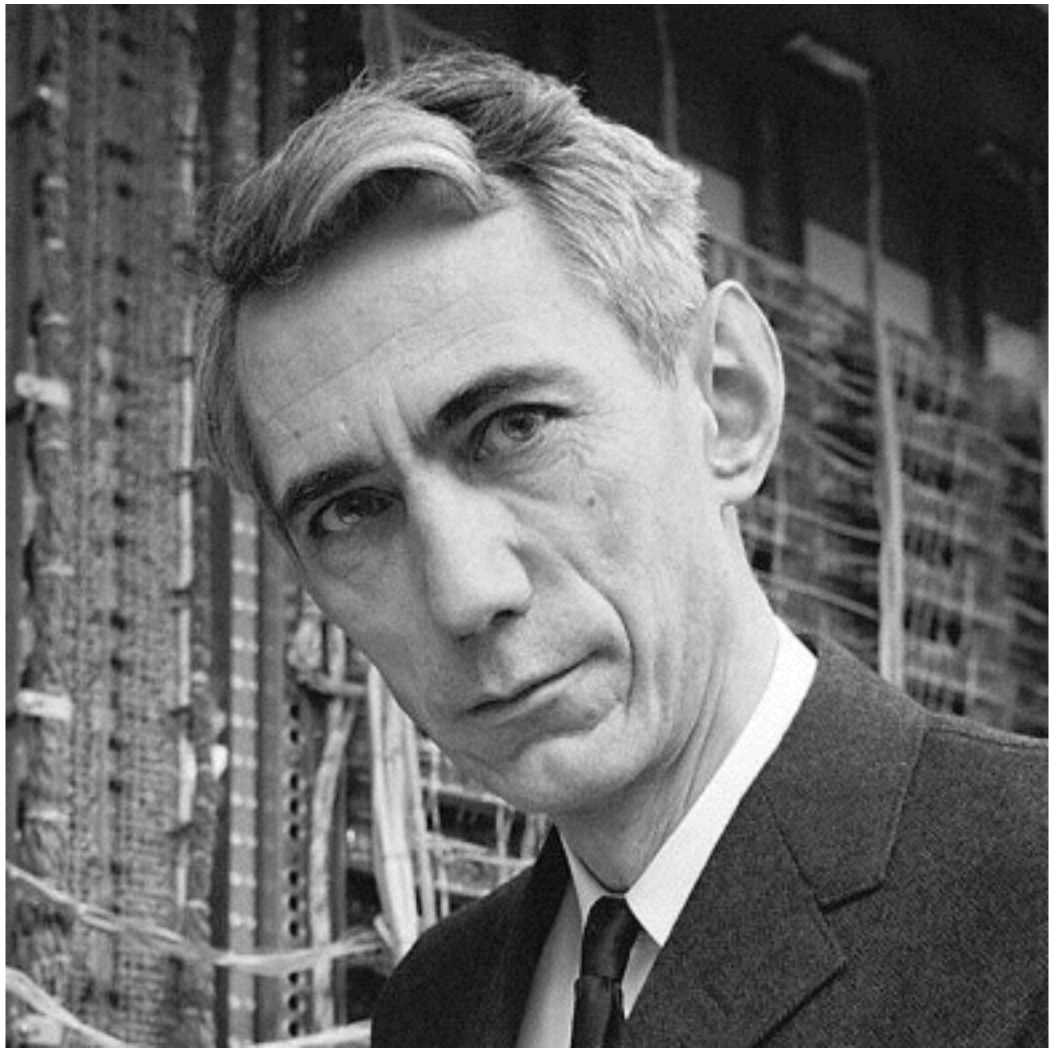
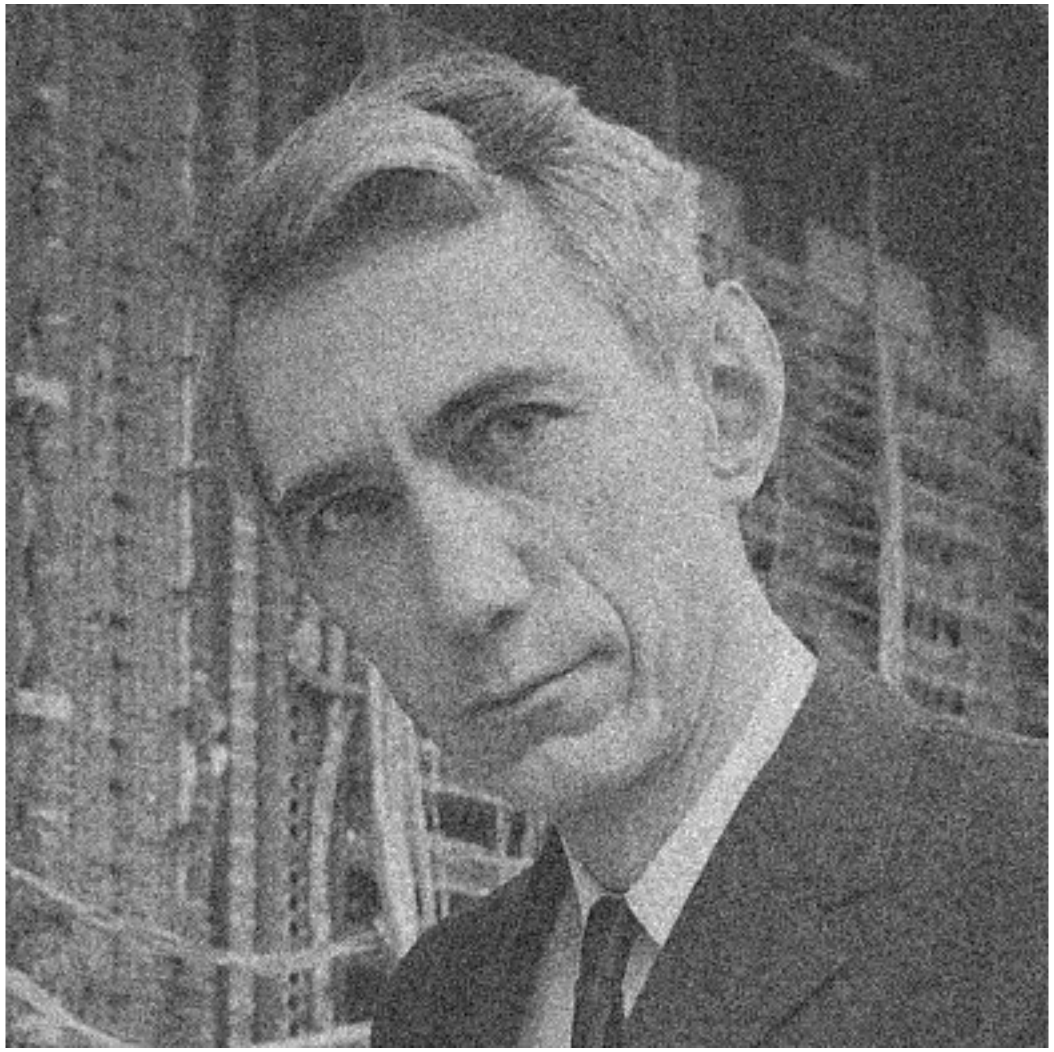
Degrees of Freedom for Partly Smooth Regularizers

Samuel Vaiter
CNRS & IMB, Dijon, France

2016/07/04
AIMS'16





many denoising methods are parametric

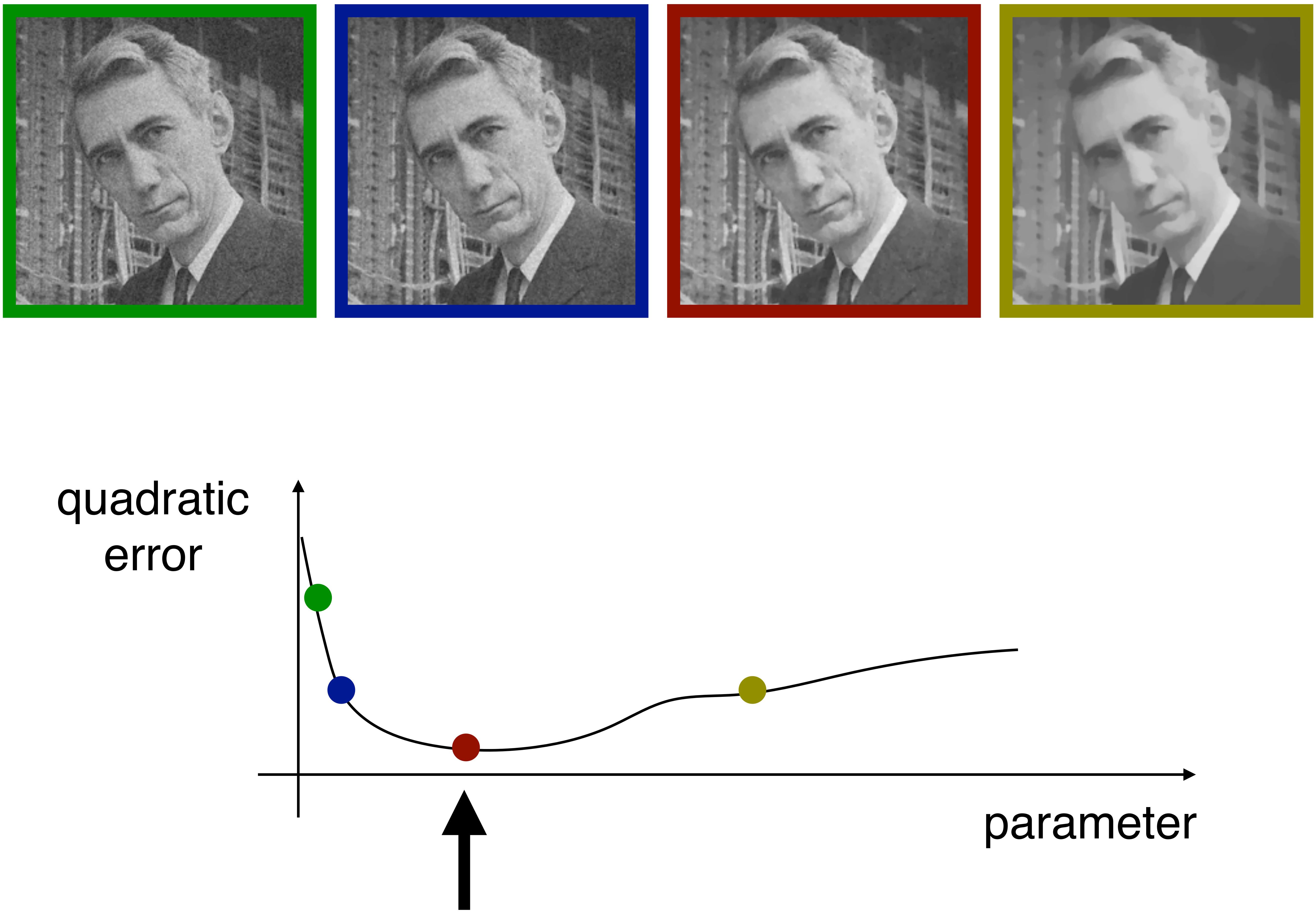
many denoising methods are parametric

parameter selection
“by hand"

many denoising methods are parametric

parameter selection
“by hand"

automatic
parameter selection


quadratic
error
parameter

Problem Statement

Inverse Problem and Variational Methods

$$Y = \mathbf{\Phi} x_0 + W$$

observations

$\mathbb{R}^n$

AWGN

$\sim \mathcal{N}(0, \sigma^2 \mathbf{Id})$

degradation operator

$\mathbb{R}^p \to \mathbb{R}^n$

ground truth

$\mathbb{R}^p$

Inverse Problem and Variational Methods

$$Y = \mathbf{\Phi} x_0 + W$$

observations $\mathbb{R}^n$

degradation operator $\mathbb{R}^p \to \mathbb{R}^n$

ground truth $\mathbb{R}^p$

AWGN $\sim \mathcal{N}(0, \sigma^2 \mathbf{Id})$

Variational methods

compromise

$$\hat{x}_\lambda(y) \in \operatorname*{Argmin}_{x \in \mathbb{R}^p} F(\mathbf{\Phi} x, y) + \lambda J(x)$$

data fidelity

convex "regularization"

Lasso

Total Variation

Nuclear

...

Our Goal

$$\hat{x}_\lambda(y) \in \underset{x \in \mathbb{R}^p}{\text{Argmin}}\ F(\mathbf{\Phi} x, y) + \lambda J(x)$$

$$\min_{\lambda \in \mathbb{R}_+} \mathrm{R}_\lambda(Y) \overset{\text{def.}}{=} \mathbb{E}_W[\|\mathbf{\Phi}\hat{x}_\lambda(Y) - \mathbf{\Phi}x_0\|_2^2]$$

Our Goal

$$\hat{x}_\lambda(y) \in \underset{x \in \mathbb{R}^p}{\operatorname{Argmin}} \; F(\mathbf{\Phi} x, y) + \lambda J(x)$$

$$\min_{\lambda \in \mathbb{R}_+} \; \mathrm{R}_\lambda(Y) \overset{\text{def.}}{=} \mathbb{E}_W[\|\mathbf{\Phi}\hat{x}_\lambda(Y) - \mathbf{\Phi} x_0\|_2^2]$$

2 issues

- x_0 is unknown
- we only have access to one realization of Y

Our Goal

$$\hat{x}_\lambda(y) \in \underset{x \in \mathbb{R}^p}{\operatorname{Argmin}}\ F(\boldsymbol{\Phi} x, y) + \lambda J(x)$$

$$\min_{\lambda \in \mathbb{R}_+} \mathrm{R}_\lambda(Y) \overset{\text{def.}}{=} \mathbb{E}_W[\|\boldsymbol{\Phi} \hat{x}_\lambda(Y) - \boldsymbol{\Phi} x_0\|_2^2]$$

2 issues

- x_0 is unknown
- we only have access to one realization of Y

create an estimator of $\mathrm{R}_\lambda(Y)$

Degrees of Freedom and Stein's Lemma

degrees of freedom
(Efron 1986)

$$df = \sum_{i=1}^{n} \frac{1}{\sigma^2} \operatorname{cov}(Y_i, \hat{\mu}_i(Y))$$

Degrees of Freedom and Stein's Lemma

degrees of freedom
(Efron 1986)

$$df = \sum_{i=1}^{n} \frac{1}{\sigma^2} \operatorname{cov}(Y_i, \hat{\mu}_i(Y))$$

empirical
degrees of freedom

$$\hat{df} = \operatorname{div}(\hat{\mu})(Y) = \operatorname{tr}(D\hat{\mu}(Y))$$

Degrees of Freedom and Stein's Lemma

degrees of freedom
(Efron 1986)

$$df = \sum_{i=1}^{n} \frac{1}{\sigma^2} \operatorname{cov}(Y_i, \hat{\mu}_i(Y))$$

empirical
degrees of freedom

$$\hat{df} = \operatorname{div}(\hat{\mu})(Y) = \operatorname{tr}(D\hat{\mu}(Y))$$

Stein's lemma (1981)

$\hat{\mu}$ weakly differentiable with essentially bounded weak derivative

$$\Downarrow$$

$$\mathbb{E}[\hat{df}] = df$$

Stein Unbiased Risk Estimation (SURE)

degrees of freedom
(Efron 1986)

$$df = \sum_{i=1}^{n} \frac{1}{\sigma^2} \mathrm{cov}(Y_i, \hat{\mu}_i(Y))$$

empirical
degrees of freedom

$$\hat{df} = \mathrm{div}(\hat{\mu})(Y) = \mathrm{tr}(D\hat{\mu}(Y))$$

$$\mathrm{SURE}(\hat{\mu})(Y) = \|Y - \hat{\mu}(Y)\|_2^2 + 2\sigma^2 \hat{df} - n\sigma^2$$

$\hat{\mu}$ weakly differentiable with essentially bounded weak derivative

$$\Downarrow$$

$$\mathbb{E}[\mathrm{SURE}(\hat{\mu})(Y)] = \mathbb{E}[\|\hat{\mu}(Y) - \boldsymbol{\Phi} x_0\|_2^2]$$

Three Missions

$$\hat{x}_\lambda(y) \in \underset{x \in \mathbb{R}^p}{\text{Argmin}}\ F(\mathbf{\Phi} x, y) + \lambda J(x)$$

Prove that $y \mapsto \hat{\mu}(y) = \mathbf{\Phi} \hat{x}_\lambda(y)$ is

single-valued

weakly differentiable

such that we know how to compute $\text{div}(\mu)(y)$

Sensitivity Analysis

An Observation

$$\hat{x}_\lambda(y) \in \underset{x \in \mathbb{R}^p}{\operatorname{Argmin}} \ \frac{1}{2}\|\mathbf{\Phi} x - y\|_2^2 + \lambda J(x)$$

$\hat{\mu}(y) = \mathbf{\Phi}\hat{x}_\lambda(y)$ uniquely defined (true when $\nabla^2 F$ positive definite)

$y \mapsto \hat{\mu}(y)$ Lipschitz, hence weakly differentiable

An Observation

$$\hat{x}_\lambda(y) \in \underset{x \in \mathbb{R}^p}{\operatorname{Argmin}} \ \frac{1}{2}\|\mathbf{\Phi} x - y\|_2^2 + \lambda J(x)$$

$\hat{\mu}(y) = \mathbf{\Phi}\hat{x}_\lambda(y)$ uniquely defined (true when $\nabla^2 F$ positive definite)

$y \mapsto \hat{\mu}(y)$ Lipschitz, hence weakly differentiable

Are we done ?

An Observation

$$\hat{x}_\lambda(y) \in \operatorname*{Argmin}_{x \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{\Phi} x - y\|_2^2 + \lambda J(x)$$

$\hat{\mu}(y) = \mathbf{\Phi}\hat{x}_\lambda(y)$ uniquely defined (true when $\nabla^2 F$ positive definite)

$y \mapsto \hat{\mu}(y)$ Lipschitz, hence weakly differentiable

Are we done ?

No, we need a formula for $\operatorname{div}(\hat{\mu})(y)$ true a.e. to compute $\mathbb{E}[\hat{df}]$

$\longrightarrow$ tricky part

Simple Example

$$\hat{x}_\lambda(y) = \underset{x \in \mathbb{R}^p}{\operatorname{argmin}}\ F(\boldsymbol{\Phi} x, y) + \lambda J(x)$$

Let $F(z, y) = \|z - y\|_2^2$ and J is C^2

Simple Example

$$\hat{x}_\lambda(y) = \underset{x \in \mathbb{R}^p}{\operatorname{argmin}}\ F(\boldsymbol{\Phi} x, y) + \lambda J(x)$$

Let $F(z, y) = \|z - y\|_2^2$ and J is C^2

First-order conditions

$$\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\hat{x}_\lambda(y) - y) + \lambda \nabla J(\hat{x}_\lambda(y)) = 0$$

Simple Example

$$\hat{x}_\lambda(y) = \underset{x \in \mathbb{R}^p}{\operatorname{argmin}}\ F(\boldsymbol{\Phi} x, y) + \lambda J(x)$$

Let $F(z, y) = \|z - y\|_2^2$ and J is C^2

First-order conditions

$$\boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\hat{x}_\lambda(y) - y) + \lambda \nabla J(\hat{x}_\lambda(y)) = 0$$

Implicit function theorem

$$D\hat{\mu}(y) = \boldsymbol{\Phi}\Gamma(y)^{-1}\boldsymbol{\Phi}^\top \quad \text{where} \quad \Gamma = \boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \lambda D^2 J$$

Simple Example

$$\hat{x}_\lambda(y) = \underset{x \in \mathbb{R}^p}{\operatorname{argmin}} \; F(\mathbf{\Phi} x, y) + \lambda J(x)$$

Let $F(z, y) = \|z - y\|_2^2$ and J is C^2

First-order conditions

$$\mathbf{\Phi}^\top (\mathbf{\Phi} \hat{x}_\lambda(y) - y) + \lambda \nabla J(\hat{x}_\lambda(y)) = 0$$

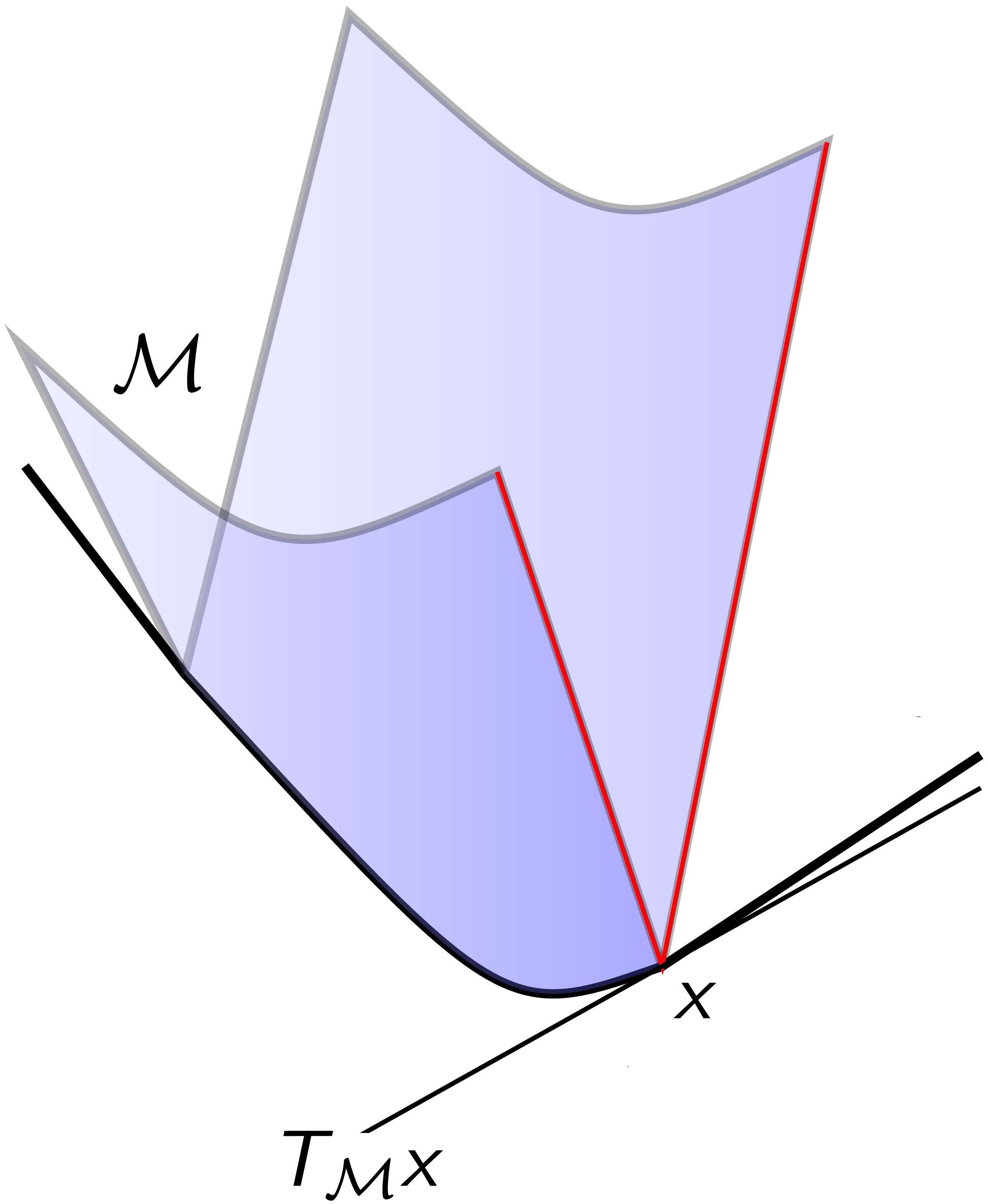
Implicit function theorem

$$D\hat{\mu}(y) = \mathbf{\Phi} \Gamma(y)^{-1} \mathbf{\Phi}^\top \quad \text{where} \quad \Gamma = \mathbf{\Phi}^\top \mathbf{\Phi} + \lambda D^2 J$$

Issues
- non-uniqueness of $\hat{x}_\lambda(y)$
- non-differentiability of J
- non-invertibility of Γ

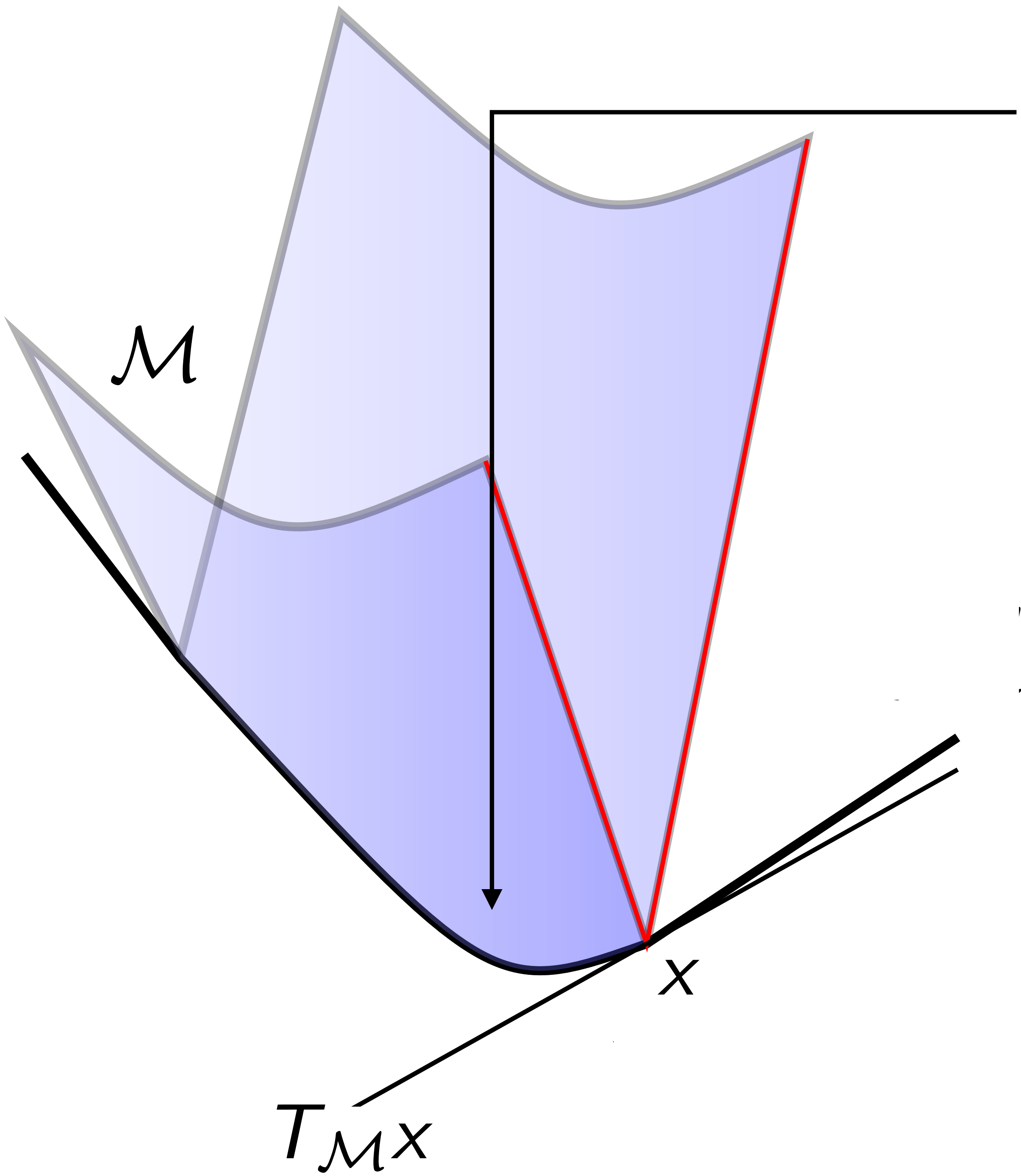
Assumption on the Regularizer

Partly smooth function [Lewis 2002]



Assumption on the Regularizer

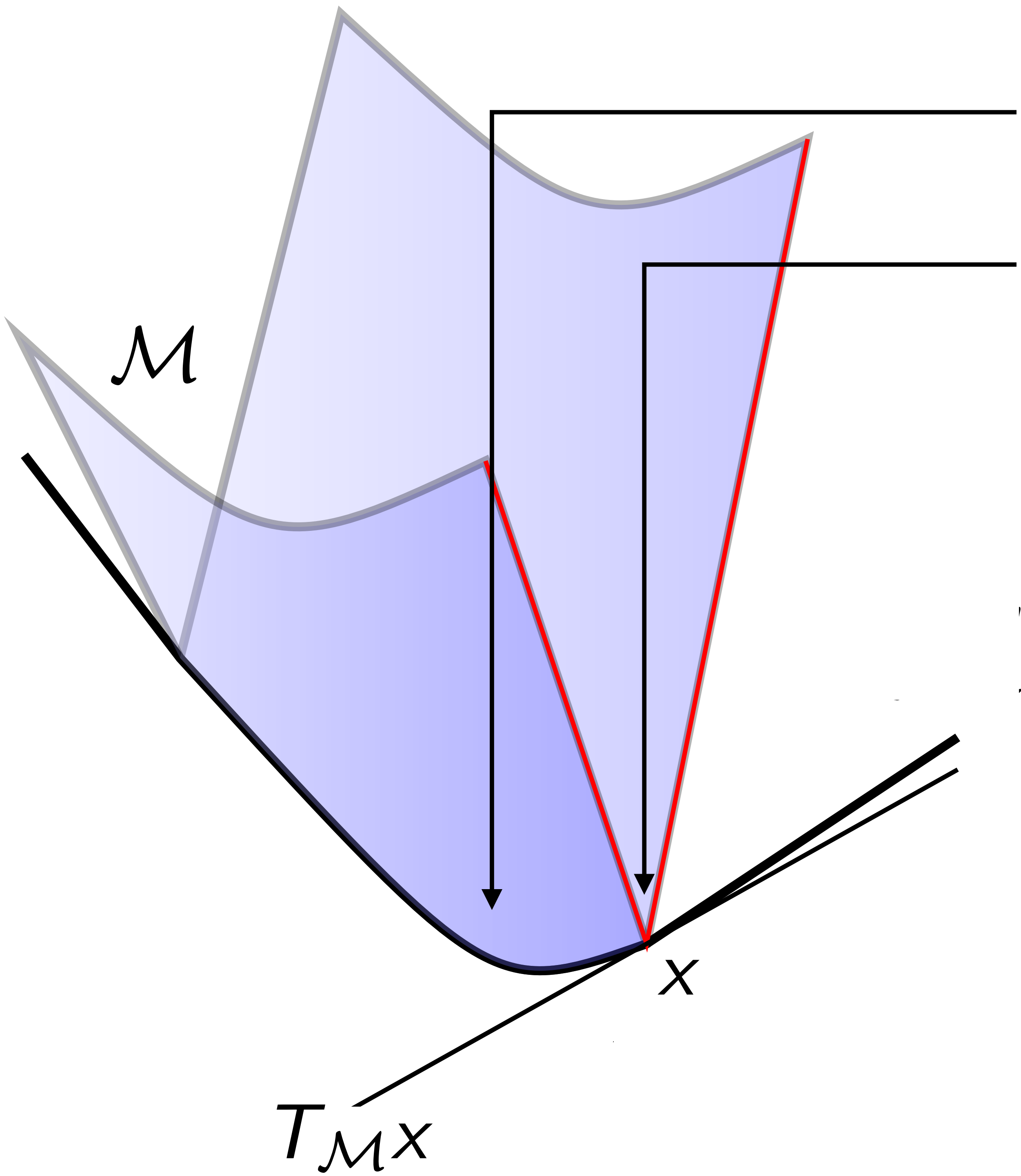
Partly smooth function [Lewis 2002]



J restricted to $\mathcal{M}$ is C^2

Assumption on the Regularizer

Partly smooth function [Lewis 2002]

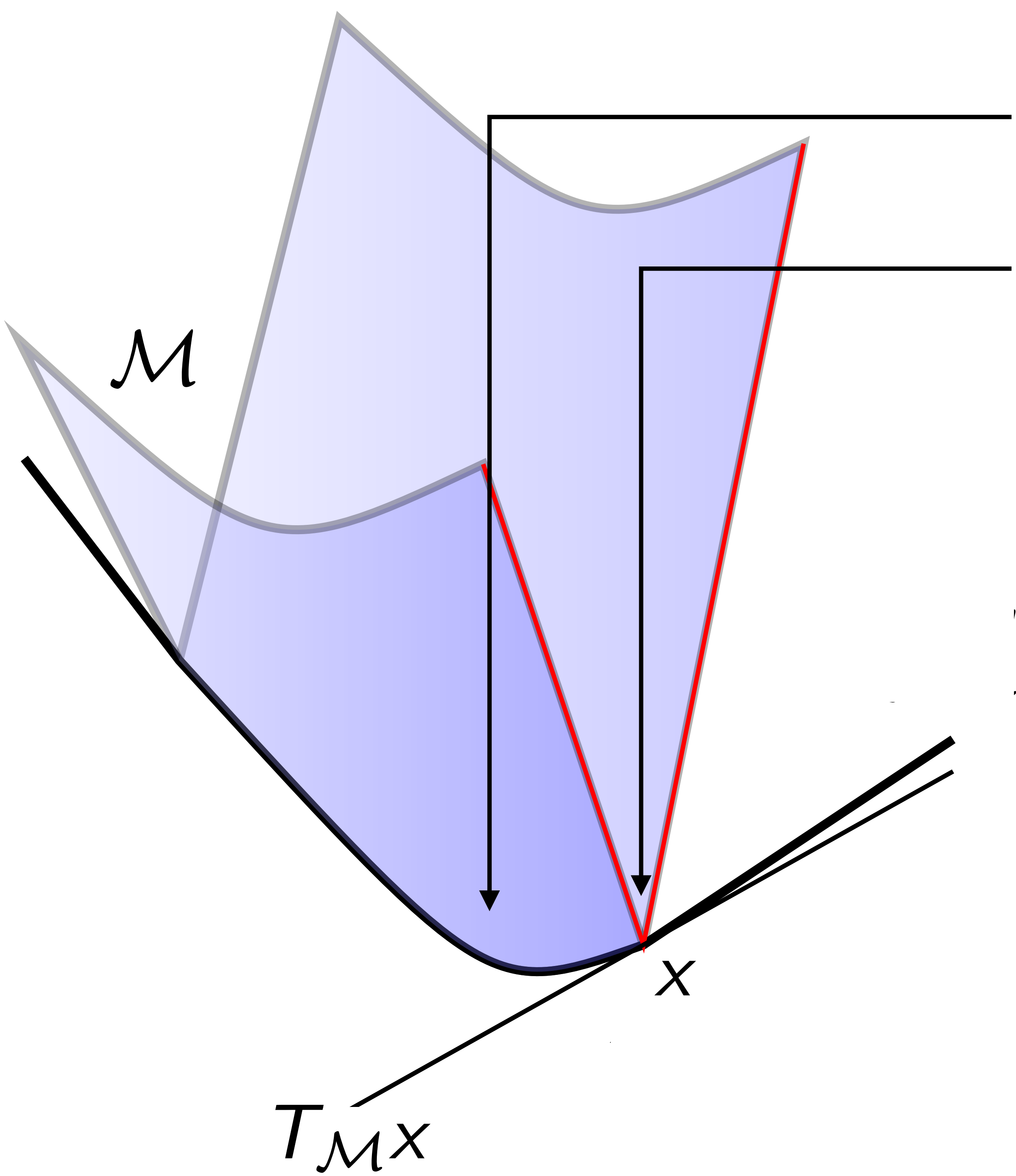


J restricted to $\mathcal{M}$ is C^2

$\forall h \in (T_{\mathcal{M}}x)^{\perp}, t \mapsto J(x+th)$
not smooth at 0

Assumption on the Regularizer

Partly smooth function [Lewis 2002]



J restricted to $\mathcal{M}$ is C^2

$\forall h \in (T_{\mathcal{M}} x)^\perp,\ t \mapsto J(x + th)$ not smooth at 0

Examples

$\|\cdot\|_1$	same support
$\|\nabla \cdot\|_{1,2}$	same jump
$\|\cdot\|_\infty$	same saturation

Sensitivity Analysis of the Prediction

$y \mapsto \hat{\mu}(y)$ is $\mathrm{C}^1(\mathbb{R}^n \setminus \mathcal{H})$ and $\forall y \notin \mathcal{H}$, $\hat{df} = \mathrm{tr}(D\hat{\mu}(y))$ where

$$D\hat{\mu}(y) = \mathbf{\Phi}_T(\mathbf{\Phi}_T^\top \mathbf{\Phi}_T + \lambda \nabla^2_{\mathcal{M}} J(\hat{x}_\lambda(y)))^+ \mathbf{\Phi}_T^\top$$

where $T = \mathcal{T}_{\mathcal{M}} \hat{x}_\lambda(y)$ and $\hat{x}_\lambda(y)$ a solution such that

$$\mathrm{Ker}\left[\nabla^2_{\mathcal{M}} J(\hat{x}_\lambda(y))\right] \cap T = \{0\}$$

Sensitivity Analysis of the Prediction

$y \mapsto \hat{\mu}(y)$ is $C^1(\mathbb{R}^n \setminus \mathcal{H})$ and $\forall y \notin \mathcal{H}$, $\hat{df} = \mathrm{tr}(D\hat{\mu}(y))$ where

$$D\hat{\mu}(y) = \boldsymbol{\Phi}_T (\boldsymbol{\Phi}_T^\top \boldsymbol{\Phi}_T + \lambda \nabla^2_{\mathcal{M}} J(\hat{x}_\lambda(y)))^+ \boldsymbol{\Phi}_T^\top$$

where $T = \mathcal{T}_{\mathcal{M}} \hat{x}_\lambda(y)$ and $\hat{x}_\lambda(y)$ a solution such that

$$\mathrm{Ker}\left[\nabla^2_{\mathcal{M}} J(\hat{x}_\lambda(y))\right] \cap T = \{0\}$$

Example

$J(x) = \|Ax\|_1$ $\qquad$ $\hat{df} = \dim \mathrm{Ker}\, A_{I^c}$

where $\mathrm{Ker}\,[\boldsymbol{\Phi}] \cap \mathrm{Ker}\,[A_{I^c}]$ and $I = \mathrm{supp}(A\hat{x}_\lambda(y))$

[Tibshirani and Taylor '12, V. et al '13]

Sensitivity Analysis of the Prediction

$y \mapsto \hat{\mu}(y)$ is $\mathrm{C}^1(\mathbb{R}^n \setminus \mathcal{H})$ and $\forall y \notin \mathcal{H}$, $\hat{df} = \mathrm{tr}(D\hat{\mu}(y))$ where

$$D\hat{\mu}(y) = \boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^\top \boldsymbol{\Phi}_T + \lambda \nabla^2_{\mathcal{M}} J(\hat{x}_\lambda(y)))^+ \boldsymbol{\Phi}_T^\top$$

where $T = \mathcal{T}_{\mathcal{M}} \hat{x}_\lambda(y)$ and $\hat{x}_\lambda(y)$ a solution such that

$$\mathrm{Ker}\left[\nabla^2_{\mathcal{M}} J(\hat{x}_\lambda(y))\right] \cap T = \{0\}$$

Sensitivity Analysis of the Prediction

$y \mapsto \hat{\mu}(y)$ is $\mathrm{C}^1(\mathbb{R}^n \setminus \mathcal{H})$ and $\forall y \notin \mathcal{H}$, $\hat{df} = \mathrm{tr}(D\hat{\mu}(y))$ where

$$D\hat{\mu}(y) = \mathbf{\Phi}_T(\mathbf{\Phi}_T^\top \mathbf{\Phi}_T + \lambda \nabla^2_{\mathcal{M}} J(\hat{x}_\lambda(y)))^+ \mathbf{\Phi}_T^\top$$

where $T = \mathcal{T}_{\mathcal{M}} \hat{x}_\lambda(y)$ and $\hat{x}_\lambda(y)$ a solution such that

$$\mathrm{Ker}\left[\nabla^2_{\mathcal{M}} J(\hat{x}_\lambda(y))\right] \cap T = \{0\}$$

Size ?

Sensitivity Analysis of the Prediction

$y \mapsto \hat{\mu}(y)$ is $C^1(\mathbb{R}^n \setminus \mathcal{H})$ and $\forall y \notin \mathcal{H}$, $\hat{df} = \mathrm{tr}(D\hat{\mu}(y))$ where

$$D\hat{\mu}(y) = \mathbf{\Phi}_T(\mathbf{\Phi}_T^\top \mathbf{\Phi}_T + \lambda \nabla^2_{\mathcal{M}} J(\hat{x}_\lambda(y)))^+ \mathbf{\Phi}_T^\top$$

where $T = \mathcal{T}_{\mathcal{M}} \hat{x}_\lambda(y)$ and $\hat{x}_\lambda(y)$ a solution such that

$$\mathrm{Ker}\left[\nabla^2_{\mathcal{M}} J(\hat{x}_\lambda(y))\right] \cap T = \{0\}$$

Size ?

Does it exist ?

Sensitivity Analysis of the Prediction

$y \mapsto \hat{\mu}(y)$ is $C^1(\mathbb{R}^n \setminus \mathcal{H})$ and $\forall y \notin \mathcal{H}$, $\hat{df} = \mathrm{tr}(D\hat{\mu}(y))$ where

$$D\hat{\mu}(y) = \boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^\top \boldsymbol{\Phi}_T + \lambda \nabla^2_{\mathcal{M}} J(\hat{x}_\lambda(y)))^+ \boldsymbol{\Phi}_T^\top$$

where $T = \mathcal{T}_{\mathcal{M}} \hat{x}_\lambda(y)$ and $\hat{x}_\lambda(y)$ a solution such that

$$\mathrm{Ker}\left[\nabla^2_{\mathcal{M}} J(\hat{x}_\lambda(y))\right] \cap T = \{0\}$$

If J is polyhedral (e.g. $\|A \cdot\|_1$, $\|A \cdot\|_\infty$, $\ldots$) or $\|A \cdot\|_{1,2}$, then

$\mathcal{H}$ is of zero Lebesgue measure

there is a solution such that $\mathrm{Ker}\left[\nabla^2_{\mathcal{M}} J(\hat{x}_\lambda(y))\right] \cap T = \{0\}$

Ingredients of the Proof

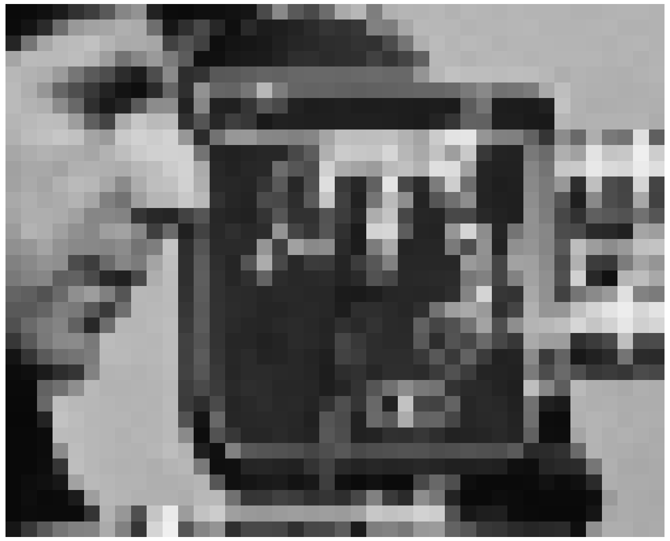
Riemmanian geometry $\longrightarrow$ provides closed-form expression

Implicit function theorem $\longrightarrow$ foundation to quantify the Jacobian

O-minimal geometry $\longrightarrow$ excludes pathological cases

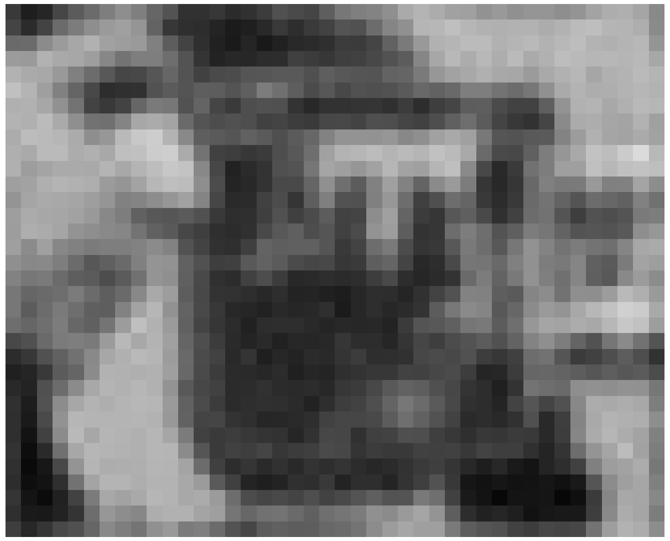
Numerical Example

A Case Study: Isotropic Total Variation



x_0

$$\xrightarrow{\mathbf{\Phi} x_0 + w}$$



y

$$\hat{x}_\lambda(y) \in \underset{x \in \mathbb{R}^{p_1 \times p_2}}{\operatorname{Argmin}} \ \frac{1}{2}\|y - \mathbf{\Phi} x\|_2^2 + \lambda \, \mathrm{TV}(x)$$

$$\mathrm{TV}(x) = \|\nabla x\|_{1,2} = \sum_{i,j} \sqrt{(x_{i+1,j} - x_{i,j})^2 + (x_{i,j+1} - x_{i,j})^2}$$

TV partly smooth at $x \in \mathbb{R}^p$ for $\mathcal{M} = \{z \ : \ \mathrm{supp}(\nabla z) = \mathrm{supp}(\nabla x)\}$

A Case Study: Isotropic Total Variation

$$\mathrm{SURE}(\hat{\mu})(Y) = \|Y - \hat{\mu}(Y)\|_2^2 + 2\sigma^2 \hat{df} - n\sigma^2$$

$$\hat{df} = \mathrm{tr}\left(\mathbf{\Phi}_I (\mathbf{\Phi}_I^\top \mathbf{\Phi}_I - \lambda\, \mathrm{div}(\delta_{\nabla \hat{x}_\lambda(y)} \circ \Pi_{(\nabla \hat{x}_\lambda(y))^\perp}) \nabla)^+ \mathbf{\Phi}_I^\top \right)$$

normalization operator

projection by block

$$I = \mathrm{supp}(\nabla \hat{x}_\lambda(y))$$

A Case Study: Isotropic Total Variation

$$\mathrm{SURE}(\hat{\mu})(Y) = \|Y - \hat{\mu}(Y)\|_2^2 + 2\sigma^2 \hat{df} - n\sigma^2$$

$$\hat{df} = \mathrm{tr}\left(\mathbf{\Phi}_I (\mathbf{\Phi}_I^\top \mathbf{\Phi}_I - \lambda\,\mathrm{div}(\delta_{\nabla \hat{x}_\lambda(y)} \circ \Pi_{(\nabla \hat{x}_\lambda(y))^\perp}) \nabla)^+ \mathbf{\Phi}_I^\top \right)$$

normalization operator

projection by block

$$I = \mathrm{supp}(\nabla \hat{x}_\lambda(y))$$

$D\hat{\mu}(y)$ potentially huge $p \times p \longrightarrow$ Monte-Carlo estimation

$$\hat{df}^{\mathrm{MC}}(z) = \langle z,\, D\mu(y) \cdot z \rangle$$

A Case Study: Isotropic Total Variation

$$\mathrm{SURE}(\hat{\mu})(Y) = \|Y - \hat{\mu}(Y)\|_2^2 + 2\sigma^2 \hat{df} - n\sigma^2$$

$$\hat{df} = \mathrm{tr}\left(\boldsymbol{\Phi}_{\mathcal{I}} (\boldsymbol{\Phi}_{\mathcal{I}}^\top \boldsymbol{\Phi}_{\mathcal{I}} - \lambda\, \mathrm{div}(\delta_{\nabla \hat{x}_\lambda(y)} \circ \Pi_{(\nabla \hat{x}_\lambda(y))^\perp}) \nabla)^+ \boldsymbol{\Phi}_{\mathcal{I}}^\top \right)$$

normalization operator

projection by block

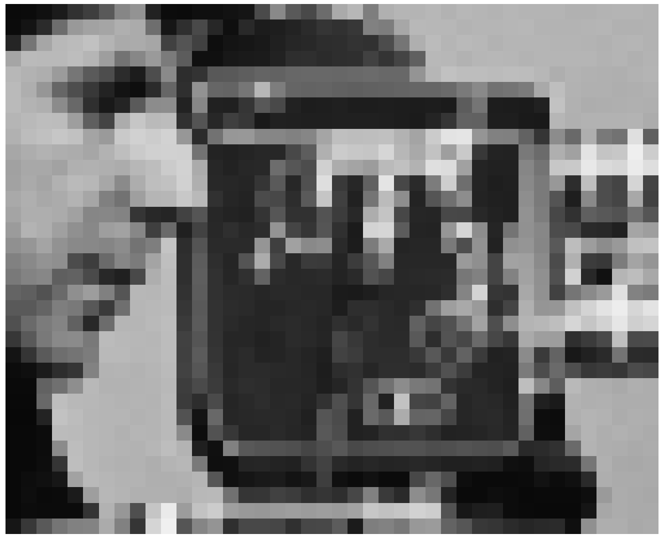
$$\mathcal{I} = \mathrm{supp}(\nabla \hat{x}_\lambda(y))$$

$D\hat{\mu}(y)$ potentially huge $p \times p \longrightarrow$ Monte-Carlo estimation

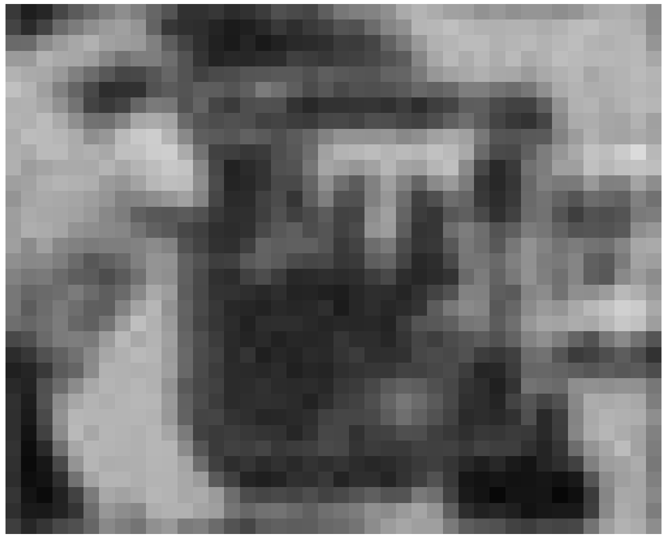
$$\hat{df}^{\mathrm{MC}}(z) = \langle z,\ D\mu(y) \cdot z \rangle$$

computable with a linear system (GMRES)

A Case Study: Isotropic Total Variation

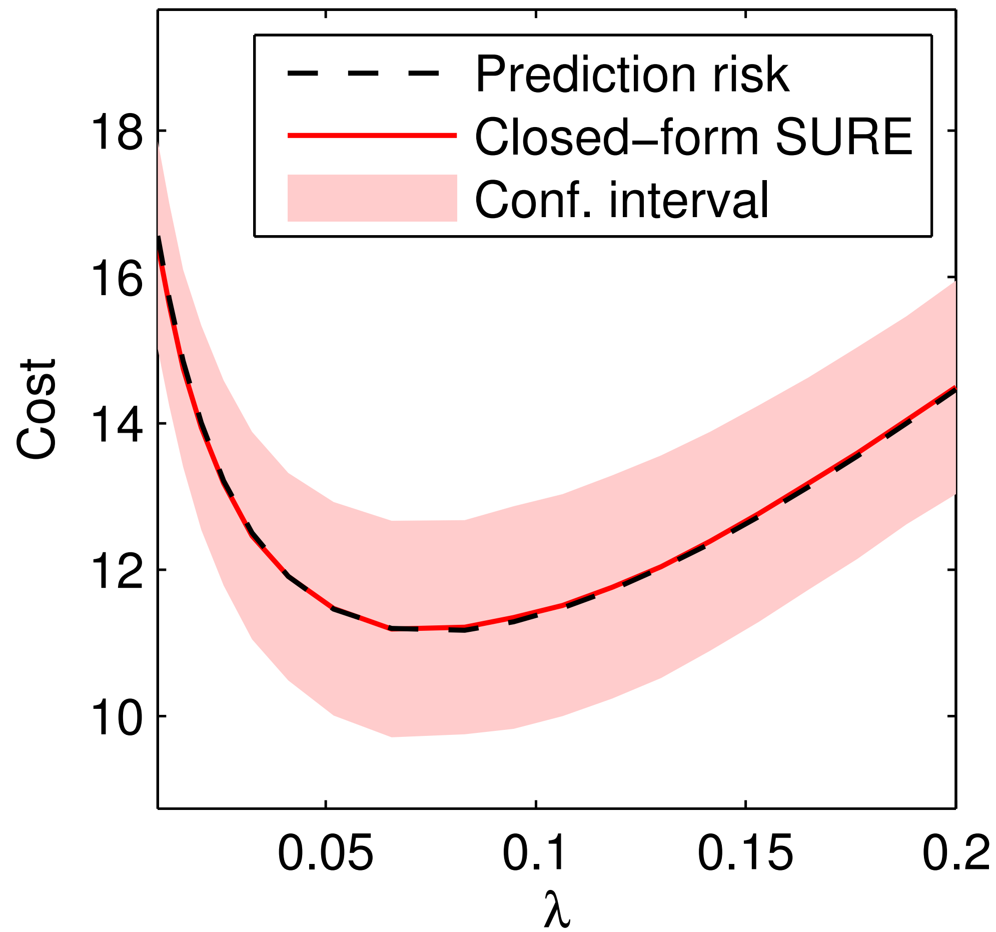


x_0



y

$\hat{x}_\lambda(y)$



Conclusion

Risk estimation $\iff$ Sensitivity of the estimator

Conclusion

Risk estimation $\Longleftrightarrow$ Sensitivity of the estimator

Practical limitations

$\longrightarrow$ Closed form expression of $\hat{df}$ unavailable for arbitrary J

$\longrightarrow$ unsuitable for non-variational methods

$\longrightarrow$ can be unstable if the model is not identified

Conclusion

Risk estimation $\Longleftrightarrow$ Sensitivity of the estimator

Practical limitations

$\longrightarrow$ Closed form expression of $\hat{df}$ unavailable for arbitrary J

$\longrightarrow$ unsuitable for non-variational methods

$\longrightarrow$ can be unstable if the model is not identified

Alternative approaches

$\longrightarrow$ Finite difference approximation SURE [Ramani et al. '08]

$\longrightarrow$ Iterative Chain Rule SUGAR [Deledalle et al. '14]

Thanks for your attention !

Any questions ?

Joint work with C. Deledalle, C. Dossal, J. Fadili and G. Peyré
The Degrees of Freedom of Partly Smooth Regularizers
Annals of the Institute of Statistical Mathematics (to appear), 2016