Sensitivity Analysis for Mirror-Stratifiable Convex Functions

Jérôme MALICK

CNRS, Laboratoire Jean Kuntzmann, Grenoble

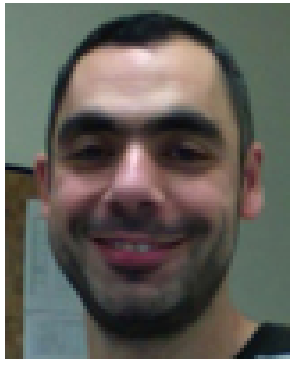
Huitièmes journées franco-chiliennes d'optimisation (JFCO)

July 2017 – Toulouse

Talk based on joint work

Gabriel Peyré

(CNRS, ENS Ulm)



Jalal Fadili

(Normandie Université - ENSICAEN)

J. Fadili, J. Malick, and G. Peyré

Sensitivity Analysis for Mirror-Stratifiable Convex Functions

to be submitted to SIAM Journal on Optimization, 2017

...partly inspired by nice ideas in "old" joint work

Aris Daniilidis

(Univ. Chile)

A. Daniilidis, W. Hare, and J. Malick

Geometrical interpretation of predictor-corrector algorithms in structured optimization

Optimization, volume 55(5), 2006

A. Daniilidis, J. Malick, and H. Sendov

On the structure of locally symmetric manifolds

Journal of Convex Analysis, volume 22(2), 2014

Outline

Outline

Motivating example

General situation in data analysis

recovering $x_0 \in \mathbb{R}^N$ from noisy observations

$$y = \Phi x_0 + w$$

Operator $\Phi : \mathbb{R}^N \to \mathbb{R}^P$ with $P < N$ (degradation operator or design matrix...)

number of observations (much) smaller than the ambient space

Motivating example

General situation in data analysis

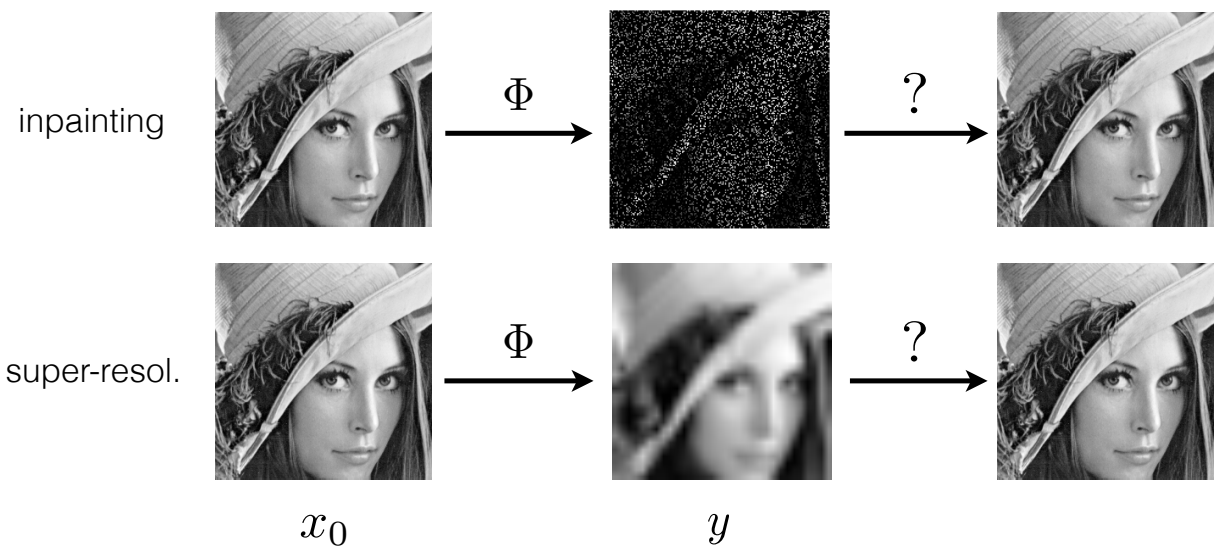
recovering $x_0 \in \mathbb{R}^N$ from noisy observations

$$y = \Phi x_0 + w$$

Operator $\Phi : \mathbb{R}^N \to \mathbb{R}^P$ with $P < N$ (degradation operator or design matrix...)

number of observations (much) smaller than the ambient space

Example in image processing (just to fix ideas)



Inverse problems

Ill-posed inverse problem: recover x_0 from $y = \Phi x_0 + w$

- Assume x_0 has a sort of "low-complexity"
- Example: sparsity of entries, of blocks, of jumps, of spectra...

Inverse problems

Ill-posed inverse problem: recover x_0 from $y = \Phi x_0 + w$

- Assume x_0 has a sort of "low-complexity"
- Example: sparsity of entries, of blocks, of jumps, of spectra...

Regularized inverse problems

$$\min_{x \in \mathbb{R}^N} \quad \frac{1}{2}\|y - \Phi x\|^2 \; + \; \lambda \, R(x)$$

data-fidelity prior regularization

- R promotes low-complexity to solutions (similar to the one of x_0)
- $\lambda > 0$ controls trade-off (depends on noise level $\|w\|$ and $R(x_0)$)

Questions: for a solution $x(y, \lambda)$

Under which conditions, can we guarantee

1. ℓ_2-recovery $\quad \|x(\lambda, y) - x_0\| = O(\|w\|^\alpha)$?
2. model recovery the low-complexity of $x(y, \lambda)$ coincides with the one of x_0 ? (when w small)

Example: compressed sensing

- Recover a sparse vector $x_0 \in \mathbb{R}^N$ from noisy observation $y = \Phi x_0 + w \in \mathbb{R}^P$
- Low-complexity: support of x_0 (= nonzeros entries $x_{0\,i}$) is small
- Regularization: $R = \|\cdot\|_1$ (= convex hull of restricted $\|\cdot\|_0 = \#\,\text{support}(\cdot)$)
- ℓ_1-regularized least-squares problem (LASSO, LARS,...)

$$\min_{x \in \mathbb{R}^N} \quad \frac{1}{2}\|y - \Phi x\|^2 + \lambda \|x\|_1$$

Example: compressed sensing

- Recover a sparse vector $x_0 \in \mathbb{R}^N$ from noisy observation $y = \Phi x_0 + w \in \mathbb{R}^P$
- Low-complexity: support of x_0 (= nonzeros entries $x_{0\,i}$) is small
- Regularization: $R = \|\cdot\|_1$ (= convex hull of restricted $\|\cdot\|_0 = \#\,\text{support}(\cdot)$)
- ℓ_1-regularized least-squares problem (LASSO, LARS,...)

$$\min_{x \in \mathbb{R}^N} \quad \frac{1}{2}\|y - \Phi x\|^2 + \lambda \|x\|_1$$

- Many answers to recovery questions [Fuchs '04] [Grasmair '10] [Vaiter '14]...
- For Φ gaussian [Candès *et al* '05] [Dossal *et al* '11]

"we have recovery when P is large enough"

ℓ_2-recovery when $P = \Omega(\|x_0\|_0 \log(N/\|x_0\|_0))$

model recovery when $P = \Omega(\|x_0\|_0 \log N)$

- What happens if P is not large enough ?

What happens in degenerate cases ?

- no idea ?! all existing results assume some kind of non-degeneracy
- In particular: the previous ones + [Lewis '06] (general sensitivity) + [Bach '08] (trace-norm recovery) + [Hare-Lewis '10] (identification) + [Candes-Recht '11] (recovery) + [Vaiter *et al* '15] (partly-smooth recovery) + [Liang *et al* '16] (identification of proximal spitting), and many others...
- However real-life problems are often degenerate (e.g. medical imaging)

What happens in degenerate cases ?

- no idea ?! all existing results assume some kind of non-degeneracy
- In particular: the previous ones + [Lewis '06] (general sensitivity) + [Bach '08] (trace-norm recovery) + [Hare-Lewis '10] (identification) + [Candes-Recht '11] (recovery) + [Vaiter *et al* '15] (partly-smooth recovery) + [Liang *et al* '16] (identification of proximal spitting), and many others...
- However real-life problems are often degenerate (e.g. medical imaging)
- Position of our work:

  – known results

  non-degenerate problems $\Longrightarrow$ (exact) recovery

  – in this talk

  general problems $\Longrightarrow$ some recovery ?

  Yes ! for some structured regularizations

  (that we called mirror-stratifiable)

Outline

Recalls on stratifications

A stratification of a set $D \subset \mathbb{R}^N$ is a finite partition $\mathcal{M} = \{M_i\}_{i \in I}$

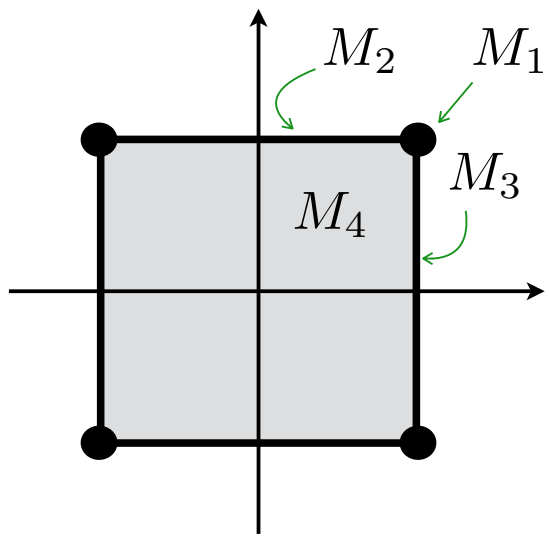
$$D = \bigcup_{i \in I} M_i$$

with "strata" which fit nicely:

$$M \cap \mathrm{cl}(M') \neq \emptyset \implies M \subset \mathrm{cl}(M')$$

Example: $\mathbb{B}_\infty$ the unit ℓ_∞-ball in $\mathbb{R}^2$

a stratification with 9 (affine) strata



Recalls on stratifications

A stratification of a set $D \subset \mathbb{R}^N$ is a finite partition $\mathcal{M} = \{M_i\}_{i \in I}$

$$D = \bigcup_{i \in I} M_i$$

with "strata" which fit nicely:

$$M \cap \mathrm{cl}(M') \neq \emptyset \implies M \subset \mathrm{cl}(M')$$

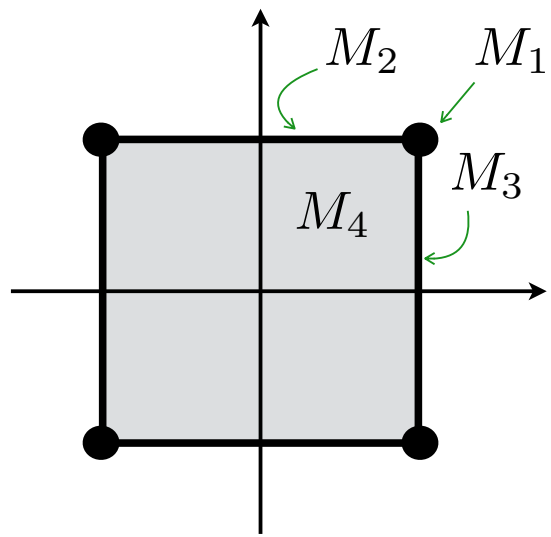
This entails a (partial) ordering $M \leqslant M'$

Example: $\mathbb{B}_\infty$ the unit ℓ_∞-ball in $\mathbb{R}^2$

a stratification with 9 (affine) strata

$$M_1 \leqslant M_2 \leqslant M_4$$

$$M_1 \leqslant M_3 \leqslant M_4$$



Mirror-stratifiable function: formal definition

A convex function $R\colon \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is mirror-stratifiable with respect to

- a (primal) stratification $\mathcal{M} = \{M_i\}_{i\in I}$ of $\mathrm{dom}(\partial R)$
- a (dual) stratification $\mathcal{M}^* = \{M_i^*\}_{i\in I}$ of $\mathrm{dom}(\partial R^*)$

if $\mathcal{J}_R$ has 2 properties

- $\mathcal{J}_R : \mathcal{M} \to \mathcal{M}^*$ is invertible with inverse $\mathcal{J}_{R^*}$

$$\mathcal{M}^* \ni M^* = \mathcal{J}_R(M) \iff \mathcal{J}_{R^*}(M^*) = M \in \mathcal{M}$$

- $\mathcal{J}_R$ is decreasing for the order relation $\leqslant$ between strata

$$M \leqslant M' \iff \mathcal{J}_R(M) \geqslant \mathcal{J}_R(M')$$

with the transfert operator $\mathcal{J}_R : \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ [Daniilidis-Drusvyatskiy-Lewis '13]

$$\mathcal{J}_R(S) = \bigcup_{x\in S} \mathrm{ri}(\partial R(x))$$

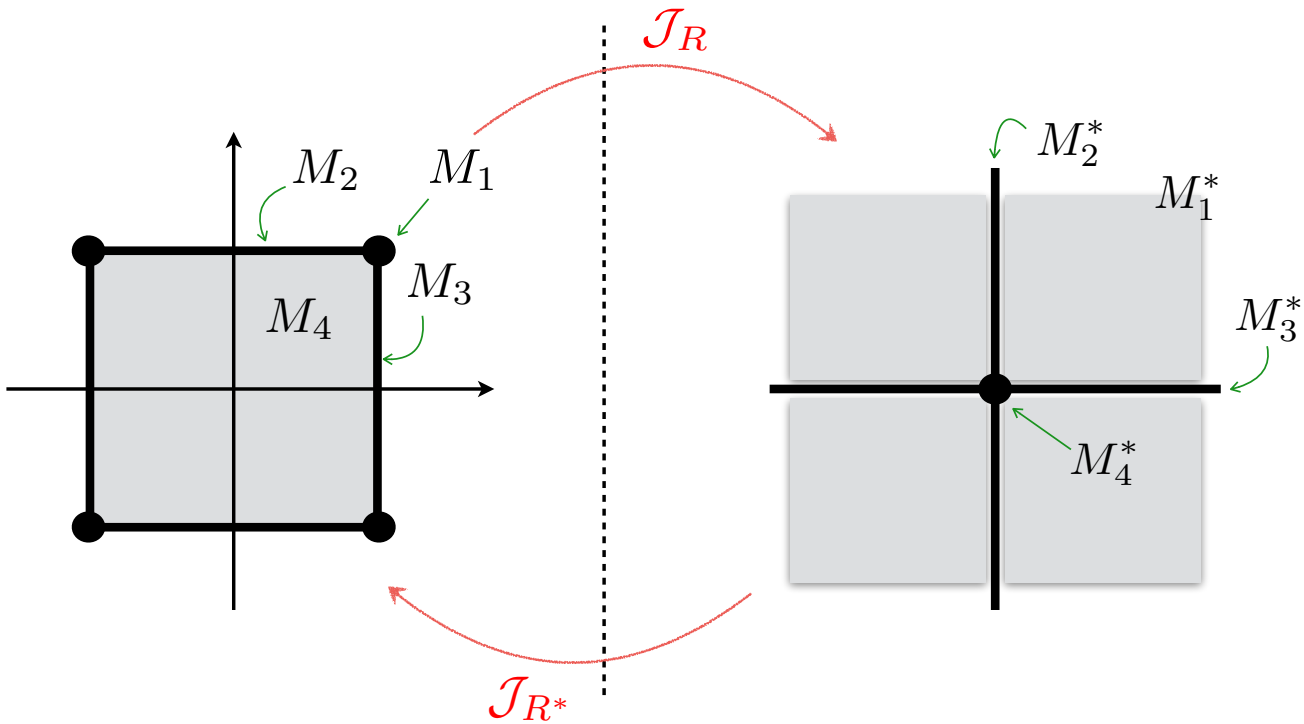
Mirror-stratifiable function: simple example

$$R = \iota_{\mathbb{B}_\infty} \qquad R^* = \|\cdot\|_1$$

$$(\operatorname{dom} R = \mathbb{B}_\infty) \qquad (\operatorname{dom} R^* = \mathbb{R}^N)$$

$$\mathcal{J}_R(M_i) = \bigcup_{x \in M_i} \operatorname{ri} \partial R(x) = \operatorname{ri} N_{\mathbb{B}_\infty}(x) = M_i^*$$

$$M_i = \operatorname{ri} \partial \|x\|_1 = \bigcup_{x \in M_i^*} \operatorname{ri} \partial R(x) = \mathcal{J}_R^*(M_i^*)$$



Mirror-stratifiable functions are everywhere !

Definition is formal, assumptions look strong... however :

All the regularizers routinely used in machine learning or image processing are mirror-stratifiable

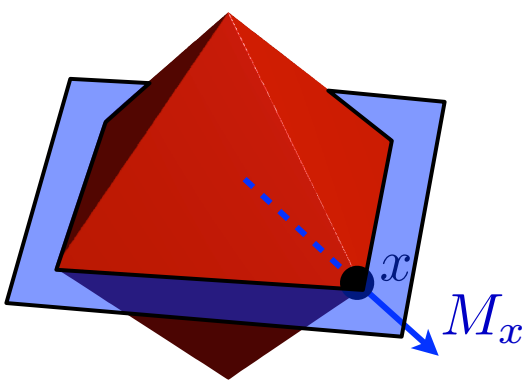
Mirror-stratifiable functions are everywhere !

Definition is formal, assumptions look strong... however :

All the regularizers routinely used in machine learning or image processing are mirror-stratifiable

Among others:

- $R = \|\cdot\|_1$ (and $\|\cdot\|_2{}^2$ or $\|\cdot\|_\infty$)



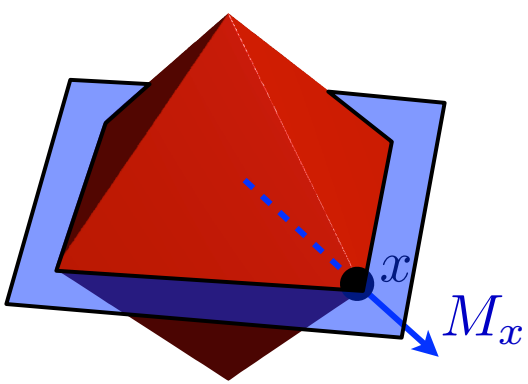
Mirror-stratifiable functions are everywhere !

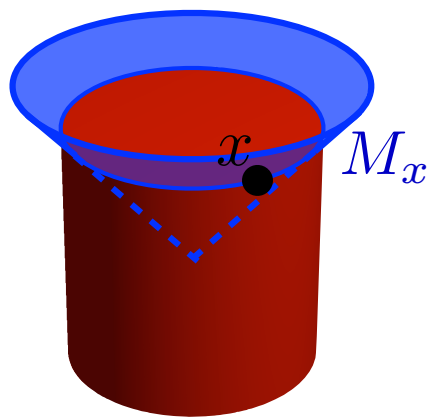
Definition is formal, assumptions look strong... however :

> All the regularizers routinely used in machine learning or image processing are mirror-stratifiable

Among others:

- $R = \|\cdot\|_1$ (and $\|\cdot\|_2{}^2$ or $\|\cdot\|_\infty$)
- nuclear norm (aka trace-norm) $R(X) = \sum_i |\sigma_i(X)| = \|\sigma(X)\|_1$





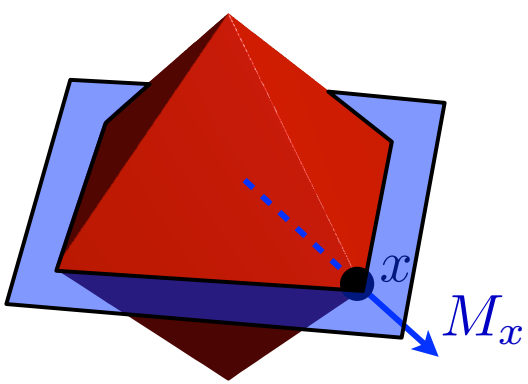
Mirror-stratifiable functions are everywhere !

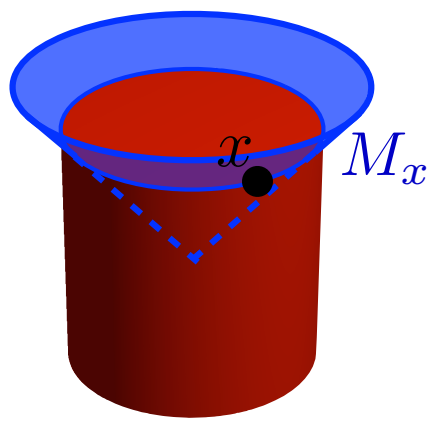
Definition is formal, assumptions look strong... however :

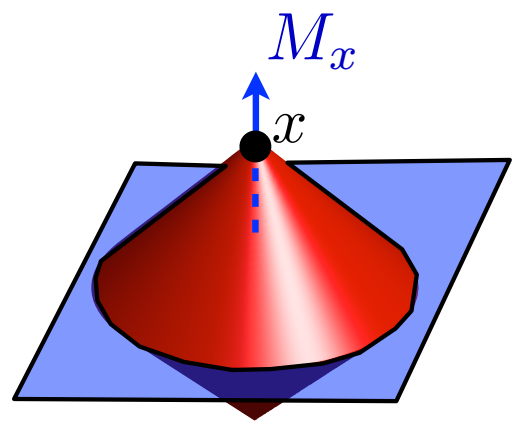
> All the regularizers routinely used in machine learning or image processing are mirror-stratifiable

Among others:

- $R = \|\cdot\|_1$ (and $\|\cdot\|_2{}^2$ or $\|\cdot\|_\infty$)
- nuclear norm (aka trace-norm) $R(X) = \sum_i |\sigma_i(X)| = \|\sigma(X)\|_1$
- group-ℓ_1 $R(x) = \sum_{b\in\mathcal{B}} \|x_b\|_2$ (e.g. $R(x) = |x_1| + \|x_{2,3}\|$)







Outline

Sensitivity of parametrized optimization problem

Parameterized composite optimization problem (smooth + nonsmooth)

$$\min_{x \in \mathbb{R}^N} \quad E(x, p) = F(x, p) + R(x),$$

Optimality condition for a primal-dual solution $(x^\star(p), u^\star(p))$

$$u^\star(p) = -\nabla F(x^\star(p), p) \ \in \ \partial R(x^\star(p))$$

Theorem (Enlarged activity)

Under mild assumptions ($E(\cdot, p_0)$ *has a unique minimizer* $x^\star(p_0)$ *and* E *is uniformly level-bounded in* x), *if* R *is mirror-stratifiable, then for* $p \sim p_0$,

$$M_{x^\star(p_0)} \leqslant M_{x^\star(p)} \leqslant \mathcal{J}_{R^*}(M^*_{u^\star(p_0)})$$

Sensitivity of parametrized optimization problem

Parameterized composite optimization problem (smooth + nonsmooth)

$$\min_{x \in \mathbb{R}^N} \quad E(x, p) = F(x, p) + R(x),$$

Optimality condition for a primal-dual solution $(x^\star(p), u^\star(p))$

$$u^\star(p) = -\nabla F(x^\star(p), p) \ \in \ \partial R(x^\star(p))$$

Theorem (Enlarged activity)

Under mild assumptions ($E(\cdot, p_0)$ *has a unique minimizer* $x^\star(p_0)$ *and* E *is uniformly level-bounded in* x), *if* R *is mirror-stratifiable, then for* $p \sim p_0$,

$$M_{x^\star(p_0)} \leqslant M_{x^\star(p)} \leqslant \mathcal{J}_{R^*}(M^*_{u^\star(p_0)})$$

In the non-degenerate case $\quad u^\star(p_0) \in \text{ri}\left(\partial R(x^\star(p_0))\right)$

$$M_{x^\star(p_0)} = M_{x^\star(p)} \quad \left(= \mathcal{J}_{R^*}(M^*_{u^\star(p_0)})\right)$$

we retrieve exactly the active strata ([Lewis '06] for partly-smooth functions)

First sensitivity result illustrated

Simple projection problem

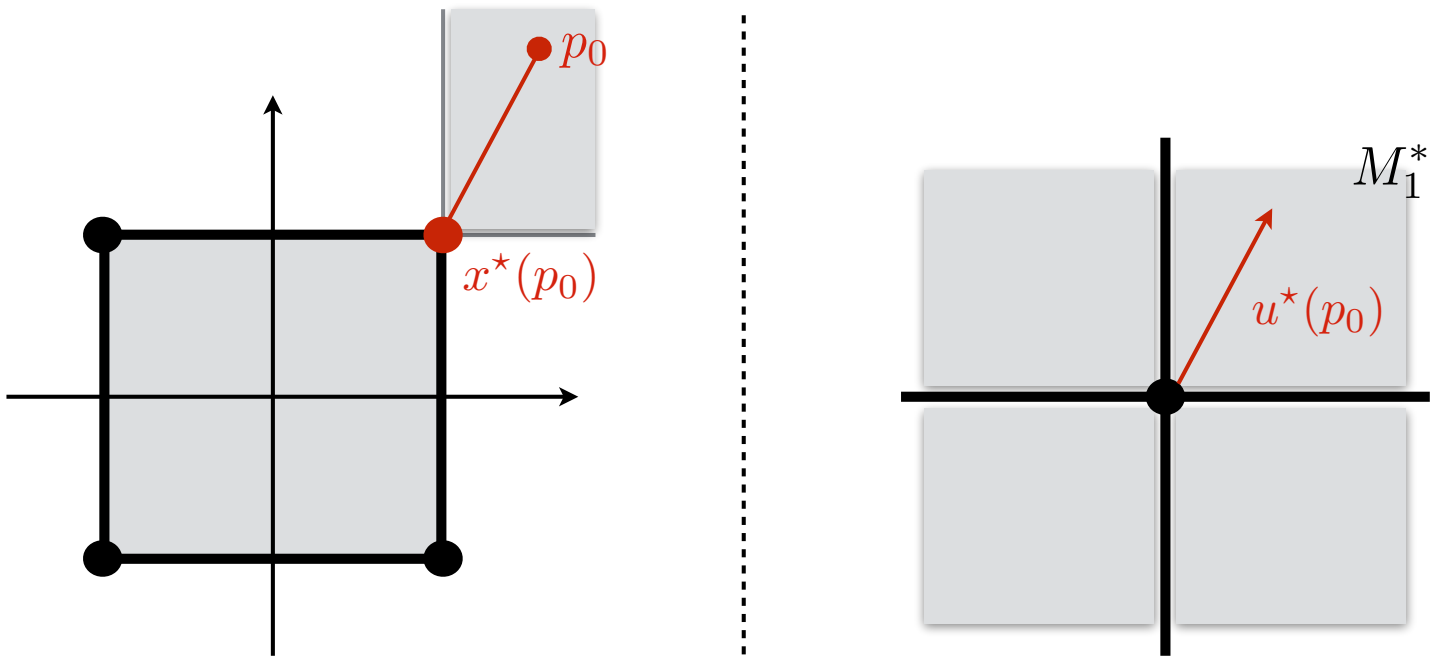
$$\begin{cases} \min & \frac{1}{2}\|x-p\|^2 \\ & \|x\|_\infty \leqslant 1 \end{cases} \qquad \begin{cases} \min & \frac{1}{2}\|u-p\|^2 + \|u\|_1 \\ & u \in \mathbb{R}^N \end{cases}$$

First sensitivity result illustrated

Simple projection problem

$$\begin{cases} \min & \frac{1}{2}\|x-p\|^2 \\ & \|x\|_\infty \leqslant 1 \end{cases} \qquad \begin{cases} \min & \frac{1}{2}\|u-p\|^2 + \|u\|_1 \\ & u \in \mathbb{R}^N \end{cases}$$

Non-degenerate case: $u^\star(p_0) = p_0 - x^\star(p_0) \in \text{ri } N_{\mathbb{B}_\infty}(x^\star(p_0))$



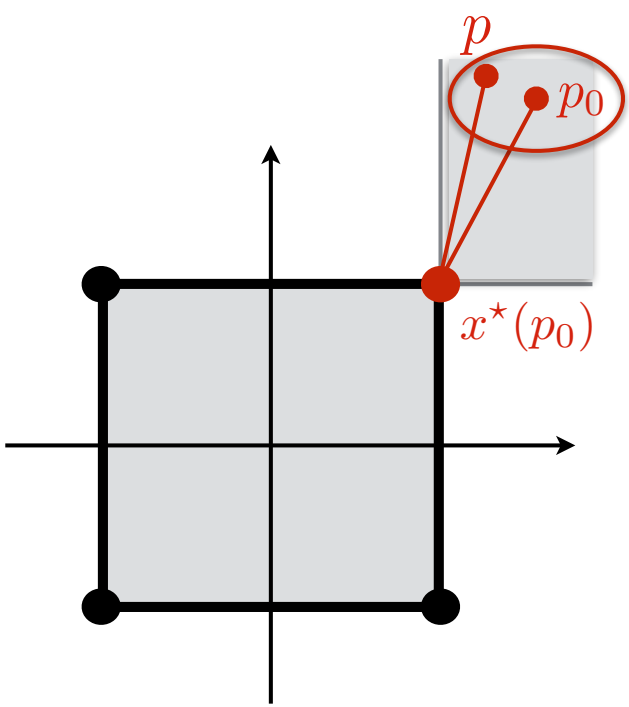
First sensitivity result illustrated

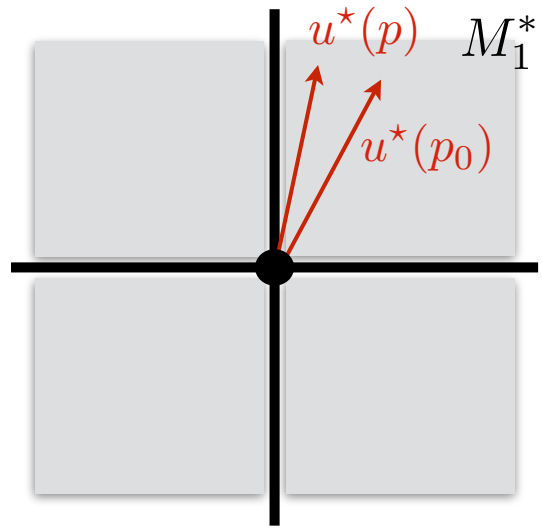
Simple projection problem

$$\begin{cases} \min & \frac{1}{2}\|x-p\|^2 \\ & \|x\|_\infty \leqslant 1 \end{cases} \qquad \begin{cases} \min & \frac{1}{2}\|u-p\|^2 + \|u\|_1 \\ & u \in \mathbb{R}^N \end{cases}$$

Non-degenerate case: $u^\star(p_0) = p_0 - x^\star(p_0) \in \text{ri } N_{\mathbb{B}_\infty}(x^\star(p_0))$

$$\Longrightarrow \quad M_1 = M_{x^\star(p_0)} = M_{x^\star(p)} \qquad \text{(in this case } x^\star(p) = x^\star(p_0)\text{)}$$



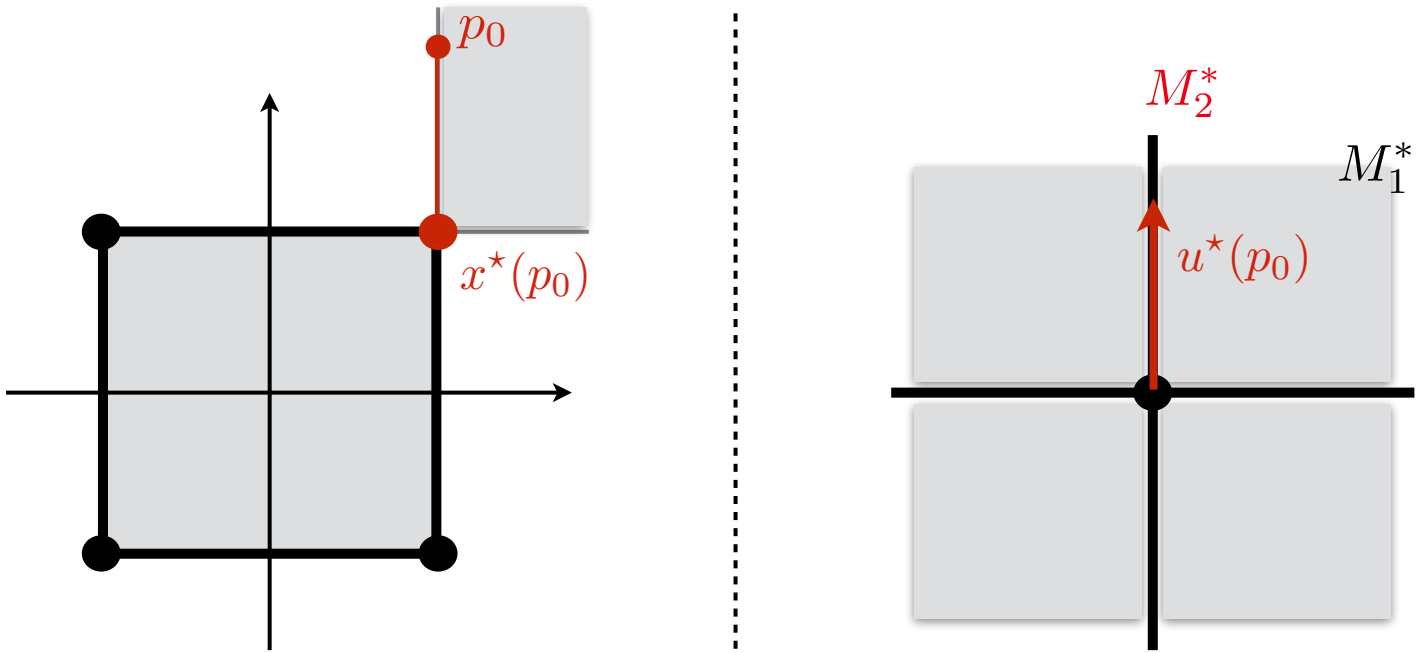


First sensitivity result illustrated

Simple projection problem

$$\begin{cases} \min & \frac{1}{2}\|x-p\|^2 \\ & \|x\|_\infty \leqslant 1 \end{cases} \qquad \begin{cases} \min & \frac{1}{2}\|u-p\|^2 + \|u\|_1 \\ & u \in \mathbb{R}^N \end{cases}$$

General case: $u^\star(p_0) = p_0 - x^\star(p_0) \in \cancel{ri}\, N_{\mathbb{B}_\infty}(x^\star(p))$



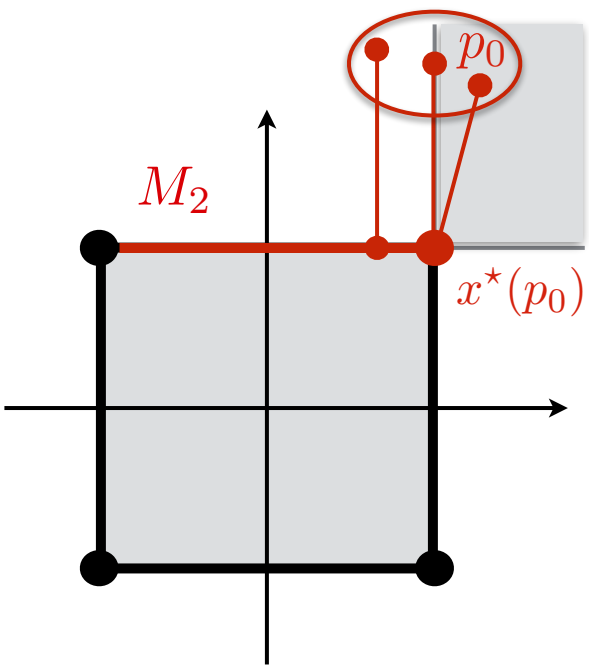
First sensitivity result illustrated

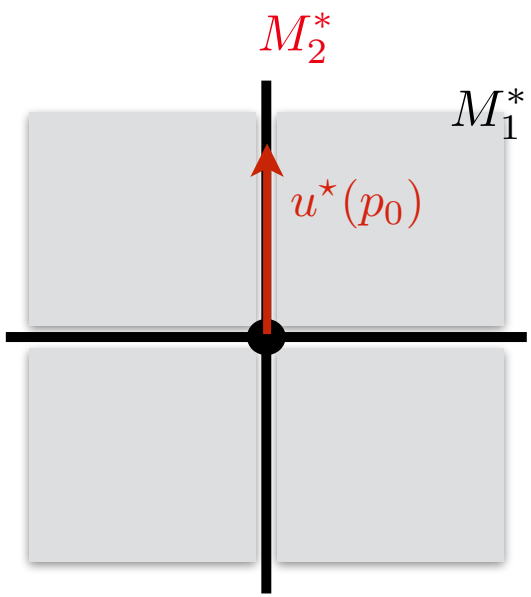
Simple projection problem

$$\begin{cases} \min & \frac{1}{2}\|x-p\|^2 \\ & \|x\|_\infty \leqslant 1 \end{cases} \qquad \begin{cases} \min & \frac{1}{2}\|u-p\|^2 + \|u\|_1 \\ & u \in \mathbb{R}^N \end{cases}$$

General case: $u^\star(p_0) = p_0 - x^\star(p_0) \in \cancel{n}\, N_{\mathbb{B}_\infty}(x^\star(p))$

$$\Longrightarrow \quad M_1 = M_{x^\star(p_0)} \leqslant M_{x^\star(p)} \leqslant \mathcal{J}_{R^*}(M^*_{u^\star(p_0)}) = M_2$$





Identification of proximal algorithms

Composite optimization problem (smooth + nonsmooth)

$$\min_{x \in \mathbb{R}^N} \quad f(x) + R(x)$$

Optimality condition $-\nabla f(x^\star) \in \partial R(x^\star)$

Proximal-gradient algorithm (aka forward-backward algorithm)

$$x_{k+1} = \mathrm{prox}_{\gamma_k R}\big(x_k - \gamma_k \nabla f(x_k)\big) \qquad (0 < \inf \gamma_k \leqslant \sup \gamma_k < 2/L)$$

Proximal mapping $\quad \mathrm{prox}_{\gamma_k R}(x) = \underset{y}{\mathrm{argmin}} \; R(y) + \dfrac{1}{2\gamma_k}\|y - x\|^2 \quad$ ($\exists$ explicit formula for $\|\cdot\|_1$)

Theorem (Enlarged identification)

Under basic assumptions, if R is mirror-stratifiable, then for k large

$$M_{x^\star} \leqslant M_{x_k} \leqslant \mathcal{J}_{R^*}\big(M^*_{-\nabla f(x^\star)}\big)$$

Identification of proximal algorithms

Composite optimization problem (smooth + nonsmooth)

$$\min_{x \in \mathbb{R}^N} \quad f(x) + R(x)$$

Optimality condition $-\nabla f(x^\star) \in \partial R(x^\star)$

Proximal-gradient algorithm (aka forward-backward algorithm)

$$x_{k+1} = \mathrm{prox}_{\gamma_k R}\big(x_k - \gamma_k \nabla f(x_k)\big) \qquad (0 < \inf \gamma_k \leqslant \sup \gamma_k < 2/L)$$

Proximal mapping $\mathrm{prox}_{\gamma_k R}(x) = \underset{y}{\mathrm{argmin}} \; R(y) + \dfrac{1}{2\gamma_k}\|y - x\|^2$ ($\exists$ explicit formula for $\|\cdot\|_1$)

Theorem (Enlarged identification)

Under basic assumptions, if R is mirror-stratifiable, then for k large

$$M_{x^\star} \leqslant M_{x_k} \leqslant \mathcal{J}_{R^*}(M^*_{-\nabla f(x^\star)})$$

In the non-degenerate case $-\nabla f(x^\star) \in \mathrm{ri}(\partial R(x^\star))$

we have exact identification $M_{x^\star} = M_{x_k}$ $\big(= \mathcal{J}_{R^*}(M^*_{-\nabla f(x^\star)})\big)$ [Liang *et al* 15]

Sensitivity of regularized inverse problems

Back to ill-posed inverse problem $\quad y = \Phi x_0 + w$

- Assume that x_0 is the unique minimizer of

$$\min_{x \in \mathbb{R}^N} \quad R(x) \qquad \text{s.t.} \quad \Phi x = \Phi x_0 \qquad (\text{“consistance”})$$

- Following [Vaiter *et al* '16], we introduce the smallest dual solution

$$q_0 = \operatorname*{argmin}_{q \in \mathbb{R}^P} \ \{\|q\|_2 \ : \ \Phi^* q \in \partial R(x_0)\} \qquad (\text{“minimum norm certificate”})$$

Solve the regularized inverse problem

$$\min_{x \in \mathbb{R}^N} \quad \frac{1}{2}\|y - \Phi x\|^2 + \lambda R(x)$$

- Can we track a solution $x^\star(\lambda, y)$ and x_k in the general case ?
- Proximal-gradient algorithm

$$x_{k+1} = \text{prox}_{\gamma_k R}\big(x_k - \gamma_k \Phi^*(\Phi x_k - y)\big)$$

Enlarged model recovery

Theorem (Sensitivity of regularized inverse problems)

If R is mirror-stratifiable, then for all (λ, y) such that

$$C_0\|y - y_0\| \leqslant \lambda \leqslant C_1$$

then $x^\star(\lambda, y)$ is localized

$$M_{x_0} \leqslant M_{x^\star(\lambda,p)} \leqslant \mathcal{J}_{R^*}(M^*_{\Phi^* q_0})$$

Theorem (Identification of proximal-gradient iterates)

Under previous assumptions, the prox-grad iterates satisfy, for k large,

$$M_{x_0} \leqslant M_{x_k} \leqslant \mathcal{J}_{R^*}(M^*_{\Phi^* q_0})$$

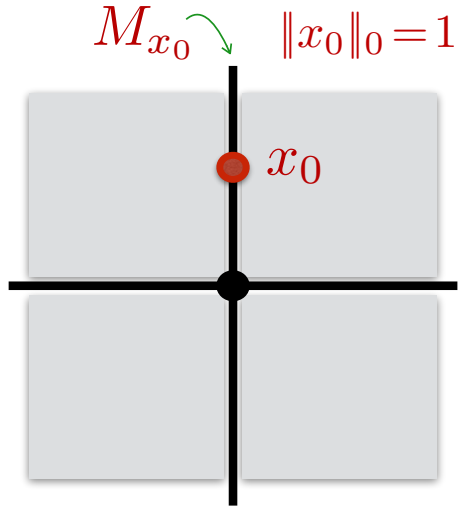
Comments:

- we track the strata when the perturbation $\|w\| = \|y - y_0\|$ is small
- $(x_k)_k$ does not converge to x_0, but still identifies strata

  $\longrightarrow$ interesting in practice when we have prior assumptions on the data x_0
- in numerical experiments, we measure $\quad \dim(\mathcal{J}_{R^*}(M^*_{\Phi^* q_0})) - \dim(M_{x_0})$

Outline

Experimental setting



Back to compressed sensing

- Recover a sparse x_0 from $y = \Phi x_0 + w$
- $M_{x_0} = \{z \in \mathbb{R}^N : \mathrm{supp}(z) \subset \mathrm{supp}(x_0)\}$
- Measure of low-complexity

$$\dim(M_{x_0}) = \#\ \mathrm{supp}(x_0) = \|x_0\|_0$$

Generate many random problems (out of the range of standard compressed sensing)

- Draw realizations (x_0, Φ, w) at random
  random $x_0 \in \{0,1\}^N$ and $\Phi \in \mathbb{R}^{P \times N}$ with gaussian entries
- Sizes: $N = 100$ $P = 50$ $\|x_0\|_0 \in \{1, ..., 30\}$
- Given N, P, the complexity $\|x_0\|_0$ too large to apply known results

Compute solutions to optimization problems

$$x(\lambda, y) \in \operatorname*{argmin}_{x \in \mathbb{R}^N} \frac{1}{2}\|y - \Phi x\|^2 + \lambda \|x\|_1$$

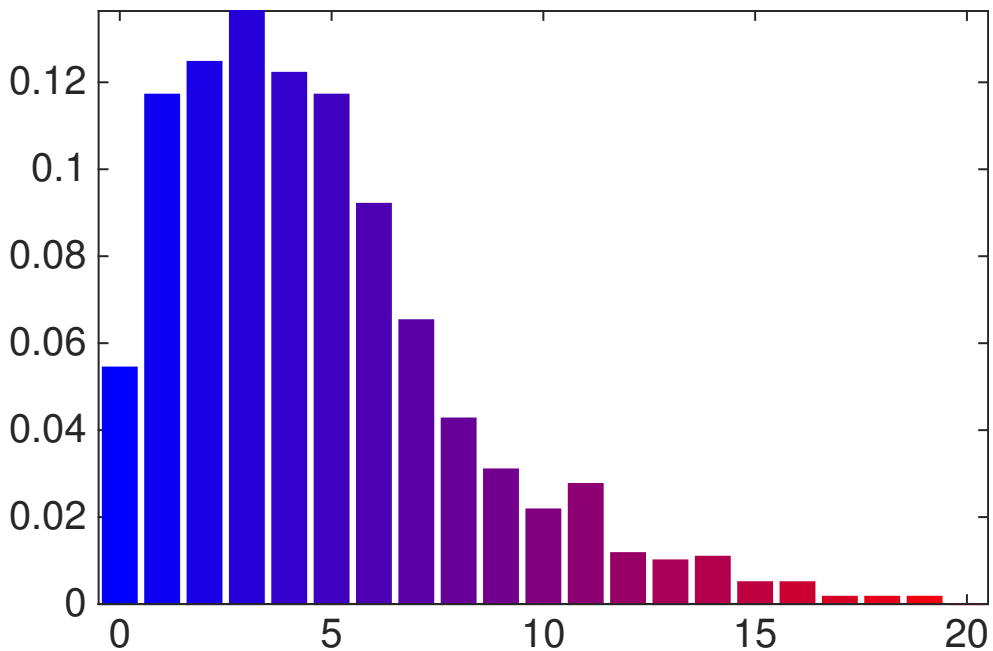
$$q_0 = \operatorname*{argmin}_{q \in \mathbb{R}^P} \{\|q\|_2 : \Phi^* q \in \partial R(x_0)\} \qquad \to \dim(\mathcal{J}_{R^*}(M^*_{\Phi^* q_0}))$$

Limits of existing results

Observe first that we do not have exact recovery in general

Histogram of the complexity index excess $\|x(\lambda, y)\|_0 - \|x_0\|_0$

(for all scenarios with fixed $\|x_0\|_0 = 10$)

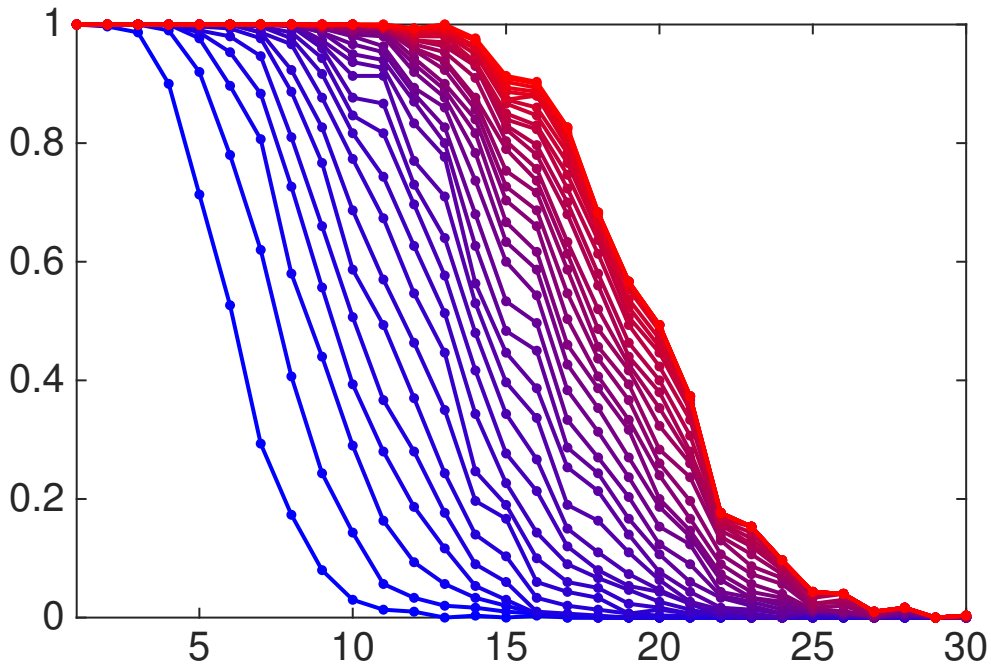


blue: exact recovery $\longrightarrow$ red: enlarged recovery

Illustration of the mirror-strat. sensitivity result

Plot of the percentage of scenarios (with respect to $\|x_0\|_0$ in horizontal axis) such that

$$\dim(\mathcal{J}_{R^*}(M^*_{\Phi^* q_0})) - \dim(M_{x_0}) \leqslant \delta$$



On the figure:

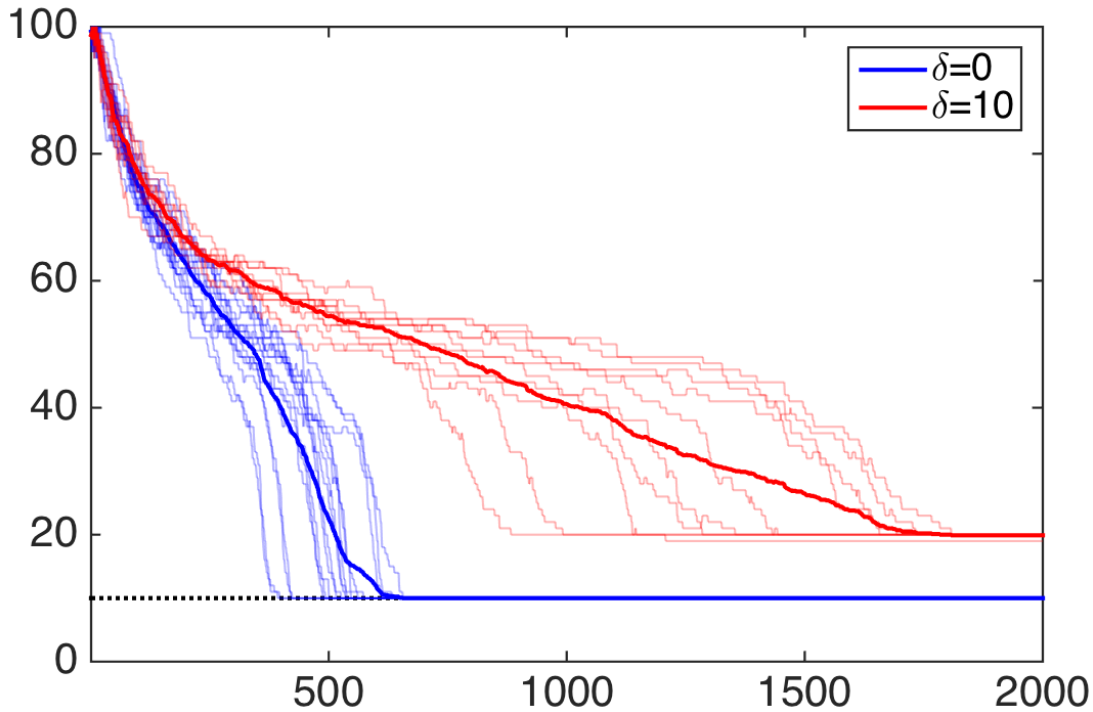
- blue curve for $\delta = 0$ (classical curve [Dossal *et al* '12] [Ameluxen *et al* '13])
- red curve for the largest δ (from which there is no ℓ_2-stability)
- intermediate transition curves illustrate the typical tradeoff

  complexity level of x_0 / instability in the presence of noise

Illustration of the identification of proximal-gradient algorithm

Plot the evolution of $\|x_k\|_0$ with $x_{k+1} = \text{prox}_{\gamma\|\cdot\|_1}\left(x_k - \gamma\Phi^*(\Phi x_k - y)\right)$

(for instances with $\|x_0\|_0 = 10$ and $\delta = 0$ or 10)



δ quantifies the degeneracy of the problem and the identification of algorithm

- $\delta = 0$: weak degeneracy $\rightarrow$ exact identification
- $\delta = 10$: strong degeneracy $\rightarrow$ enlarged identification

Conclusions, perspectives

Take-home message

- Previous localization results: exact, but restricted to non-degenerate cases
  vs. real-life problem are often degenerate, as in medical imaging
- General localization results in enlarged strata (explaining observed phenomena)
- Exploit the strong primal-dual structure of regularizers used in machine learning and image processing applications

Extensions

- Many possible refinements of sensitivity results
  other data fidelity terms, a priori control on strata dimension, explaining transition curves...
- Identification for other algorithms
  relaxed and inertial versions of proximal-gradient, other splitting methods
- Identification to be exploited by accelerate algorithms
  $\rightarrow$ see the talk of Jalal, this afternoon !

Conclusions, perspectives

Take-home message

- Previous localization results: exact, but restricted to non-degenerate cases
  vs. real-life problem are often degenerate, as in medical imaging
- General localization results in enlarged strata (explaining observed phenomena)
- Exploit the strong primal-dual structure of regularizers used in machine learning and image processing applications

Extensions

- Many possible refinements of sensitivity results
  other data fidelity terms, a priori control on strata dimension, explaining transition curves...
- Identification for other algorithms
  relaxed and inertial versions of proximal-gradient, other splitting methods
- Identification to be exploited by accelerate algorithms
  $\rightarrow$ see the talk of Jalal, this afternoon !

thanks !!