Nonsmooth differentiation of parametric fixed points

Edouard Pauwels (Toulouse School of Economics, France)

joint work with

Jérôme Bolte, Tâm Lê, Antonio Silveti-Falls, Samuel Vaiter

Sigma-Mode (January, 2024)

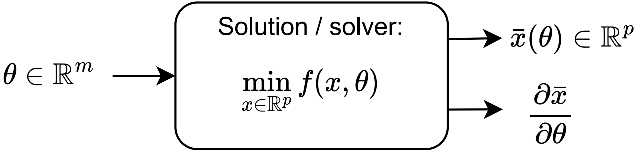
Parametric optimization, differentiation of solution mapping

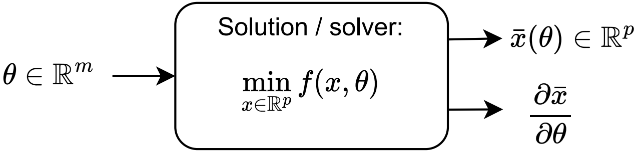
Static convex optimization:
$f, g_1, \ldots, g_q \colon \mathbb{R}^p \to \mathbb{R}$, convex

$$\bar{x} \in \arg\min_{x \in \mathbb{R}^p} \quad f(x) \qquad\qquad \text{s.t.} \quad g_i(x) \leq 0, \qquad i = 1, \ldots, q.$$

Parametric convex optimization:
$f, g_1, \ldots, g_q \colon \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$, continuous, convex in first variable

$$\bar{x}(\theta) \in \arg\min_{x \in \mathbb{R}^p} \quad f(x, \theta) \qquad\qquad \text{s.t.} \quad g_i(x, \theta) \leq 0, \qquad i = 1, \ldots, q.$$



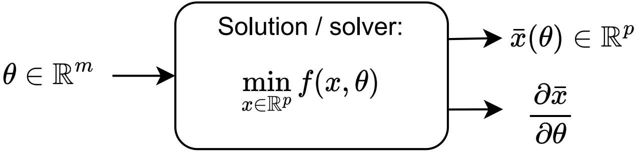


Sensitivity analysis: stability of minimizers under perturbation.
Bonnans, Fiacco, Jittorntrum, Robinson, Shapiro, . . .

Bilevel optimization: $\arg\min_x f(x; \theta)$ as a constraint.
Bracken, Dempe, Luo, McGill, Pang, Stackelberg . . .
Renewed interest in ML/signal: hyperparameter tuning, meta learning.

Ablin, Blondel, Chambolle, Duvenaud, Moreau, Pedregosa, Pock, Lorraine, Vaiter . . .
Abbeel, Finn, Franceschi, Levine, Pontil, Rajeswaran, Salzo . . .

Differentiable programming: $\bar{x}$ as an elementary component of a larger model.
OptNet, Deep Equilibrium networks (DEQ), cvxpylayers, QPlayers . . .



Optimality condition: $\nabla_x f(\bar{x}(\theta), \theta) = 0$ and many extensions, ...

Algorithm: $x_{k+1}(\theta) = F(x_k(\theta), \theta) \to \bar{x}(\theta)$.

Fixed point formulation: $\bar{x}(\theta) = F(\bar{x}(\theta), \theta)$.

Roadmap: 1/ the story in the smooth setting 2/ recent nonsmooth extensions.

Plan

A result from Gilbert (92, simplified)

Optimization Methods and Software, 1992, Vol. 1, pp.13–21
Reprints available directly from the publisher
Photocopying permitted by license only

AUTOMATIC DIFFERENTIATION AND ITERATIVE PROCESSES*

JEAN CHARLES GILBERT
INRIA, Domaine de Voluceau, Rocquencourt, Le Chesnay Cedex, France.

Proposition: $F\colon \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}^p$, C^1, $F(\cdot,\theta)$ $\rho < 1$ Lipschitz, for all $\theta \in \mathbb{R}^m$.
Assume $x_0\colon \mathbb{R}^m \to \mathbb{R}^p$ is C^1 and consider the recursion

$$x_{k+1}(\theta) = F(x_k(\theta),\theta) \qquad \Rightarrow \qquad x_k(\theta) \underset{k\to\infty}{\to} \bar{x}(\theta), \qquad \frac{\partial x_k}{\partial \theta} \underset{k\to\infty}{\to} \frac{\partial \bar{x}}{\partial \theta}.$$

Proof sketch: Banach fixed point theorem: $F(\cdot,\theta)$ contraction, $x_k(\theta) \to \bar{x}(\theta)$
Implicit functions theorem: $I - J_x F(\bar{x}(\theta),\theta)$ invertible, $\bar{x}$ is C^1.

$$\bar{x}(\theta) = F(\bar{x}(\theta),\theta) \qquad \frac{\partial \bar{x}}{\partial \theta} = J_x F(\bar{x}(\theta),\theta)\frac{\partial \bar{x}}{\partial \theta} + J_\theta F(\bar{x}(\theta),\theta)$$

$$x_{k+1}(\theta) = F(x_k(\theta),\theta) \qquad \frac{\partial x_{k+1}}{\partial \theta} = J_x F(x_k(\theta),\theta)\frac{\partial x_k}{\partial \theta} + J_\theta F(x_k(\theta),\theta)$$

Derivative of fixed point $\sim$ fixed point of derivative.
$M \to J_x F(\bar{x}(\theta),\theta) M + J_\theta F(\bar{x}(\theta),\theta)$ contraction ($\|J_x F(x,\theta)\|_{\mathrm{op}} \leq \rho$).
Continuity argument.

Examples: studied in details by Mehmood and Ochs (2020)

Assumption: $f\colon \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$, C^2 and $\mu > 0$ strongly convex and L-Lipschitz gradient with respect to the first variable.

$$\bar{x}(\theta) = \arg\min_x f(x, \theta)$$

Gradient descent: $0 < \alpha < 2/L$

$$F(x, \theta) = x - \alpha \nabla_x f(x, \theta) \qquad C^1, \quad \rho = \max\{1 - \alpha\mu, \alpha L - 1\}$$

Polyak's heavy ball: $0 \leq \beta < 1$, $0 < \alpha < 2(1 + \beta)/L$

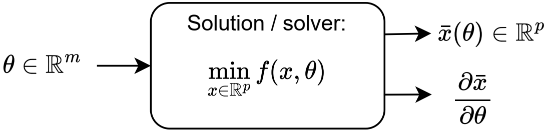
$$F(x, z, \theta) = (x - \alpha \nabla_x f(x, \theta) + \beta(x - z), x) \qquad C^1$$

Not a contraction: but spectral radius of $J_{x,z}F < 1$ (*e.g.* Polyak's book).
Change metric: local contraction after a linear change of variable.

Actual result from Gilbert:

- Spectral radius of Jacobian < 1 at fixed point (in a neighborhood by continuity).
- Assume iteration converge

Plan



- f may not be differentiable (Lasso hyperparameter tuning, learning TV regularizer).
- Algorithms / optimality condition involving projections / prox operators (cvxpylayers).
- Already implemented (equilibrium nets, opt layers, `TensorFlow`, `PyTorch`, `Jax`).

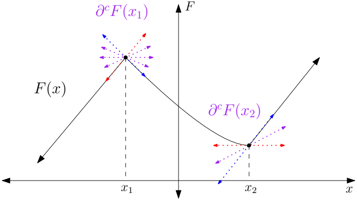
Algorithmic differentiation:

- differentiate solutions $\subset$ differential calculus $\sim$ algorithmic differentiation.
- nonsmooth $\rightarrow$ generalized derivatives.

Clarke's generalized derivatives: $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ locally Lipschitz (Rademacher, differentiable a.e.).

$$\begin{aligned} &\mathrm{Jac}^c\, F(x) \\ = \;&\mathrm{conv}\left\{\lim \mathrm{Jac}\, F(x_k) : x_k \rightarrow x, k \rightarrow +\infty\right\} \end{aligned}$$

$\mathrm{Jac}^c\, F \colon \mathbb{R}^n \rightrightarrows \mathbb{R}^{n \times m}$. $m = 1$: subdifferential $\partial^c F$



$g_1, g_2 \colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz, then $\partial^c(g_1 + g_2) \subset \partial^c g_1 + \partial^c g_2$.

- Equality if g_1 and g_2 are convex or C^1.
- No equality in general: $g \colon x \mapsto |x|$

$$\partial^c(g - g) = \partial^c(x \mapsto 0) = \{0\} \subset \quad \partial^c(g) + \partial^c(-g) = \quad \begin{cases} \{0\} & \text{if } x \neq 0 \\ [-2, 2] & \text{if } x = 0 \end{cases}.$$

- Take $f \colon \mathbb{R}^p \to \mathbb{R}$ Lipschitz, composition of elementary Lipschitz blocks $g_1, \ldots, g_L$

$$f = g_L \circ \ldots \circ g_1$$

- $\mathrm{autodiff}\, f \colon \mathbb{R}^p \to \mathbb{R}^p$, **formal chain rule:** a selection in the set valued field

$$\mathrm{Jac}^c g_L \circ \ldots \circ \mathrm{Jac}^c g_1 \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p \qquad \neq \quad \partial^c f \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$$

Conservative gradients (Bolte, Pauwels 2020, long story, long history)

Definition [Conservative gradient] :
$f\colon \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz, $D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, closed graph, non empty, locally bounded,
For any Lipschitz curve $\gamma\colon [0,1] \mapsto \mathbb{R}^p$

$$\frac{d}{dt} f(\gamma(t)) = \langle v, \dot{\gamma}(t) \rangle \qquad \forall v \in D(\gamma(t)), \qquad \text{a.e.} \quad t \in [0,1]$$

f is path-differentiable $\Leftrightarrow$ $\exists D$ conservative for f (could be many) $\Leftrightarrow$ $\partial^c f$ conservative.
Conservative Jacobians defined similarly.

Path-differentiability generic in applications:
semi-algebraic (or tame) $\Rightarrow$ $\mathrm{Jac}^c f$ is conservative.

Chain rule: $g_1, \ldots, g_L$ path-differentiable, conservative Jacobians $D_1, \ldots, D_L$, then $D_L \circ \ldots \circ D_1$ is conservative for $f = g_L \circ \ldots \circ g_1$.
$\mathrm{autodiff}\, f$ is a selection in a conservative gradient.

Optimization: Generically, as $\alpha_k \to 0$ (under appropriate mild assumptions).

$$\begin{array}{rcll} \theta_{k+1} & = & \theta_k - \alpha_k \,\mathrm{autodiff}\, f(\theta_k) & \text{selection in a conservative gradient} \\ \mathrm{dist}(\theta_k, \mathrm{crit}_f) & \underset{k\to\infty}{\to} & 0 & \mathrm{crit}_f = \{\theta,\, 0 \in \partial^c f(\theta)\} \end{array}$$

Implicit differentiation of fixed points (Bolte, Le, Pauwels, Silvetti, 2021)

$F : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}^m$ Lipschitz and $\bar{x} = F(\bar{x}, \theta)$

Classical implicit differentiation:

F **smooth,** assume

$[A, B] = \mathrm{Jac}\, F(\bar{x}, \theta), \quad I - A$ invertible.

$\bar{x} : U \to \mathbb{R}^p$, **smooth locally:**

$$F(\bar{x}(\theta), \theta) = \bar{x}(\theta).$$

Implicit **jacobian** of $\bar{x}$:

$$\theta \to (I - A)^{-1} B : [A, B] = \mathrm{Jac}\, F(\bar{x}(\theta), \theta).$$

Nonsmooth implicit differentiation:

F **path-differentiable**, assume

$\forall [A\ B] \in \mathrm{Jac}^c F(\bar{x}, \theta),\ I - A$ invertible.

$\bar{x} : U \to \mathbb{R}^p$, **path-differentiable:**

$$F(\bar{x}(\theta), \theta) = \bar{x}(\theta).$$

Implicit **conservative jacobian** for $\bar{x}$:

$$\theta \rightrightarrows \left\{ (I - A)^{-1} B : [A\ B] \in \mathrm{Jac}^c F(\bar{x}(\theta), \theta) \right\}$$

Invertibility: $F(\cdot, \theta)$, $\rho < 1$ Lipschitz, $\|A\|_{\mathrm{op}} \leq \rho, \forall [A\ B] \in \mathrm{Jac}^c F(\bar{x}(\theta), \theta)$.
Extends to any D conservative for F, in place of $\mathrm{Jac}^c F$.

$F : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}^p$, algorithmic recursion, $x_0(\theta) \in \mathbb{R}^p$

$$x_{k+1}(\theta) = F(x_k(\theta), \theta).$$

For all θ, $F(\cdot, \theta)$ is ρ Lipschitz, $\rho < 1$: $$x_k(\theta) \underset{k\to\infty}{\to} \bar{x}(\theta).$$

Classical asymptotics (Gilbert 92):

F **smooth**.

Forward **jacobian** propagation:

$$\begin{aligned} \mathrm{Jac}\, x_{k+1}(\theta) &= A \mathrm{Jac}\, x_k(\theta) + B \\ [A, B] &= \mathrm{Jac}\, F(x_k(\theta), \theta) \end{aligned}$$

Limiting **jacobian.**

$$\mathrm{Jac}\, x_k(\theta) \underset{k\to\infty}{\to} \mathrm{Jac}\, \bar{x}(\theta)$$

Nonsmooth unrolling :

F **path-differentiable**.

Conservative jacobian propagation:

$$D_{k+1}(\theta) = \Big\{ A D_k(\theta) + B \quad [A, B] \in \mathrm{Jac}^{\,c} F(x_k(\theta), \theta) \Big\}$$

Limiting **conservative jacobian:**

$$D_k(\theta) \underset{k\to\infty}{\to} \bar{D}(\theta) \quad \text{conservative for } \bar{x}$$

Remark: $\|A\|_{\mathrm{op}} \leq \rho, \forall [A\ B] \in \mathrm{Jac}^c\, F(\bar{x}(\theta), \theta)$, crucial for set valued fixed point. Extends to any D conservative for F, in place of $\mathrm{Jac}^c\, F$.

Two different conservative Jacobians:

Implicit differentiation:

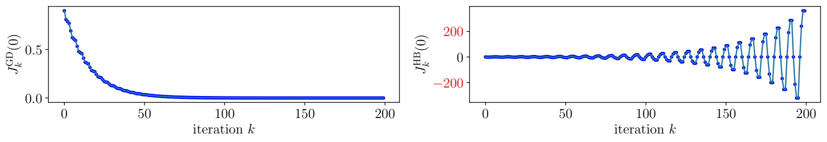
$$D_{\text{imp}}\,\bar{x}(\theta) = \{M,\ \exists[A,B] \in \operatorname{Jac}^c F(\bar{x}(\theta),\theta),\ M = AM + B\}$$

Iterative differentiation: unique $\bar{D}(\theta)$ such that

$$\forall M \in \bar{D}(\theta),\ \forall[A,B] \in \operatorname{Jac}^c F(\bar{x}(\theta),\theta),\ AM + B \in \bar{D}(\theta)$$

Examples: f strongly convex, Lipschitz path-differentiable gradient (not C^2)

Gradient descent | Heavy-Ball



Operator norm condition cannot be extended to spectral radius.

Application to monotone inclusions (Bolte, Pauwels, Silvetti-Falls 2023)

For all θ, $\mathcal{A}_\theta = \mathcal{A}(\cdot,\theta) : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ and $\mathcal{B}_\theta = \mathcal{B}(\cdot,\theta) : \mathbb{R}^p \to \mathbb{R}^p$ maximal monotone.

$$0 \in \mathcal{A}(\cdot,\theta) + \mathcal{B}(\cdot,\theta) \qquad \text{solution set non-empty}$$

Assumption: For all $\gamma > 0$, $\mathcal{R}_{\gamma\mathcal{A}_\theta} = (I + \gamma\mathcal{A}_\theta)^{-1}$ and $\mathcal{B}$ Lipschitz and path-differentiable, jointly in (x,θ).

$$F(x,\theta) := \mathcal{R}_{\gamma\mathcal{A}_\theta}(x - \gamma\mathcal{B}_\theta(x)) \qquad \text{path-differentiable jointly in } (x,\theta)$$

Theorem: Assume that $\mathcal{A}_\theta$ or $\mathcal{B}_\theta$ is strongly monotone.
Then for small γ, F is $\rho < 1$ Lipschitz and for any $[A,B] \in \mathrm{autodiff} F$, $\|A\|_{\mathrm{op}} \leq \rho$.

Applications: f_θ, g_θ, convex, lower semi continuous, proper, value in $\mathbb{R} \cup \{+\infty\}$.
Forward-backward: f_θ Lipschitz gradient, f_θ or g_θ strongly convex.

$$\min_{x\in\mathbb{R}^p} \quad f_\theta(x) + g_\theta(x) \qquad\qquad 0 \in \nabla_x f_\theta + \partial_x g_\theta$$

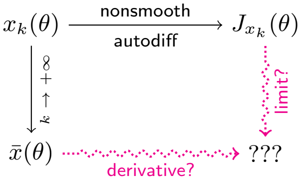
Primal-dual: f_θ and g_θ strongly convex.

$$\min_{x\in\mathbb{R}^p} g_\theta(x) + \max_{y\in\mathbb{R}^q} \langle K_\theta x, y\rangle - f_\theta(y) \qquad \begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \partial g_\theta & 0 \\ 0 & \partial f_\theta \end{pmatrix} + \begin{pmatrix} 0 & K_\theta \\ -K_\theta & 0 \end{pmatrix}$$

Plan

Conclusion

$$F(\bar{x}(\theta), \theta) = \bar{x}(\theta) \qquad\qquad x_{k+1}(\theta) = F(x_k(\theta), \theta)$$



- Extend implicit and iterative differentiation to the nonsmooth setting.
- Conservative Jacobians.
- $\sim$ smooth setting:
  autodiff, convergence, strong convexity, optimization ...

References:

- Conservative set valued fields, automatic differentiation, stochastic gradient method and deep learning. Bolte, Pauwels. Math. Prog. 2020.
- Nonsmooth Implicit Differentiation for Machine Learning and Optimization. Bolte, Le, Pauwels, Silveti-Falls Neurips 2021.
- Automatic differentiation of nonsmooth iterative algorithms. Bolte, Pauwels, Vaiter Neurips 2022.
- Differentiating Nonsmooth Solutions to Parametric Monotone Inclusion Problems. Bolte, Pauwels, Silveti-Falls SIOPT, 2023

Thanks.