theta), \theta), \det(H_0) \neq 0$ [Bol+21].
Continuity of det: nonsingularity $\Rightarrow$ uniform range; **(R)** can fail.

Relaxing Nonsingularity

Goal: bounded derivatives without assuming nonsingularity.

Idea: uniform range assumption **(R)**

$$(M_\varepsilon \in \operatorname{range}(H_\varepsilon)) \quad M_\varepsilon = H_\varepsilon W_\varepsilon, \quad \|W_\varepsilon\| \leqslant C$$

$$-(H_\varepsilon + \varepsilon I_n)^{-1} M_\varepsilon = -\underbrace{(H_\varepsilon + \varepsilon I_n)^{-1} H_\varepsilon}_{0 \preceq \cdot \preceq I} \underbrace{W_\varepsilon}_{\|\cdot\| \leqslant C}.$$

Least-Squares $M_\varepsilon \in \operatorname{range}(H_\varepsilon)$: $\underbrace{-A^\mathsf{T}}_{=M_\varepsilon} = -(\underbrace{A^\mathsf{T} A}_{H_\varepsilon}) \underbrace{A^\dagger}_{W_\varepsilon \equiv W}$ (SVD)

Nonsingularity: $\forall\ (H_0\ M_0) \in J_F(x^\star_{\min}(\theta), \theta), \det(H_0) \neq 0$ [Bol+21].
Continuity of det: nonsingularity $\Rightarrow$ uniform range; **(R)** can fail.

Relaxing Nonsingularity

Goal: bounded derivatives without assuming nonsingularity.

Idea: uniform range assumption **(R)**

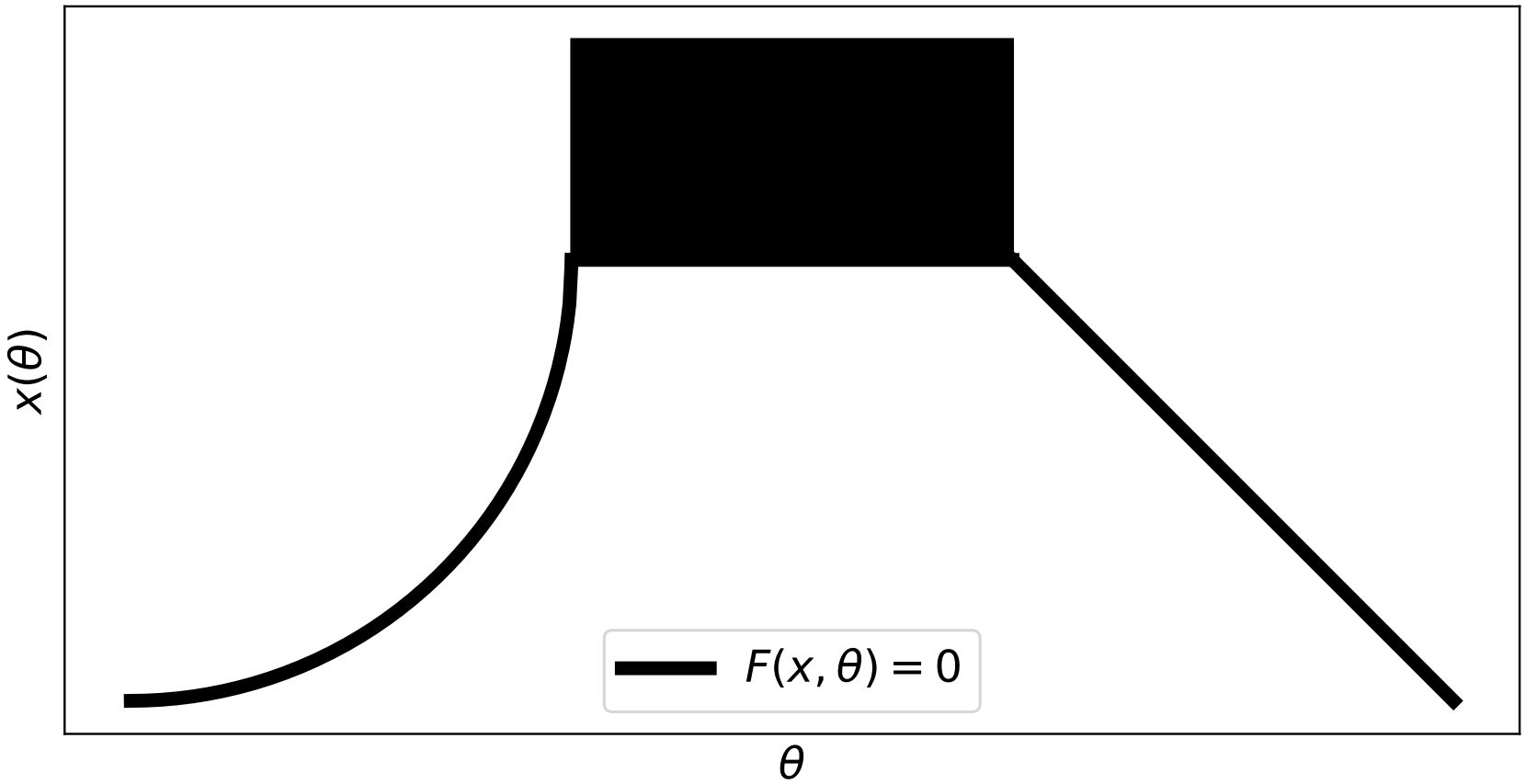
$$(M_\varepsilon \in \operatorname{range}(H_\varepsilon)) \quad M_\varepsilon = H_\varepsilon W_\varepsilon, \quad \|W_\varepsilon\| \leqslant C$$

$$-(H_\varepsilon + \varepsilon I_n)^{-1} M_\varepsilon = -\underbrace{(H_\varepsilon + \varepsilon I_n)^{-1} H_\varepsilon}_{0 \preceq \cdot \preceq I} \underbrace{W_\varepsilon}_{\|\cdot\| \leqslant C}.$$

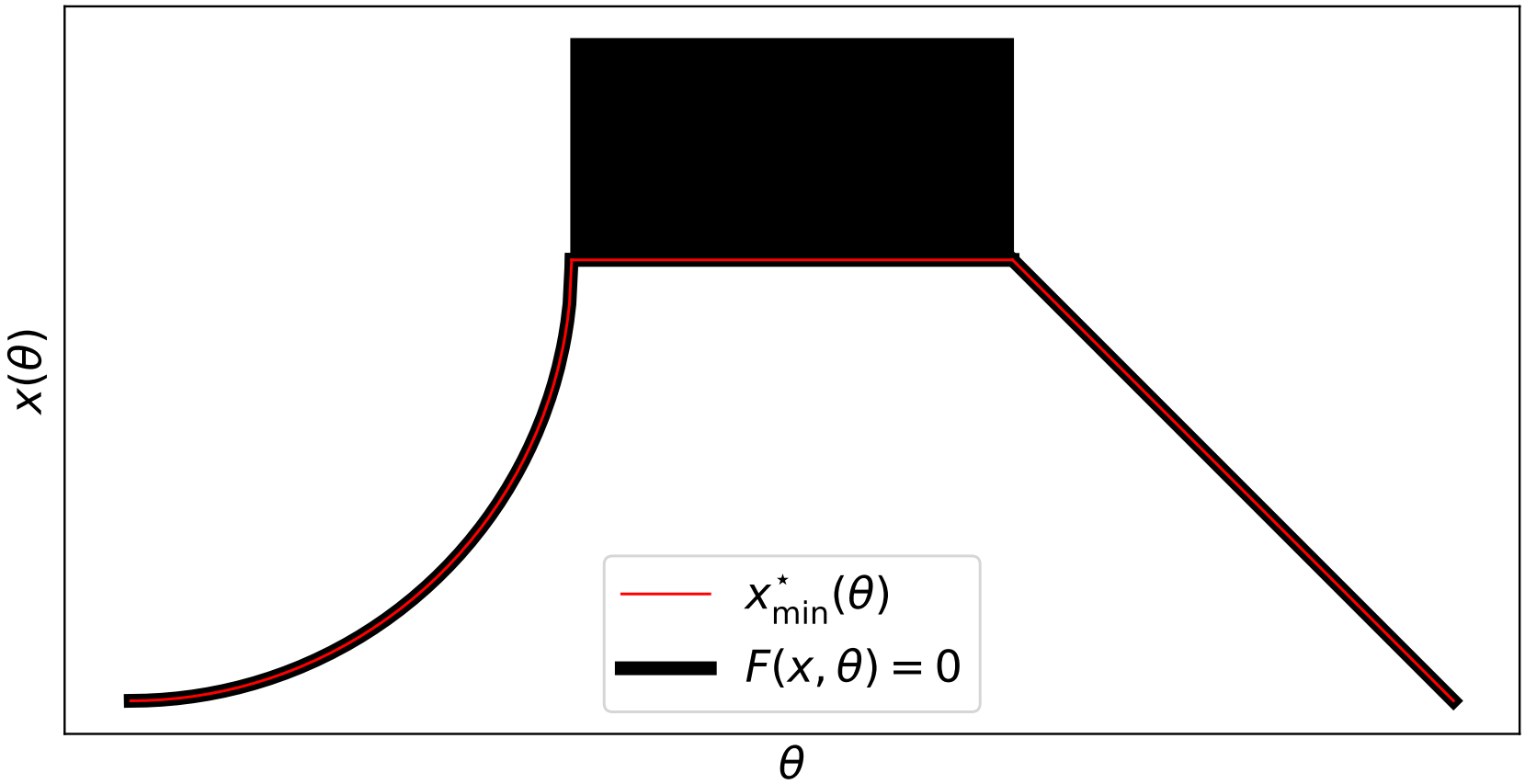
Least-Squares $M_\varepsilon \in \operatorname{range}(H_\varepsilon)$: $\underbrace{-A^\mathsf{T}}_{=M_\varepsilon} = -(\underbrace{A^\mathsf{T} A}_{H_\varepsilon}) \underbrace{A^\dagger}_{W_\varepsilon \equiv W}$ (SVD)

Nonsingularity: $\forall\ (H_0\ M_0) \in J_F(x^\star_{\min}(\theta), \theta), \det(H_0) \neq 0$ [Bol+21].
Continuity of det: nonsingularity $\Rightarrow$ uniform range; **(R)** can fail.

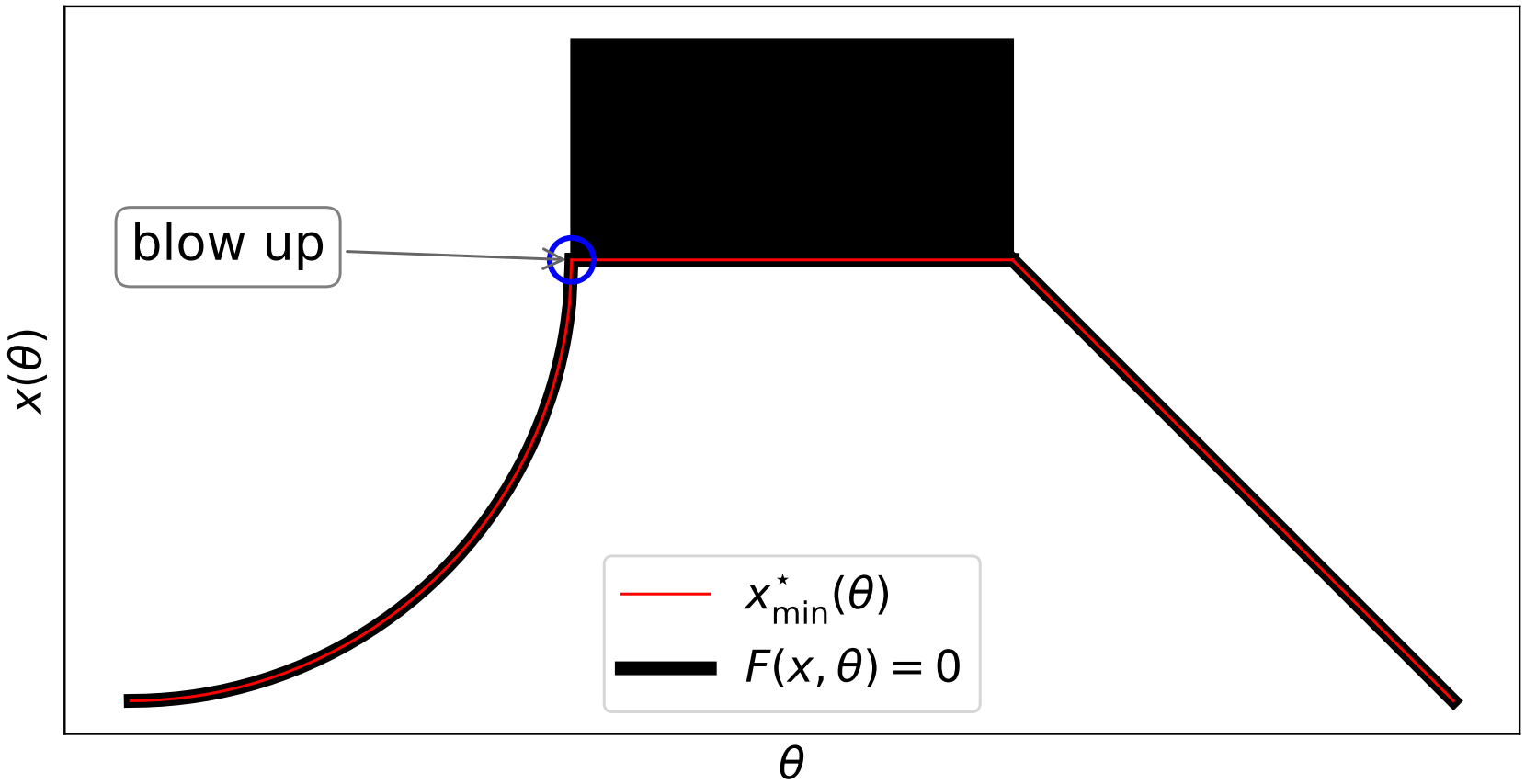
Example



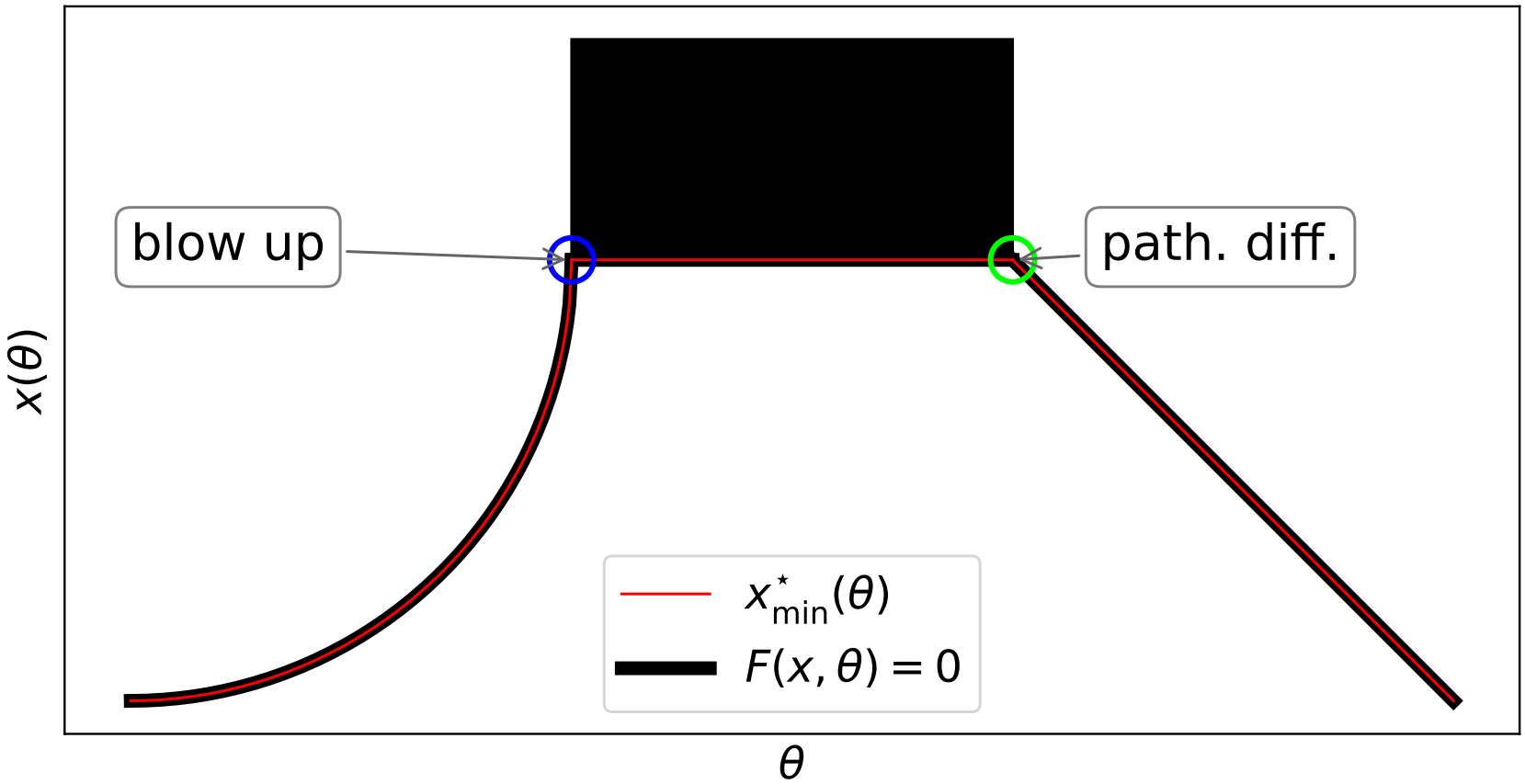
Example



Example



Example



Summary

- Regularization $\Rightarrow$ strong convexity $\Rightarrow$ conservative IFT.
- At the limit, uniform boundedness required.
- Provided by uniform range assumption.

Summary

- Regularization ⇒ strong convexity ⇒ conservative IFT.
- At the limit, uniform boundedness required.
- Provided by uniform range assumption.

About limits

- Boundedness ⇒ converging subsequences.
- What about the pseudoinverse?
- For the limit(s), a possible calculation mismatch.
- (Also a nonsmoothness issue.)

Outline

1 Introduction

2 Implicit Framework

3 At the Limit?

4 Applications

Actual Limits vs Pseudoinverse Formula

Let $0 < q < 1 < p$, $c, c' > 0$, and $H_\varepsilon = \mathrm{diag}(0, \varepsilon^p, c\varepsilon, \varepsilon^q, c')$,

$$\text{[with range]} \quad (H_\varepsilon + \varepsilon I_n)^{-1} H_\varepsilon \longrightarrow \mathrm{diag}(0, 0, c/(1+c), 1, 1)$$
$$\neq \mathrm{diag}(0, 0, 0, 0, 1) = H_0^\dagger H_0.$$

Actual Limits vs Pseudoinverse Formula

Let $0 < q < 1 < p$, $c, c' > 0$, and $H_\varepsilon = \mathrm{diag}(0, \varepsilon^p, c\varepsilon, \varepsilon^q, c')$,

$$\text{[with range]} \quad (H_\varepsilon + \varepsilon I_n)^{-1} H_\varepsilon \longrightarrow \mathrm{diag}(0, 0, c/(1+c), 1, 1)$$
$$\neq \mathrm{diag}(0, 0, 0, 0, 1) = H_0^\dagger H_0.$$

Observations:

- eigenvalues $\equiv 0$ (rank deficiency) or $\equiv c'$ ($\nrightarrow 0$) ✓;
- otherwise, curvature *vanishes* with ε;
- if $\lambda_i(H_\varepsilon) = o(\varepsilon)$ "fast", equals pseudoinverse,
- if $\varepsilon = o(\lambda_i(H_\varepsilon))$ "slow", mismatch,
- if $\lambda_i(H_\varepsilon) \sim \varepsilon$ "critical", smaller mismatch.

Actual Limits vs Pseudoinverse Formula

Let $0 < q < 1 < p$, $c, c' > 0$, and $H_\varepsilon = \mathrm{diag}(0, \varepsilon^p, c\varepsilon, \varepsilon^q, c')$,

$$\text{[with range]} \quad (H_\varepsilon + \varepsilon I_n)^{-1} H_\varepsilon \longrightarrow \mathrm{diag}(0, 0, c/(1+c), 1, 1)$$
$$\neq \mathrm{diag}(0, 0, 0, 0, 1) = H_0^\dagger H_0.$$

Observations:

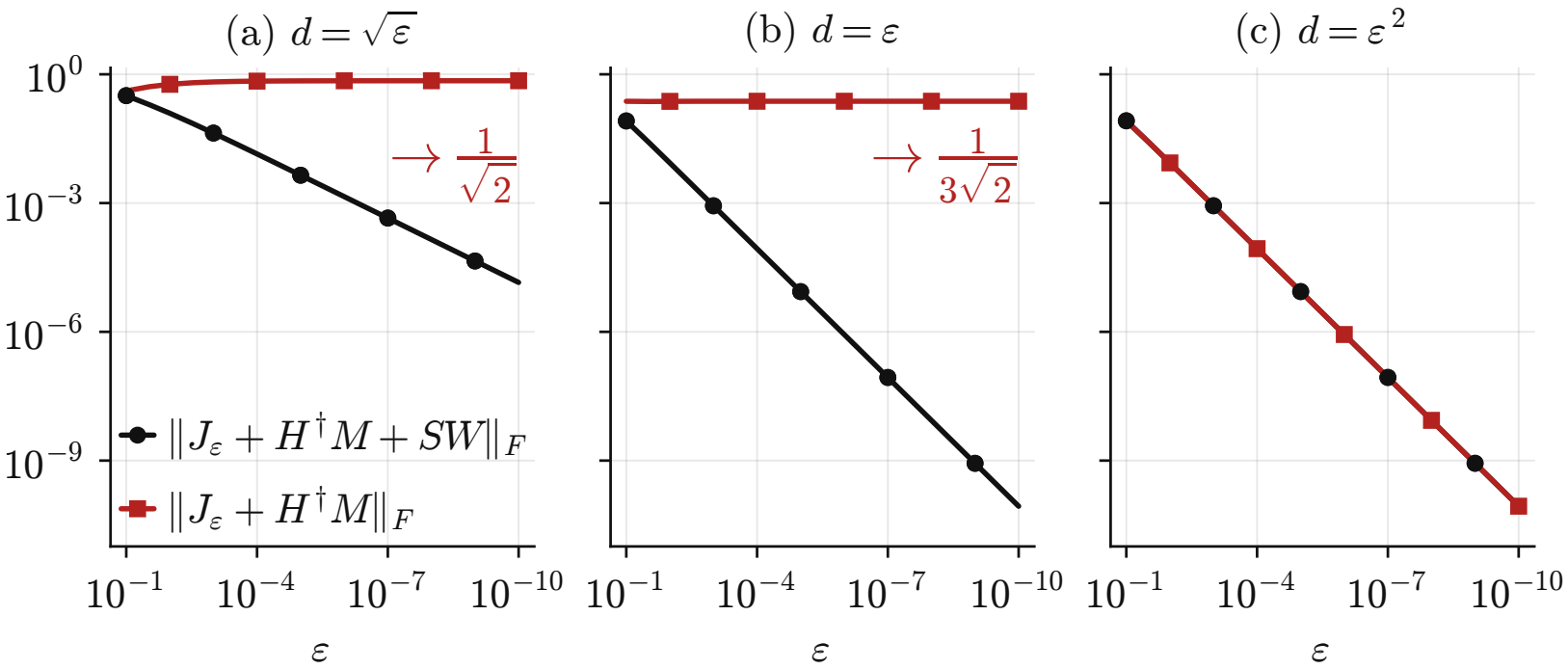
- eigenvalues $\equiv 0$ (rank deficiency) or $\equiv c'$ ($\nrightarrow 0$) ✓;
- otherwise, curvature *vanishes* with ε;
- if $\lambda_i(H_\varepsilon) = o(\varepsilon)$ "fast", equals pseudoinverse,
- if $\varepsilon = o(\lambda_i(H_\varepsilon))$ "slow", mismatch,
- if $\lambda_i(H_\varepsilon) \sim \varepsilon$ "critical", smaller mismatch.

Illustration

- Trivial LASSO / Soft-thresholding $\min_m \|m-\theta\|_2^2/2 + \alpha\|m\|_1$
- duplicate m: $\min_x \|x_1 + x_2 - \theta\|_2^2/2 + \alpha(\|x_1\|_1 + \|x_2\|)$ [Mey01; Auj+05]
- enforce nonsmooth coefficients $1 \in J_1$ and $q(\varepsilon) \in J_2$

Illustration

- Trivial LASSO / Soft-thresholding $\min_m \|m-\theta\|_2^2/2 + \alpha\|m\|_1$
- duplicate m: $\min_x \|x_1 + x_2 - \theta\|_2^2/2 + \alpha(\|x_1\|_1 + \|x_2\|)$ [Mey01; Auj+05]
- enforce nonsmooth coefficients $1 \in J_1$ and $q(\varepsilon) \in J_2$



Huber Regression and LASSO (lower level)

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{n_{\text{lower}}} h_\delta((Ax-\theta)_i), h_\delta(r) = \begin{cases} \frac{1}{2}r^2, & |r| \leqslant \delta, \\ \delta|r| - \frac{1}{2}\delta^2, & |r| > \delta, \end{cases}, \delta > 0$$

$$\text{or} \quad \min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-\theta\|_2^2 + e^\lambda \|x\|_1.$$

Good news

Global bound for the regularized Jacobians ✓

Bad news: this proof is specific for Huber/LASSO.

Huber Regression and LASSO (lower level)

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{n_{\text{lower}}} h_\delta\big((Ax-\theta)_i\big), h_\delta(r) = \begin{cases} \frac{1}{2}r^2, & |r| \leqslant \delta, \\ \delta|r| - \frac{1}{2}\delta^2, & |r| > \delta, \end{cases}, \delta > 0$$

$$\text{or} \quad \min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-\theta\|_2^2 + e^\lambda \|x\|_1.$$

Good news

Global bound for the regularized Jacobians ✓

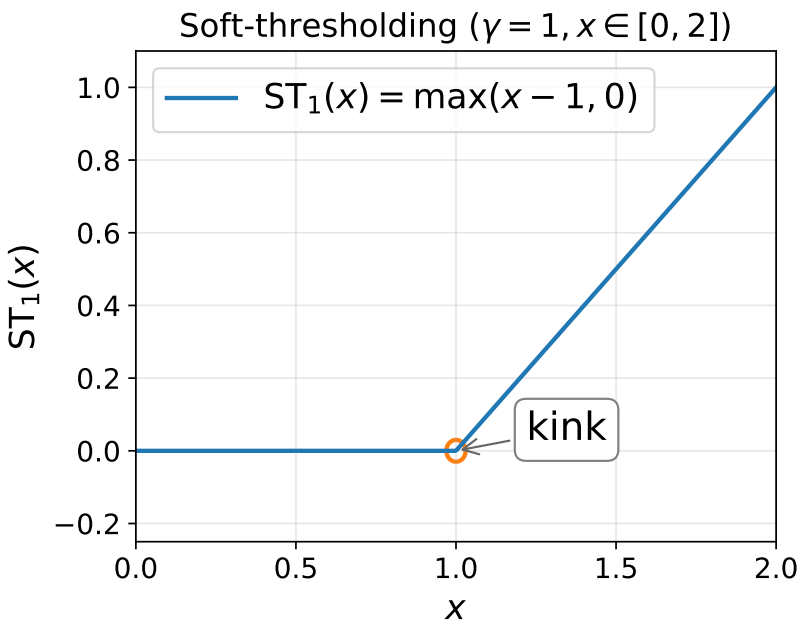
Bad news: this proof is specific for Huber/LASSO.

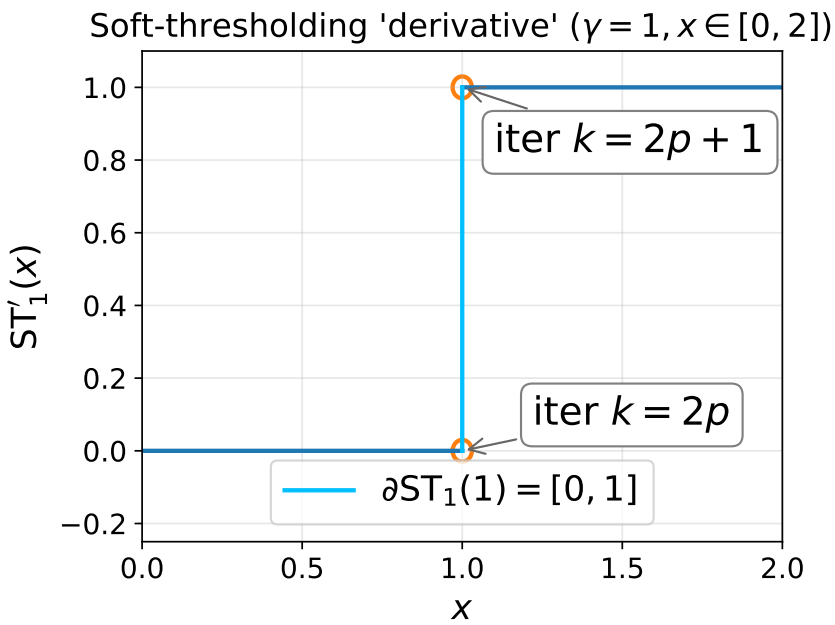
Continuation-type Technique

- Numerically, finite number of $\varepsilon > 0$;
- compute $J_{x_\varepsilon^\star}$, verify if they "stabilize" when $\varepsilon \to 0$
- compute the limit $\varepsilon = 0$ with stabilized sequence

Continuation-type Technique

- Numerically, finite number of $\varepsilon > 0$;
- compute $J_{x_\varepsilon^\star}$, verify if they "stabilize" when $\varepsilon \to 0$
- compute the limit $\varepsilon = 0$ with stabilized sequence





What should NOT happen.

Illustration: Huber Experiment

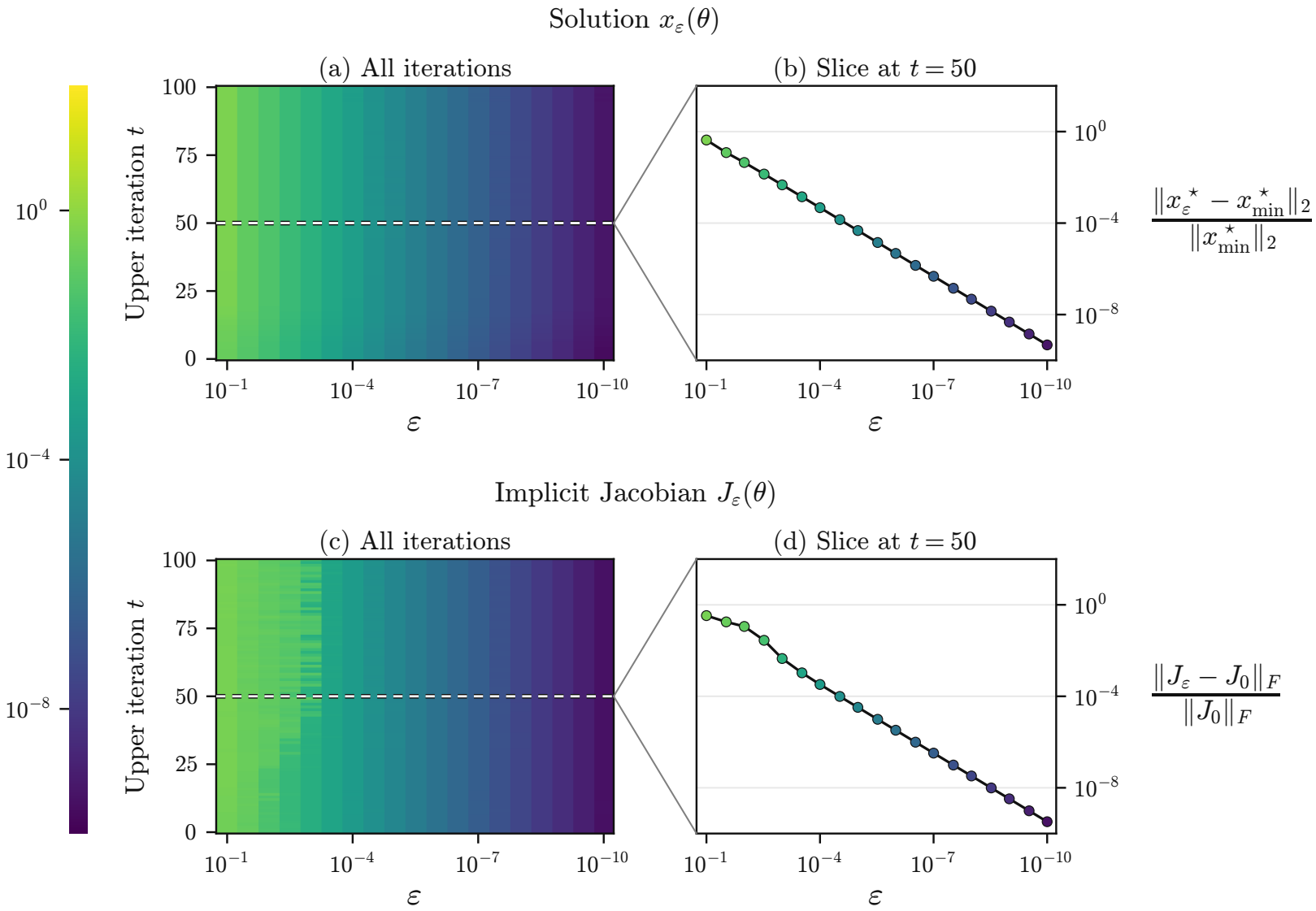
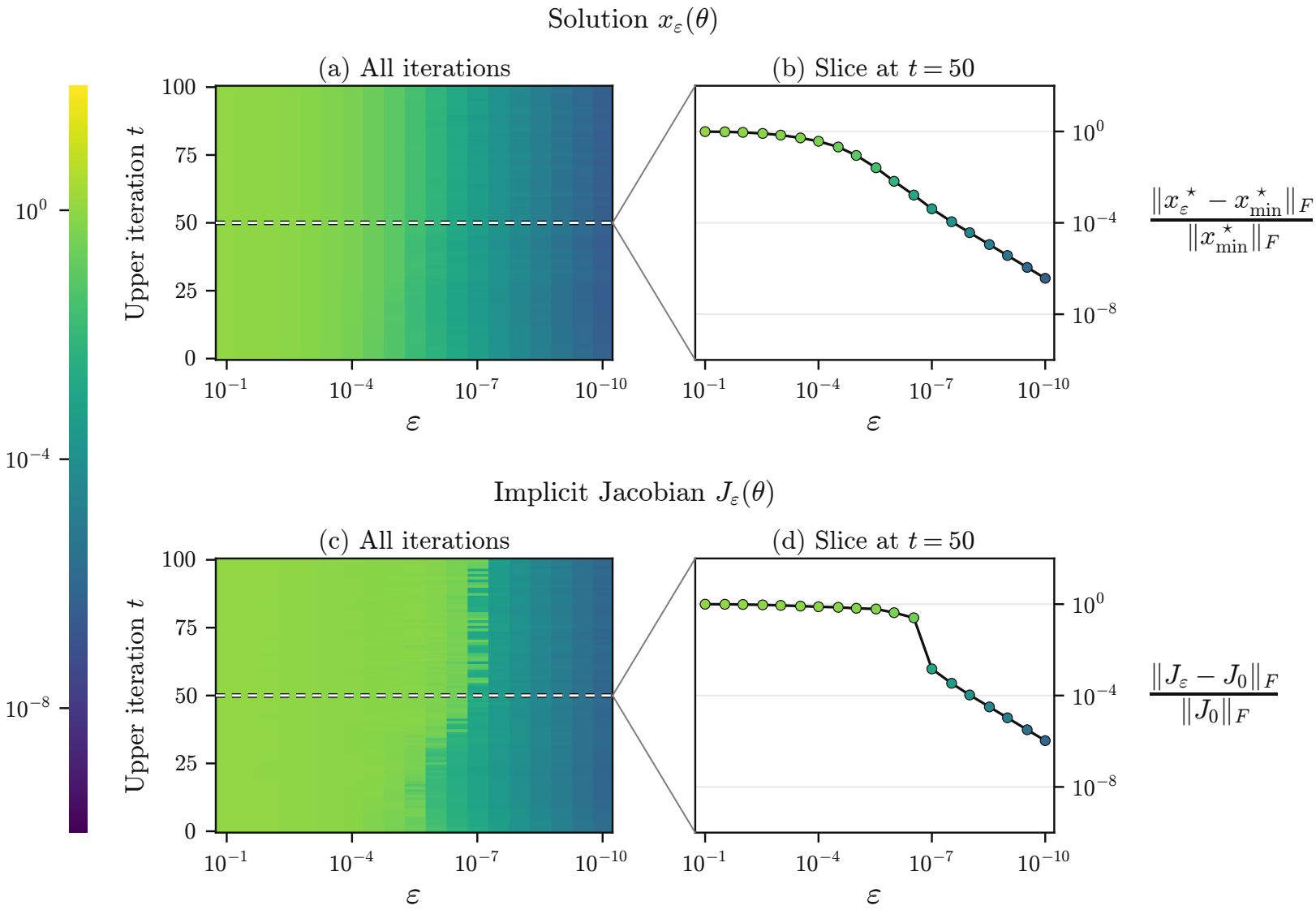


Illustration: LASSO Experiment



Conclusion

Summary

- theoretical justification of relevant practical methods;
- applicable to Huber & LASSO;
- some convincing numerical evidence.

Conclusion

Summary

- theoretical justification of relevant practical methods;
- applicable to Huber & LASSO;
- some convincing numerical evidence.

Uncharted directions

- assumptions may be hard to verify;
- ε-regularization limited by constraints; see also [MO24];
- deeper understanding of numerical behavior.

Bibliographic elements I

[AM22] Michael Arbel and Julien Mairal. "Non-Convex Bilevel Games with Critical Point Selection Maps". In: NeurIPS 2022 - Proceedings of the 36th Conference on Neural Information Processing Systems. Nov. 2022, pp. 1–34. DOI: 10.48550/arXiv.2207.04888.

[Auj+05] Jean-François Aujol et al. "Image Decomposition into a Bounded Variation Component and an Oscillating Component". In: Journal of Mathematical Imaging and Vision 22.1 (2005), pp. 71–88. DOI: 10.1007/s10851-005-4783-8.

[Ber+20] Quentin Bertrand et al. "Implicit differentiation of Lasso-type models for hyperparameter optimization". In: International Conference on Machine Learning. Feb. 2020, pp. 810–821. DOI: 10.48550/arXiv.2002.08943.

[BHS20] Pascal Bianchi, Walid Hachem, and Sholom Schechtman. "Convergence of Constant Step Stochastic Gradient Descent for Non-Smooth Non-Convex Functions". In: arXiv preprint arXiv:2005.08513 (2020). DOI: 10.48550/arXiv.2005.08513. URL: https://arxiv.org/abs/2005.08513.

[Blo+21] Mathieu Blondel et al. "Efficient and Modular Implicit Differentiation". In: Proceedings of the 36th Conference on Neural Information Processing Systems. May 2021. DOI: 10.48550/arXiv.2105.15183.

[Bol+21] Jérôme Bolte et al. "Nonsmooth Implicit Differentiation for Machine Learning and Optimization". In: Advances in Neural Information Processing Systems. Vol. 34. 2021, pp. 13537–13549. DOI: 10.48550/arXiv.2106.04350.

[Bol+26] Jérôme Bolte et al. "Bilevel Gradient Methods and the Morse Parametric Qualification Condition". In: Mathematics of Operations Research (Apr. 2026). DOI: 10.1287/moor.2025.0914.

[BP20] Jérôme Bolte and Edouard Pauwels. "Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning". In: Mathematical Programming 188 (Apr. 2020), pp. 19–51. DOI: 10.1007/s10107-020-01501-5.

Bibliographic elements II

[BPS24] Jérôme Bolte, Edouard Pauwels, and Antonio Silveti-Falls. "Differentiating Nonsmooth Solutions to Parametric Monotone Inclusion Problems". In: SIAM Journal on Optimization 34 (Jan. 2024), pp. 71–97. DOI: 10.1137/22M1541630.

[BPV22] Jérôme Bolte, Edouard Pauwels, and Samuel Vaiter. "Automatic differentiation of nonsmooth iterative algorithms". In: 36th Conference on Neural Information Processing Systems (NeurIPS 2022). May 2022. DOI: 10.48550/arXiv.2206.00457.

[Bro66] Felix E. Browder. "EXISTENCE AND APPROXIMATION OF SOLUTIONS OF NONLINEAR VARIATIONAL INEQUALITIES". In: Proceedings of the National Academy of Sciences 56.4 (1966), pp. 1080–1086. DOI: 10.1073/pnas.56.4.1080.

[Cla90] Francis H. Clarke. Optimization and Nonsmooth Analysis. second. Vol. 1. Classics in Applied Mathematics. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics (SIAM), 1990. ISBN: 0-89871-256-4.

[DD22] Damek Davis and Dmitriy Drusvyatskiy. "Conservative and Semismooth Derivatives are Equivalent for Semialgebraic Maps". In: Set-Valued and Variational Analysis 30 (2022), pp. 453–463. DOI: 10.1007/s11228-021-00594-0.

[Eng+25] Logan Engstrom et al. "Optimizing ML training with metagradient descent". In: arXiv preprint arXiv:2503.13751 (2025).

[Fra+18] Luca Franceschi et al. "Bilevel Programming for Hyperparameter Optimization and Meta-Learning". In: International Conference on Machine Learning. July 2018.

Bibliographic elements III

[GPS24] Riccardo Grazzi, Massimiliano Pontil, and Saverio Salzo. "Nonsmooth implicit differentiation: deterministic and stochastic convergence rates". In: 41st International Conference on Machine Learning. July 2024, pp. 16250–16274. DOI: 10.48550/arXiv.2403.11687.

[LT21] Adrian S. Lewis and Tonghua Tian. "The Structure of Conservative Gradient Fields". In: SIAM Journal on Optimization 31.3 (2021), pp. 2080–2083. DOI: 10.1137/21M1393637. URL: https://doi.org/10.1137/21M1393637.

[LVD19] Jonathan Lorraine, Paul Vicol, and David Duvenaud. "Optimizing Millions of Hyperparameters by Implicit Differentiation". In: International Conference on Artificial Intelligence and Statistics. Nov. 2019, pp. 1540–1552. DOI: 10.48550/arXiv.1911.02590.

[Mas+26] Saeed Masiha et al. "Select-then-differentiate: Solving Bilevel Optimization with Manifold Lower-level Solution Sets". In: arXiv preprint arXiv:2605.09209 (May 2026).

[MBP11] Julien Mairal, Francis Bach, and Jean Ponce. "Task-Driven Dictionary Learning". In: IEEE transactions on pattern analysis and machine intelligence 34 (July 2011), pp. 791–804. DOI: 10.1109/TPAMI.2011.156.

[Mey01] Yves Meyer. Oscillating Patterns in Image Processing and Nonlinear Evolution Equations. Vol. 22. University Lecture Series. Providence, Rhode Island: American Mathematical Society, 2001. DOI: 10.1090/ulect/022.

[MO24] Sheheryar Mehmood and Peter Ochs. "Automatic differentiation of optimization algorithms with time-varying updates". In: arXiv preprint arXiv:2410.15923 (2024).

[Ped16] Fabian Pedregosa. "Hyperparameter optimization with approximate gradient". In: Proceedings of the 33rd International Conference on Machine Learning. Feb. 2016, pp. 737–746. DOI: 10.48550/arXiv.1602.02355.

Bibliographic elements IV

[PF11] Gabriel Peyré and Jalal Fadili. "Learning Analysis Sparsity Priors". In: International Conference on Sampling Theory and Applications (SampTA). Singapour, 2011.

[Rad19] Hans Rademacher. "Über partielle und totale differenzierbarkeit von funktionen mehrerer variablen und über die transformation der doppelintegrale". In: Matematische Annalen 79 (Jan. 1919). DOI: 10.1007/BF01498415.

[Rob91] Stephen Robinson. "An Implicit-Function Theorem for a Class of Nonsmooth Functions". In: Mathematics of Operations Research - MOR 16 (May 1991), pp. 292–309. DOI: 10.1287/moor.16.2.292.

[Sch26] Sholom Schechtman. "The gradient's limit of a definable family of functions admits a variational stratification". In: SIAM Journal on Optimization 36.2 (2026), pp. 1075–1099.

[SV26] Ensio Suonperä and Tuomo Valkonen. "Single-loop approaches to nonsmooth bilevel optimisation". In: arXiv preprint arXiv:2606.19143 (2026).

[Vai+15] Samuel Vaiter et al. "Model selection with low complexity priors". In: Information and Inference: A Journal of the IMA 4.3 (2015), pp. 230–287.

[Vai+17] Samuel Vaiter et al. "The degrees of freedom of partly smooth regularizers". In: Annals of the Institute of Statistical Mathematics 69.4 (2017), pp. 791–832.

[Vic+22] Paul Vicol et al. "On Implicit Bias in Overparameterized Bilevel Optimization". In: 39th International Conference on Machine Learning PMLR. Dec. 2022, pp. 22234–22259. DOI: 10.48550/arXiv.2212.14032.

[Wan+18] Tongzhou Wang et al. "Dataset Distillation". In: International Conference on Learning Representations. Sept. 2018. DOI: 10.48550/arXiv.1811.10959.

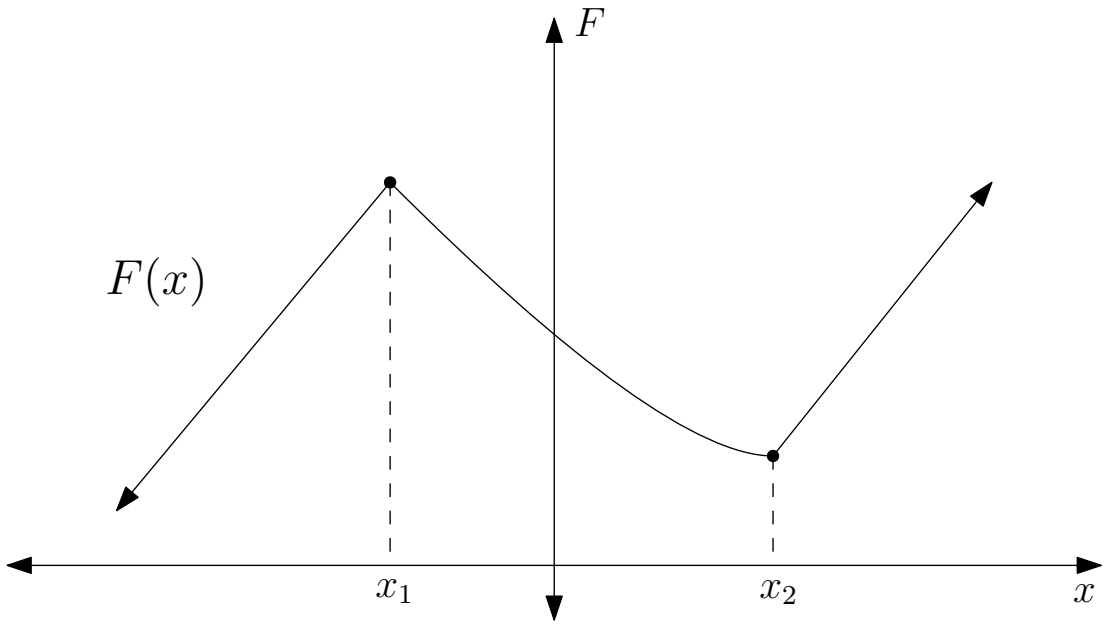
Bibliographic elements V

[WK20] Ezra Winston and J. Zico Kolter. "Monotone operator equilibrium networks". In: Advances in Neural Information Processing Systems. Vol. 33. June 2020. DOI: 10.48550/arXiv.2006.08591.

Illustration: captions

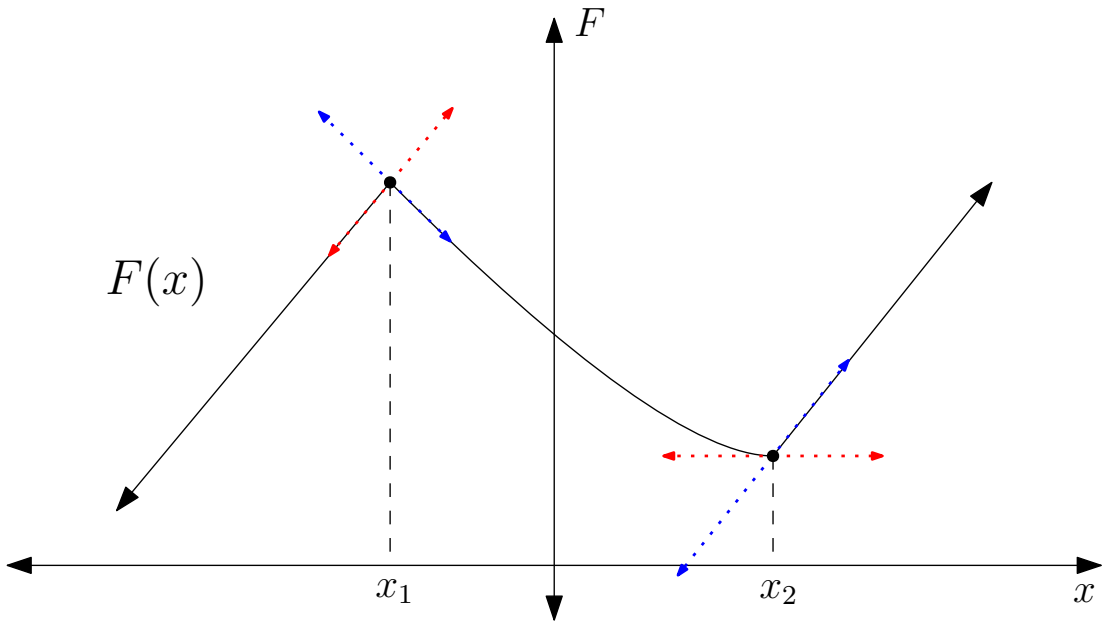
- Huber: *Top*, convergence of $x_\varepsilon^\star(\theta)$. *Bottom*, error on $J_{x_\varepsilon^\star}(\theta)$. Abscissa: ε; ordinate: upper iteration. Heatmaps: $\text{errors}(\varepsilon, t)$ ($\simeq$ independent of t). Curves: iteration $t = 50$. Linear regime is observed for $x_\varepsilon^\star$ all along, observed for $J_{x_\varepsilon^\star}$ with smaller ε.
- LASSO: *Top*, convergence of $x_\varepsilon^\star(\theta)$. *Bottom*, error on $J_{x_\varepsilon^\star}(\theta)$. Abscissa: ε; ordinate: upper iteration. Heatmaps: $\text{errors}(\varepsilon, t)$ ($\simeq$ independent of t). Curves: iteration $t = 50$. Linear regime observed for $x_\varepsilon^\star$ and $J_{x_\varepsilon^\star}$ at lower ε – support identification?

Nonsmooth but Locally Lipschitz Functions



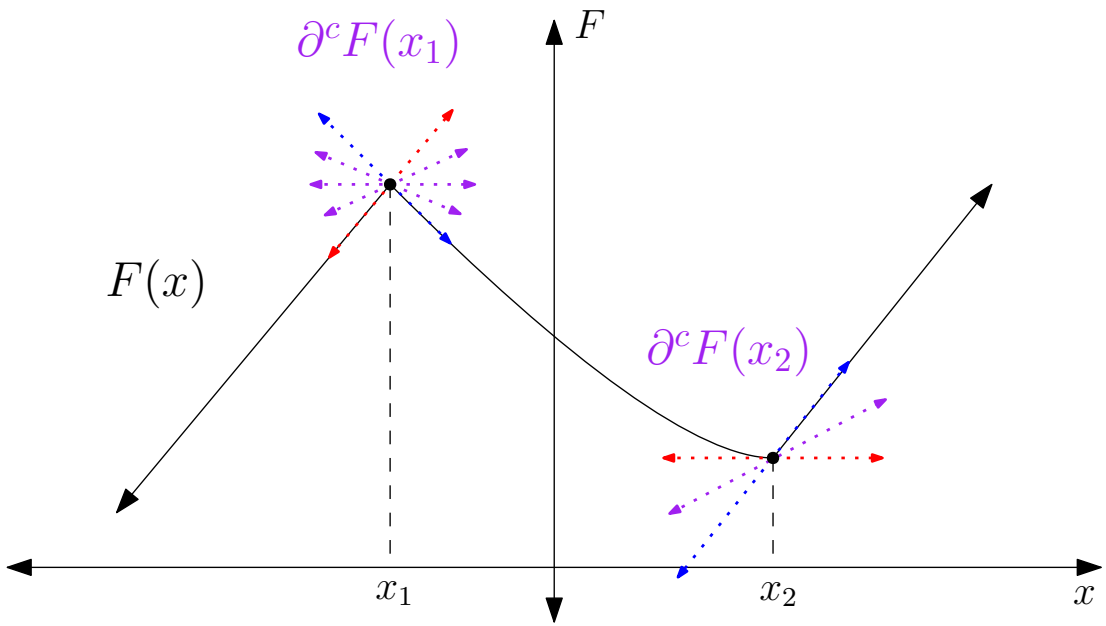
$$\partial^c F(x) := \operatorname{conv}\left(\{V : F'(x_k) \to V \text{ for } \{x_k\}_k \to x, F'(x_k) \text{ exists } \}\right)$$

Nonsmooth but Locally Lipschitz Functions



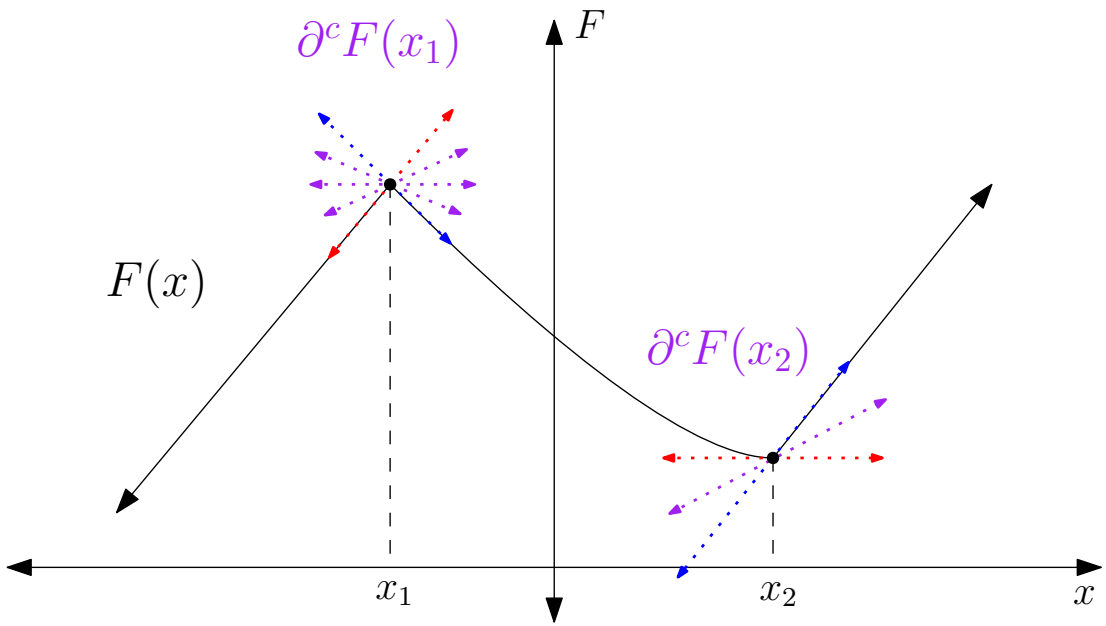
$$\partial^c F(x) := \operatorname{conv}\left(\{V : F'(x_k) \to V \text{ for } \{x_k\}_k \to x, F'(x_k) \text{ exists } \}\right)$$

Nonsmooth but Locally Lipschitz Functions



$$\partial^c F(x) := \operatorname{conv}\left(\{ V : F'(x_k) \to V \text{ for } \{x_k\}_k \to x, F'(x_k) \text{ exists } \}\right)$$

Nonsmooth but Locally Lipschitz Functions



$$\partial^c F(x) := \operatorname{conv}\left(\{ V : F'(x_k) \to V \text{ for } \{x_k\}_k \to x, F'(x_k) \text{ exists } \}\right)$$

- IFT **Existence**: theorems [Cla90; Rob91]...
- IFT **Calculus**: some issues.

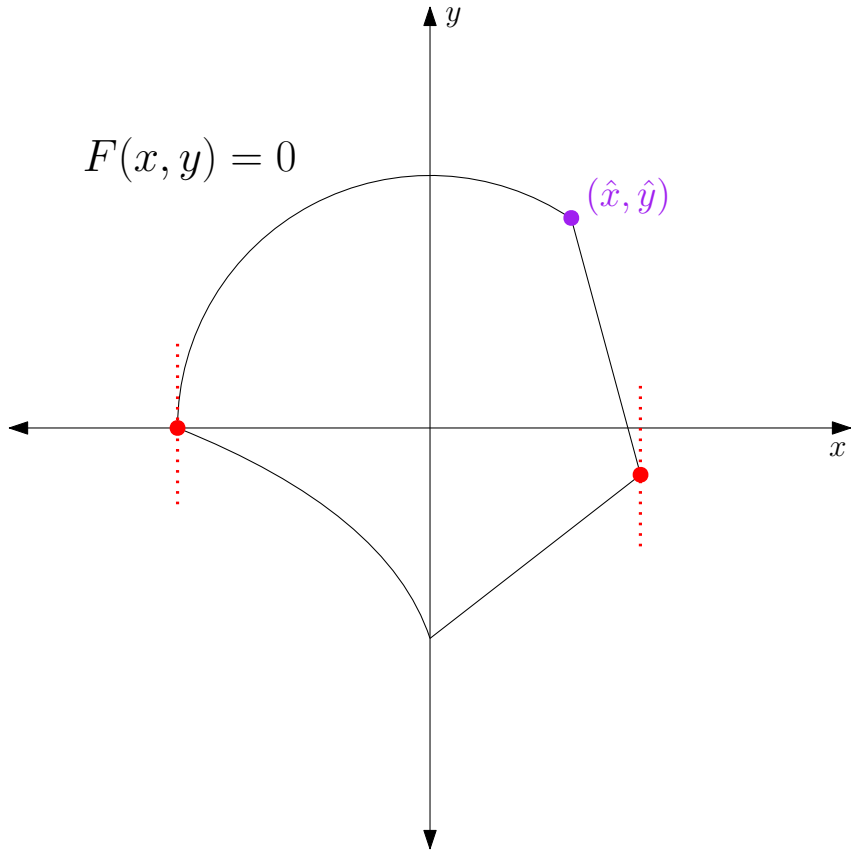
Nonsmooth IFT with Clarke Differential

Theorem ([Cla90]...)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ *be locally Lipschitz and* $F(\hat{x}, \hat{y}) = 0$.

If, $\forall (A\ B) \in \partial^c F(\hat{x}, \hat{y}), B$ *is invertible, then* $\exists\ U \subseteq \mathbb{R}^n$ *a neighborhood of* $\hat{x}$ *and a locally Lipschitz function* $G(x)$ *so that*

$$F(x, G(x)) = 0 \qquad \forall x \in U.$$



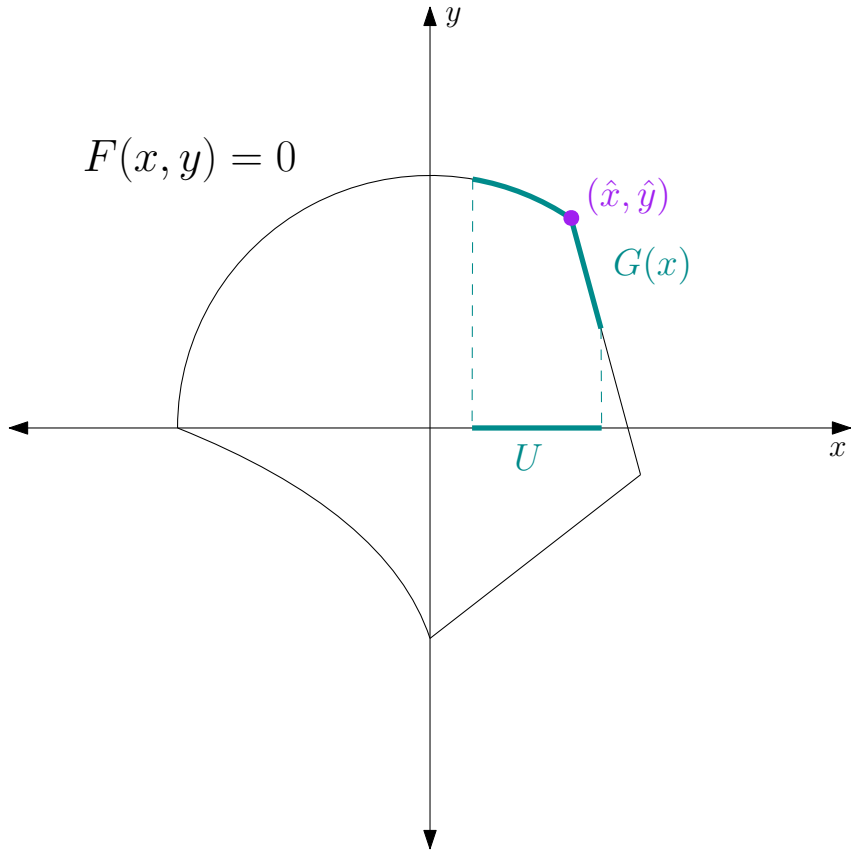
Nonsmooth IFT with Clarke Differential

Theorem ([Cla90]...)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz and $F(\hat{x}, \hat{y}) = 0$.

If, $\forall (A\ B) \in \partial^c F(\hat{x}, \hat{y}), B$ is invertible, then $\exists\ U \subseteq \mathbb{R}^n$ a neighborhood of $\hat{x}$ and a locally Lipschitz function $G(x)$ so that

$$F(x, G(x)) = 0 \qquad \forall x \in U.$$



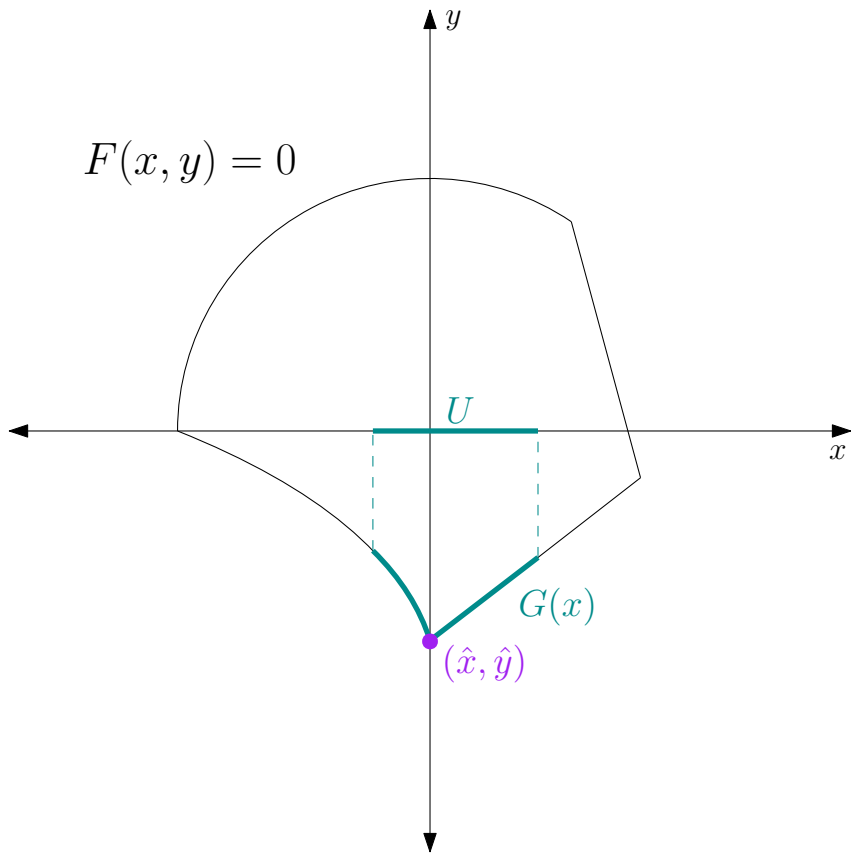
Nonsmooth IFT with Clarke Differential

Theorem ([Cla90]...)

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be locally Lipschitz and $F(\hat{x}, \hat{y}) = 0$.

If, $\forall (A\ B) \in \partial^c F(\hat{x}, \hat{y}), B$ is invertible, then $\exists\ U \subseteq \mathbb{R}^n$ a neighborhood of $\hat{x}$ and a locally Lipschitz function $G(x)$ so that

$$F(x, G(x)) = 0 \qquad \forall x \in U.$$



But by conv: many B elements.

Conservative Fields (/Mappings/Jacobians)

Definition (conservative field of a locally Lipschitz function [BP20])

Let $F : \mathbb{R}^n \to \mathbb{R}$, $D_F : \mathbb{R}^n \rightrightarrows \mathbb{R}$ is a conservative field of F if

- $\forall\, x, D_F(x) \neq \varnothing$
- D_F is locally bounded and has closed graph;
- $\forall\, \gamma : [0,1] \to \mathbb{R}^n$ absolutely continuous ($\dot{\gamma}(t)$ exists a.e. t),

$$\underset{\text{rule}}{\text{chain}} \quad \frac{\mathrm{d}}{\mathrm{d}t} F(\gamma(t)) = (u, \dot{\gamma}(t)), \quad \forall\, u \in D_F(\gamma(t)), \text{a.e. } t$$

Chain rule justifies backpropagation (Jax, Pytorch...)

Conservative Fields (/Mappings/Jacobians)

Definition (conservative field of a locally Lipschitz function [BP20])

Let $F : \mathbb{R}^n \to \mathbb{R}$, $D_F : \mathbb{R}^n \rightrightarrows \mathbb{R}$ is a conservative field of F if

- $\forall\, x, D_F(x) \neq \varnothing$
- D_F is locally bounded and has closed graph;
- $\forall\, \gamma : [0,1] \to \mathbb{R}^n$ absolutely continuous ($\dot\gamma(t)$ exists a.e. t),

$$\underset{\text{rule}}{\text{chain}} \qquad \frac{\mathrm{d}}{\mathrm{d}t} F(\gamma(t)) = (u, \dot\gamma(t)), \quad \forall\, u \in D_F(\gamma(t)), \text{a.e. } t$$

Chain rule justifies backpropagation (Jax, Pytorch...)

Conservative Fields (/Mappings/Jacobians)

Definition (conservative field of a locally Lipschitz function [BP20])

Let $F : \mathbb{R}^n \to \mathbb{R}$, $D_F : \mathbb{R}^n \rightrightarrows \mathbb{R}$ is a conservative field of F if

- $\forall\, x, D_F(x) \neq \varnothing$
- D_F is locally bounded and has closed graph;
- $\forall\, \gamma : [0,1] \to \mathbb{R}^n$ absolutely continuous ($\dot{\gamma}(t)$ exists a.e. t),

$$\text{chain rule} \qquad \frac{\mathrm{d}}{\mathrm{d}t} F(\gamma(t)) = (u, \dot{\gamma}(t)), \quad \forall\, u \in D_F(\gamma(t)), \text{a.e. } t$$

Chain rule justifies backpropagation (Jax, Pytorch...)

Conservative Fields (/Mappings/Jacobians)

Definition (conservative field of a locally Lipschitz function [BP20])

Let $F : \mathbb{R}^n \to \mathbb{R}$, $D_F : \mathbb{R}^n \rightrightarrows \mathbb{R}$ is a conservative field of F if

- $\forall\, x, D_F(x) \neq \varnothing$
- D_F is locally bounded and has closed graph;
- $\forall\, \gamma : [0,1] \to \mathbb{R}^n$ absolutely continuous ($\dot{\gamma}(t)$ exists a.e. t),

$$\text{chain rule} \qquad \frac{\mathrm{d}}{\mathrm{d}t} F(\gamma(t)) = (u, \dot{\gamma}(t)), \quad \forall\, u \in D_F(\gamma(t)), \text{a.e. } t$$

Chain rule justifies backpropagation (Jax, Pytorch...)

Some Properties of Conservative Fields

- Generalization of ∂^c but not a unique D_F.
- Usual calculus rules: sum, product, **composition**.
- Almost everywhere, $D_F(x) = \nabla F(x)$ (see [Rad19]).
- $\partial^c F(x) \subseteq \mathrm{conv}(D_F(x))$ $(\forall\ D_F, \forall\ x)$.
- Subgradient-type convergence properties.
- See also [Bol+21; DD22; LT21; BHS20; BPV22; GPS24].

Some Properties of Conservative Fields

- Generalization of ∂^c but not a unique D_F.
- Usual calculus rules: sum, product, **composition**.
- Almost everywhere, $D_F(x) = \nabla F(x)$ (see [Rad19]).
- $\partial^c F(x) \subseteq \mathrm{conv}(D_F(x))$ ($\forall\ D_F, \forall\ x$).
- Subgradient-type convergence properties.
- See also [Bol+21; DD22; LT21; BHS20; BPV22; GPS24].

Some Properties of Conservative Fields

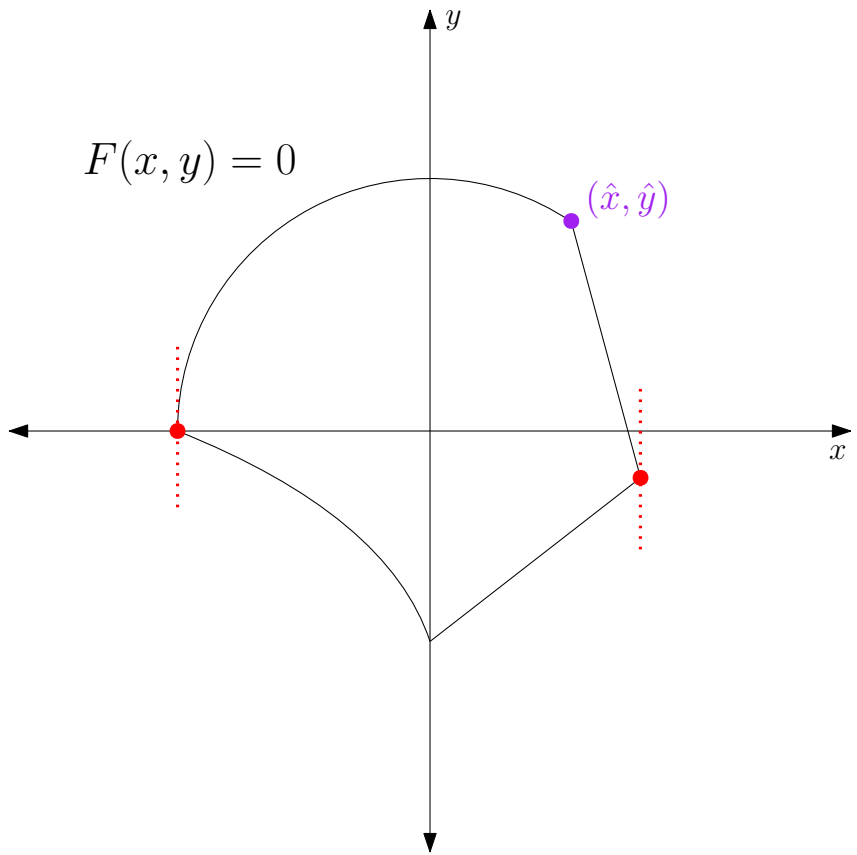
- Generalization of ∂^c but not a unique D_F.
- Usual calculus rules: sum, product, **composition**.
- Almost everywhere, $D_F(x) = \nabla F(x)$ (see [Rad19]).
- $\partial^c F(x) \subseteq \mathrm{conv}(D_F(x))$ ($\forall\ D_F, \forall\ x$).
- Subgradient-type convergence properties.
- See also [Bol+21; DD22; LT21; BHS20; BPV22; GPS24].

Path-differentiable/conservative Nonsmooth IFT

Theorem ([Bol+21])

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ *be path diff. and* $F(\hat{x}, \hat{y}) = 0$.
If $D_F(\hat{x}, \hat{y})$ *is convex and* B *is invertible* $\forall\ (A\ B) \in D_F(\hat{x}, \hat{y})$,
then $\exists\ U \subseteq \mathbb{R}^n$ *a neighborhood of* $\hat{x}$ *and a path diff. function* G

$$\forall x \in U, \quad F(x, G(x)) = 0,$$

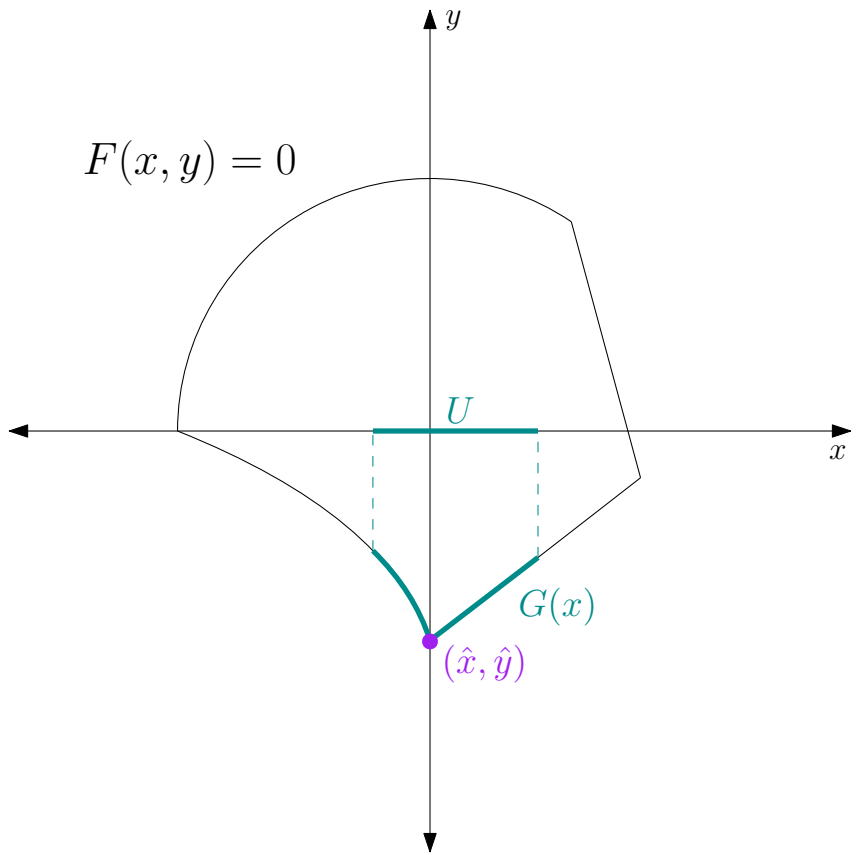


Path-differentiable/conservative Nonsmooth IFT

Theorem ([Bol+21])

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ *be path diff. and* $F(\hat{x}, \hat{y}) = 0$.
If $D_F(\hat{x}, \hat{y})$ *is convex and* B *is invertible* $\forall\ (A\ B) \in D_F(\hat{x}, \hat{y})$, *then* $\exists\ U \subseteq \mathbb{R}^n$ *a neighborhood of* $\hat{x}$ *and a path diff. function* G

$$\forall x \in U, \quad F(x, G(x)) = 0,$$



Path-differentiable/conservative Nonsmooth IFT

Theorem ([Bol+21])

Let $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ *be path diff. and* $F(\hat{x}, \hat{y}) = 0$.
If $D_F(\hat{x}, \hat{y})$ *is convex and* B *is invertible* $\forall\ (A\ B) \in D_F(\hat{x}, \hat{y})$, *then* $\exists\ U \subseteq \mathbb{R}^n$ *a neighborhood of* $\hat{x}$ *and a path diff. function* G *with conservative mapping* D_G *verifying*

$$\forall x \in U, \quad F(x, G(x)) = 0,$$

$$D_G(x) = \left\{ \begin{array}{l} -B^{-1}A : (A\ B) \\ \in D_F(x, G(x)) \end{array} \right\}.$$

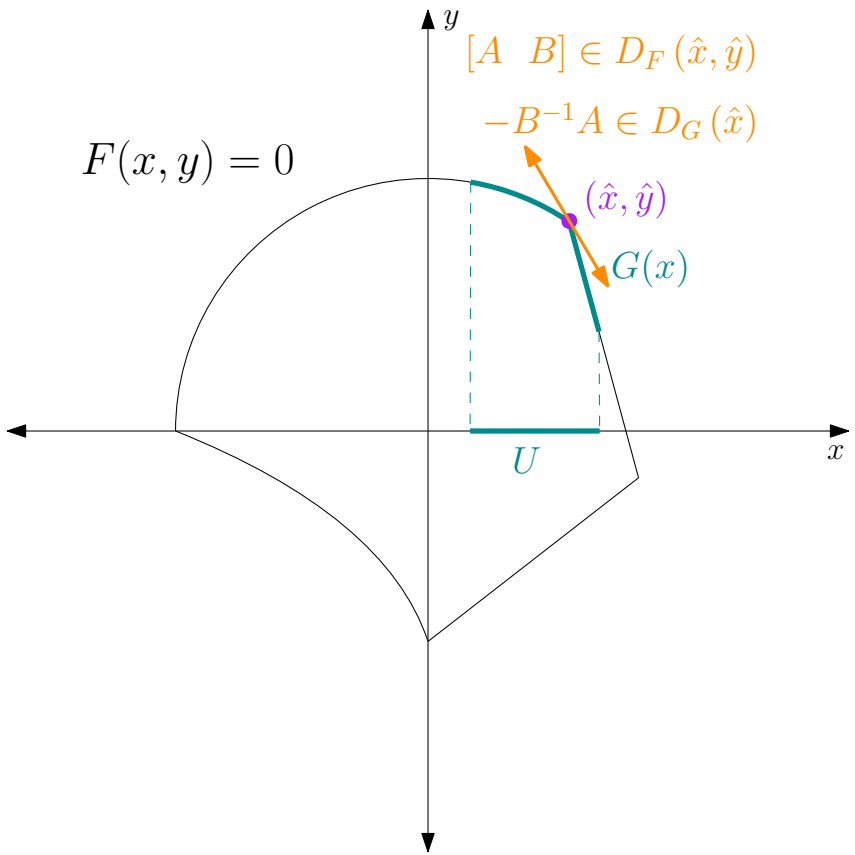
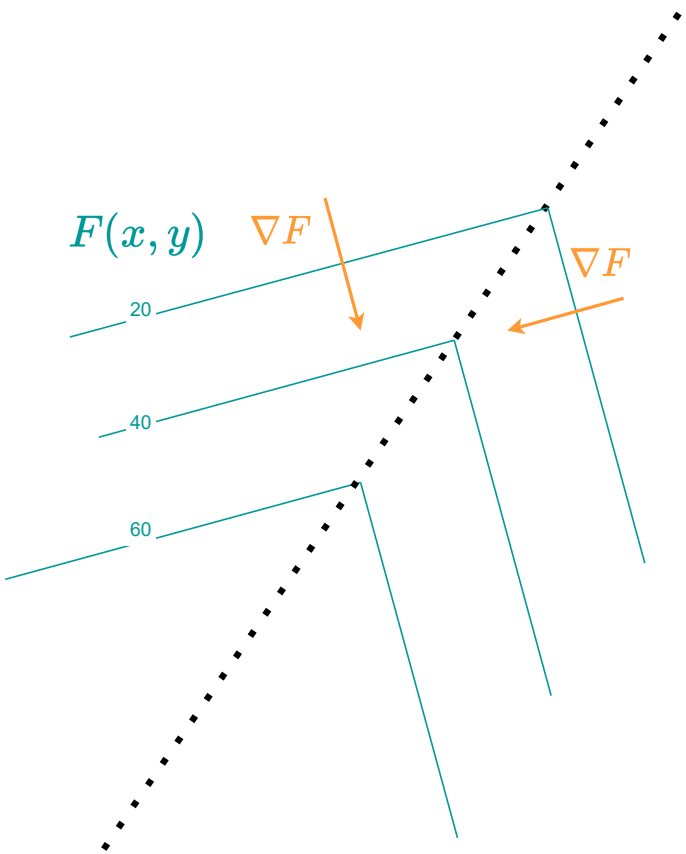
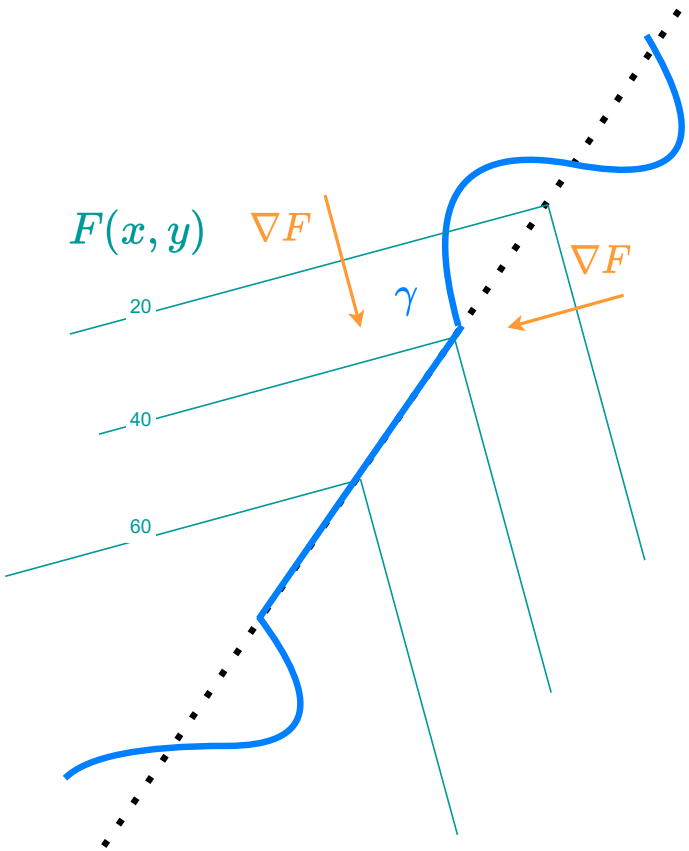


Illustration on Piecewise Affine Functions



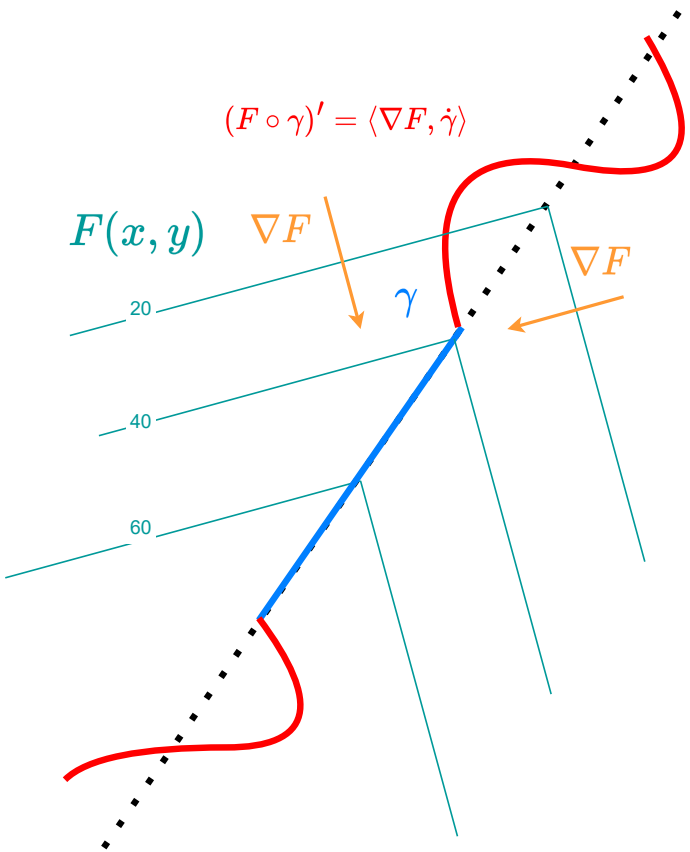
Locally Lipschitz function $F : \mathbb{R}^2 \to \mathbb{R}$ with piecewise constant gradient and kink in black.

Illustration on Piecewise Affine Functions



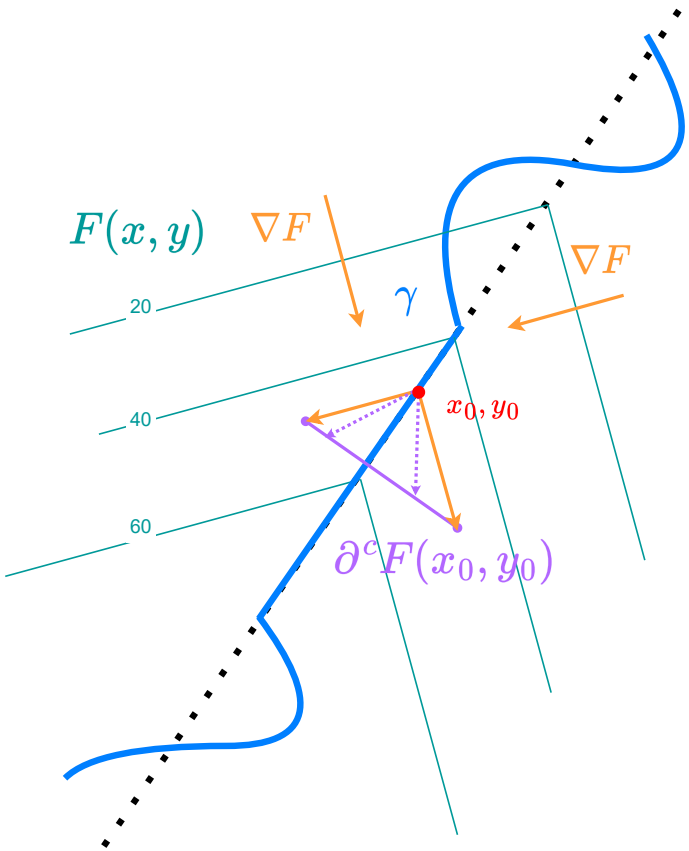
Consider some absolutely continuous curve $\gamma\colon [0,1] \to \mathbb{R}^2$.

Illustration on Piecewise Affine Functions



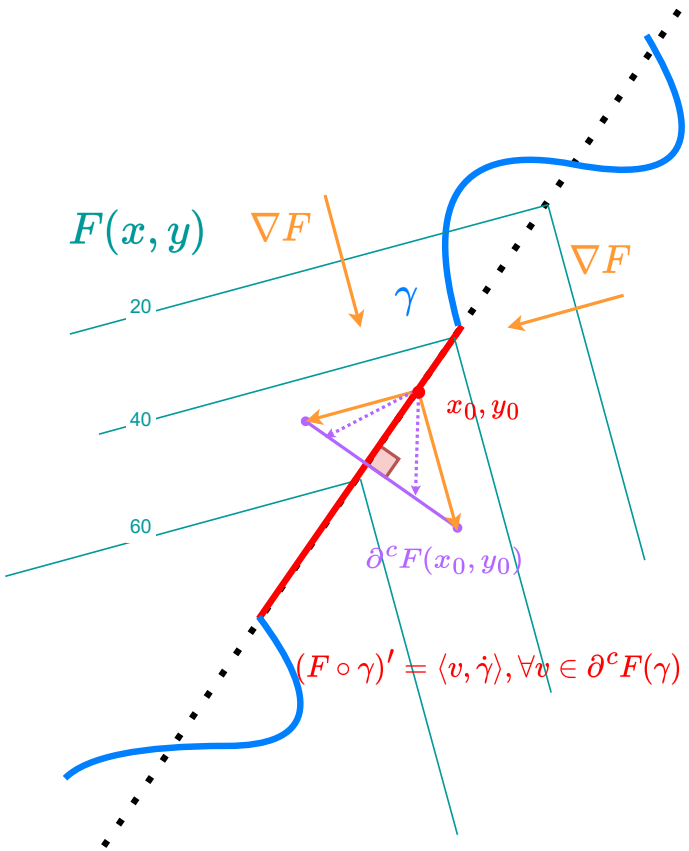
The red parts of $\gamma([0, 1])$ only see the bad region for a set of measure zero in t.

Illustration on Piecewise Affine Functions



For a point in the bad region, the Clarke subdifferential has this form.

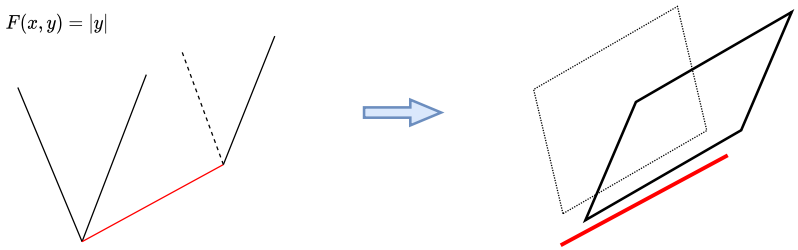
Illustration on Piecewise Affine Functions



The formula holds for any point in the bad region.

Relation with Definability and o-minimal Geometry

Piecewise affine functions are easily decomposed as earlier

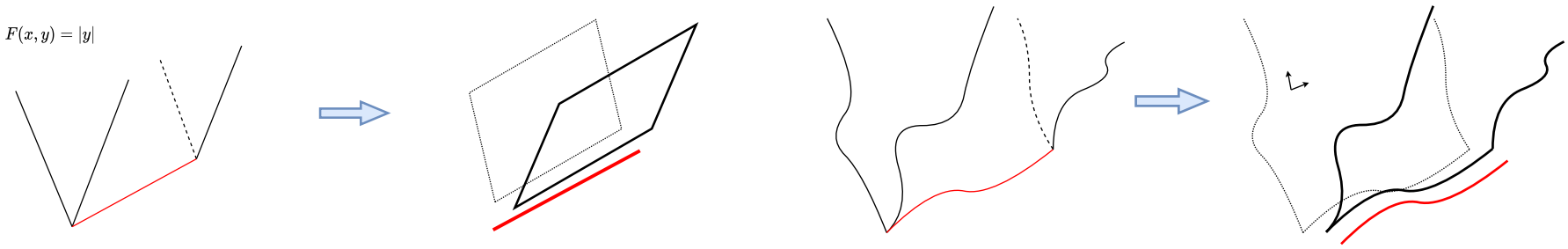


Definability in our setting

- definable geometry generalizes this beyond affine functions;
- graphs of definable functions are decomposable into manifolds that are locally like piecewise affine ("Whitney stratification");
- locally Lipschitz definable $\Rightarrow$ path-differentiable;
- we will assume "everything" is definable.

Relation with Definability and o-minimal Geometry

Piecewise affine functions are easily decomposed as earlier

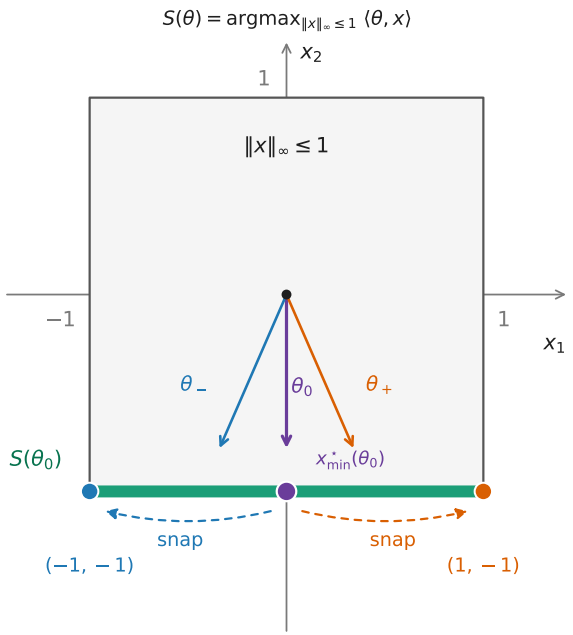


Definability in our setting

- definable geometry generalizes this beyond affine functions;
- graphs of definable functions are decomposable into manifolds that are locally like piecewise affine ("Whitney stratification");
- locally Lipschitz definable $\Rightarrow$ path-differentiable;
- we will assume "everything" is definable.

Failure for constrained/composite problems

Consider: $\max \theta^\mathsf{T} x, \|x\|_\infty \leqslant 1 \Leftrightarrow \max \theta^\mathsf{T} x - \iota_{B_\infty(0,1)}(x)$.



We have $x^\star_{\min}(\theta) = \operatorname{sgn}(\theta)$ ($\operatorname{sgn}(0) = 0$). For $\theta_0 = (0, -1)$, $\mathcal{S}(\theta_0)$ is in green and $x^\star_{\min}(\theta_0) = (0, -1)$. There is no continuity thus no path-differentiability. Perturbing θ_0 collapses the edge $\mathcal{S}(\theta_0)$ to a corner.

Illustration: Different Regimes (1)

Test problem: duplicated soft-thresholding (trivial LASSO)

$$\min_m h(m) := \frac{1}{2}\|m-\theta\|_2^2 + \alpha\|m\|_1 \tag{2}$$

Version with duplicate variable ($m = x_1 + x_2$):

$$\min_{x_1,x_2} \tilde{h}(x_1,x_2) := \frac{1}{2}\|x_1+x_2-\theta\|_2^2 + \alpha(\|x_1\|_1 + \|x_2\|_1) \tag{3}$$

with ε regularization; we take $\theta \in \mathbb{R}^n_+$ and $\alpha = \theta_k$.

Illustration: Different Regimes (1)

Test problem: duplicated soft-thresholding (trivial LASSO)

$$\min_m h(m) := \frac{1}{2}\|m-\theta\|_2^2 + \alpha\|m\|_1 \tag{2}$$

Version with duplicate variable ($m = x_1 + x_2$):

$$\min_{x_1,x_2} \tilde{h}(x_1,x_2) := \frac{1}{2}\|x_1+x_2-\theta\|_2^2 + \alpha(\|x_1\|_1 + \|x_2\|_1) \tag{3}$$

with ε regularization; we take $\theta \in \mathbb{R}^n_+$ and $\alpha = \theta_k$.

$\alpha = \theta_k \Rightarrow$ nonsmoothness on index k; $J^1_{i\neq k} = J^2_{i\neq k}$, $J^1_k \neq J^2_k$.
We **intentionally** construct a mismatch on coordinate k.

Detailed range assumption

Uniform range assumption **(R)**

Let $O \subseteq \Theta$ open, $\exists\ \bar{\varepsilon} > 0$ s.t. $\forall\ K \subseteq O$ compact, $\exists\ C_K > 0$, s.t. $\forall\ \theta \in K, \forall\ 0 < \varepsilon \leqslant \bar{\varepsilon}$ and $\forall\ (H_\varepsilon\ M_\varepsilon) \in J_F(x_\varepsilon^\star(\theta), \theta)$, $H_\varepsilon \succeq 0$ and $M_\varepsilon = H_\varepsilon W_\varepsilon$ with $\|W_\varepsilon\| \leqslant C$:

$$-\left(H_\varepsilon + \varepsilon I_n\right)^{-1} M_\varepsilon = -\underbrace{\left(H_\varepsilon + \varepsilon I_n\right)^{-1} H_\varepsilon}_{0 \preceq \cdot \preceq I} \underbrace{W_\varepsilon}_{\|\cdot\| \leqslant C}.$$

Vanishing curvature

Counterexample (in 1D)

$$f(x,\theta) = \psi(x-\theta), \psi(t) = \begin{cases} t^4/4 & |t| \leqslant 1, \\ |t| - 3/4 & |t| > 1. \end{cases}$$

- $x(\theta) = \theta, J_{x^\star_{\min}} \equiv 1 \neq -H_0^\dagger M_0 = -0^\dagger 0 = 0$;
- $H_\varepsilon \sim \varepsilon^{2/3}, M_\varepsilon = -H_\varepsilon$: $\lim_{\varepsilon \to 0} (\varepsilon^{2/3} + \varepsilon)^{-1} \varepsilon^{2/3} = 1$.

Assumptions to use the pseudoinverse

Eigenvalues of Hessian blocks must not vanish or vanish fast.

Statement for the pseudoinverse formula

- **(G)** $\exists\ \lambda^*, \forall\ \varepsilon \leqslant \varepsilon_G, \mathrm{spectrum}(H_\varepsilon) \subseteq \{0\} \cup [\lambda^*; +\infty[$ ("gap")
- **(V)** for all converging sequences H_ε with ordered eigenvalues $\lambda_1(H_\varepsilon) \geqslant \ldots$, $\lambda_i(H_\varepsilon) \to 0 \Rightarrow \lambda_i(H_\varepsilon) = o(\varepsilon)$ ("fast vanishing")

Without slow/critical vanishing (and previous assumptions), the pseudoinverse formula is valid.

Setting

$$\theta \in O \subseteq \Theta, \quad \min_{x \in \mathbb{R}^n} \left\{ f(x, \theta) + g(x, \theta) \quad \left(+ \frac{\varepsilon}{2} \|x\|_2^2 \right) \right\} \quad \text{(P-Comp)}$$

- everything is definable;
- $\forall\, \theta \in O$, $f(\cdot, \theta) \in \mathcal{C}^{1,1}(\mathbb{R}^n) \cap \Gamma_0(\mathbb{R}^n)$, $g(\cdot, \theta) \in \Gamma_0(\mathbb{R}^n)$;
- $\forall\, \theta \in O$, $\mathcal{S}_c(\theta) := \arg\min_{x \in \mathbb{R}^n} \{ f(x, \theta) + g(x, \theta) \} \neq \varnothing$;
- $\nabla_x f$ is jointly locally Lipschitz and verifies $\|\nabla_x f(x, \theta) - \nabla_x f(x', \theta)\| \leqslant L \|x - x'\|$;
- $G^\gamma(z, \theta) := \operatorname{prox}_{\gamma g(\cdot, \theta)}(z)$ is jointly locally Lipschitz ($\gamma > 0$).

Setting

$$\theta \in O \subseteq \Theta, \quad \min_{x \in \mathbb{R}^n} \left\{ f(x, \theta) + g(x, \theta) \quad \left(+ \frac{\varepsilon}{2} \|x\|_2^2 \right) \right\} \quad \text{(P-Comp)}$$

- everything is definable;
- $\forall\ \theta \in O$, $f(\cdot, \theta) \in \mathcal{C}^{1,1}(\mathbb{R}^n) \cap \Gamma_0(\mathbb{R}^n)$, $g(\cdot, \theta) \in \Gamma_0(\mathbb{R}^n)$;
- $\forall\ \theta \in O$, $\mathcal{S}_c(\theta) := \arg\min_{x \in \mathbb{R}^n} \{ f(x, \theta) + g(x, \theta) \} \neq \varnothing$;
- $\nabla_x f$ is jointly locally Lipschitz and verifies $\|\nabla_x f(x, \theta) - \nabla_x f(x', \theta)\| \leqslant L \|x - x'\|$;
- $G^\gamma(z, \theta) := \operatorname{prox}_{\gamma g(\cdot, \theta)}(z)$ is jointly locally Lipschitz ($\gamma > 0$).

Composite Optimality Conditions

Fixed-point reformulation

$$x \text{ solution} \iff x - \underbrace{G^\gamma}_{\text{backward}} \Big(\underbrace{x - \gamma\big(\nabla_x f(x,\theta) + \varepsilon x\big)}_{\text{forward}}, \theta \Big) = 0$$

- conservative differentiation and nonsmooth IFT
- $J_{\nabla_x f} (\simeq [\partial_x \nabla_x f \ \ \partial_\theta \nabla_x f])$ and $J_{G^\gamma} (\simeq [\partial_x G^\gamma \ \ \partial_\theta G^\gamma])$
- complicated equations involving "$\partial \mathrm{prox}_{g(\cdot,\theta)} / \partial \theta$"

Requires stronger assumptions + one for "$\partial_\theta G^\gamma \simeq \partial \mathrm{prox}_{g(\cdot,\theta)} / \partial \theta$".

Still issues of $H_0^\dagger$ vs $\lim_{\varepsilon \to 0} \dots$ and which elements reached.

Composite Optimality Conditions

Fixed-point reformulation

$$x \text{ solution} \iff x - \underbrace{G^\gamma}_{\text{backward}}\Big(\underbrace{x - \gamma\big(\nabla_x f(x,\theta) + \varepsilon x\big)}_{\text{forward}}, \theta\Big) = 0$$

- conservative differentiation and nonsmooth IFT
- $J_{\nabla_x f}\big(\simeq [\partial_x \nabla_x f \;\; \partial_\theta \nabla_x f]\big)$ and $J_{G^\gamma}\big(\simeq [\partial_x G^\gamma \;\; \partial_\theta G^\gamma]\big)$
- complicated equations involving "$\partial \mathrm{prox}_{g(\cdot,\theta)}/\partial\theta$"

Requires stronger assumptions + one for "$\partial_\theta G^\gamma \simeq \partial \mathrm{prox}_{g(\cdot,\theta)}/\partial\theta$".

Still issues of $H_0^\dagger$ vs $\lim_{\varepsilon \to 0} \dots$ and which elements reached.