Smooth over-parameterized solvers for non-smooth structured optimisation

Clarice Poon
University of Bath

Joint work with Gabriel Peyré (ENS Paris)

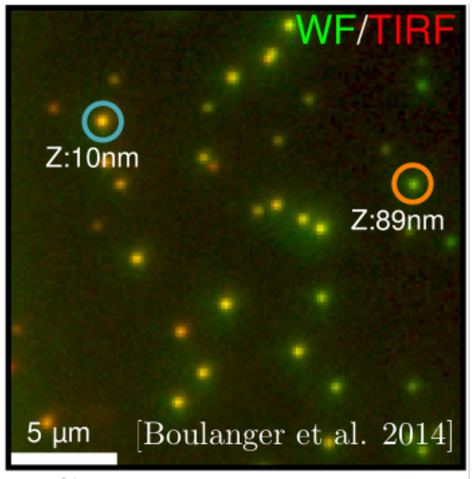
References:

1. *Smooth Bilevel Programming for Sparse Regularization, Neurips 2021.*
2. *Smooth over-parameterized solvers for non-smooth structured optimization. (preprint)*

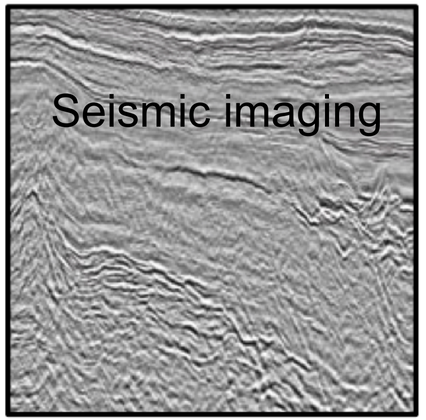
Structure promoting regularisation

Given $A \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^m$, solve $Ax = y$.

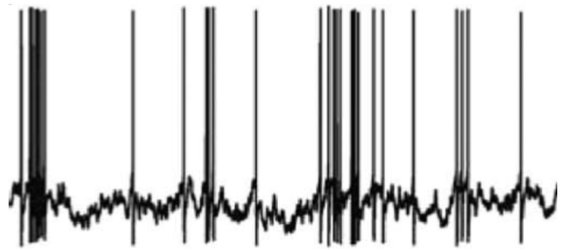
- $m \ll n$ is ill conditioned setting.
- $n \ll m$ with noise is typical ML setting



Fluorescence microscopy



Neural spikes(1D)

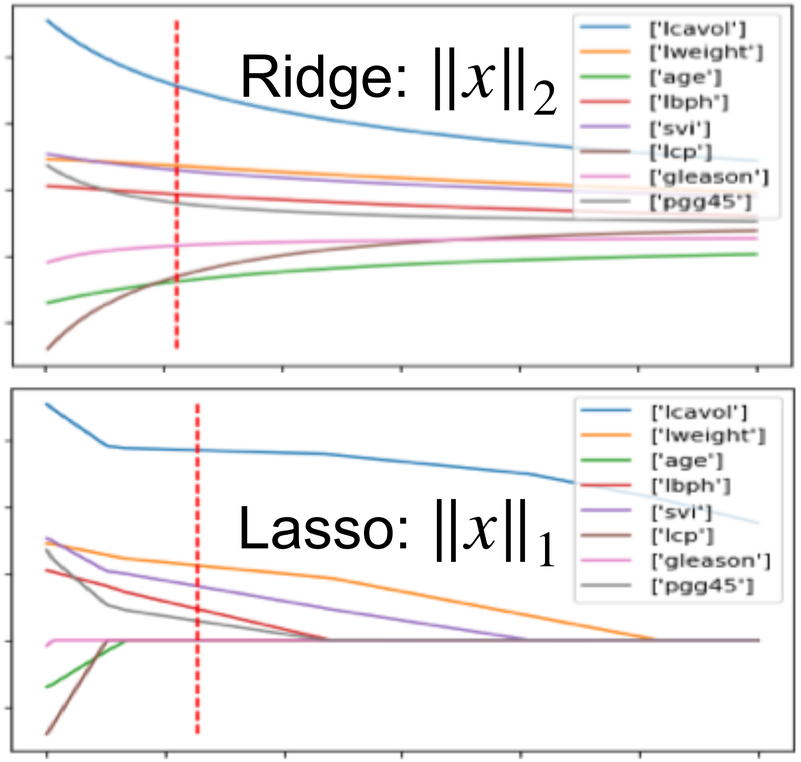


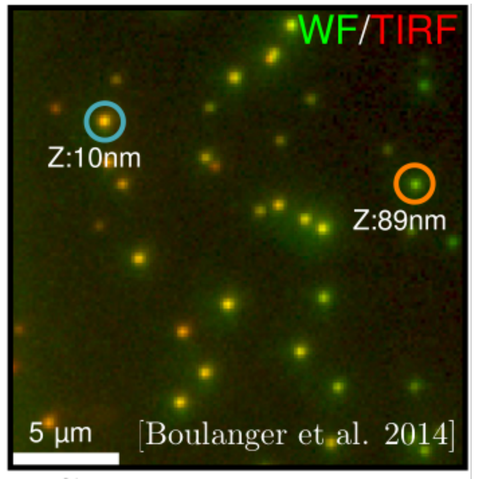
Structure promoting regularisation

Given $A \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^m$, solve $Ax = y$.

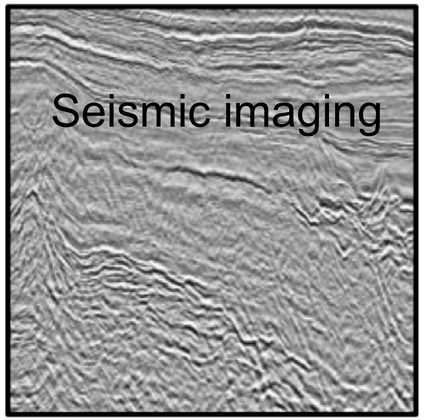
$$\min_{x \in \mathbb{R}^n} \Phi(x) := R(x) + F(Ax, y)$$

- $m \ll n$ is ill conditioned setting.
- $n \ll m$ with noise is typical ML setting

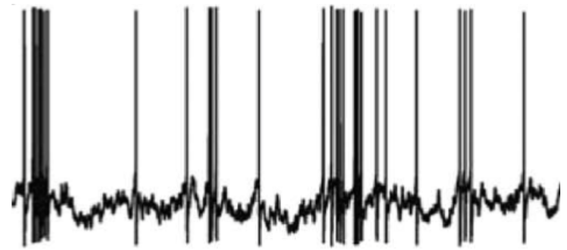




Fluorescence microscopy



Neural spikes(1D)

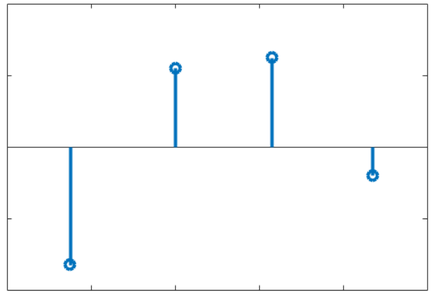


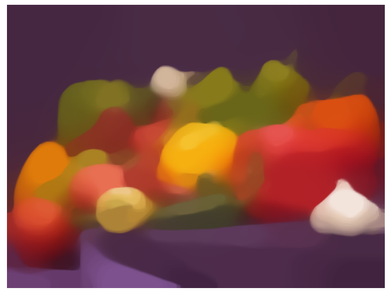
Structure promoting regularisation

$$\min_{x \in \mathbb{R}^n} R(x) + F(Ax, y)$$

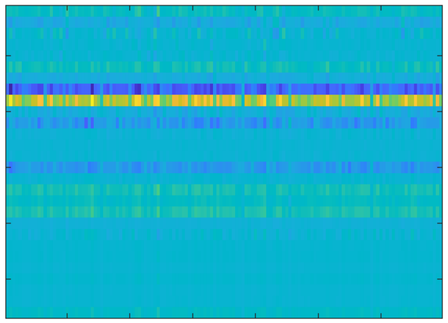
Nonsmooth R promotes structure:

- $R(x) = \|Dx\|_1$





- $R(x) = \|x\|_* = \sum_{i=1}^{n} \sigma_i(x)$

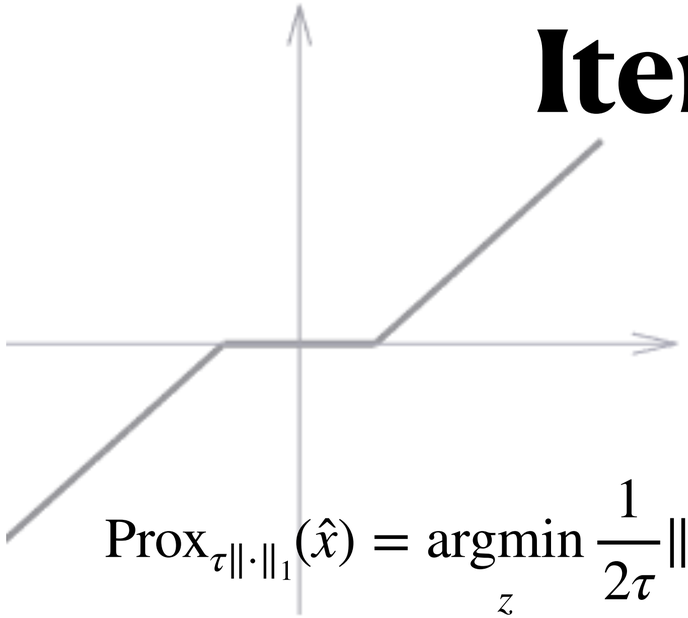


F is loss function:

- $F(x, y) = \frac{1}{2}\|x - y\|^2$
- $F(x, y) = \iota_{x=y}$
- $F(x, y) = \|x - y\|_1$

Typical optimisation approach: Proximal-based methods such as Primal-Dual, ADMM, ISTA, …

Iterative Soft Thresholding

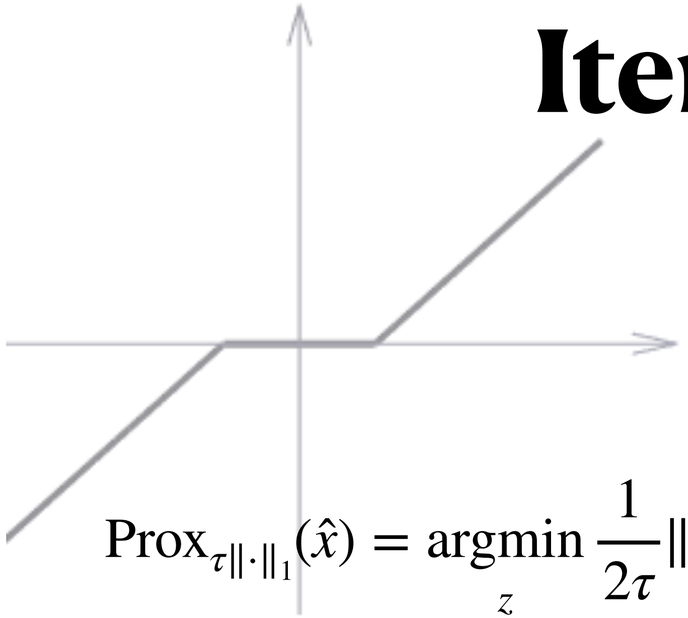


$$\min_{x \in \mathbb{R}^n} \Phi(x) = \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1$$

$$\begin{aligned} \mathrm{Prox}_{\tau\|\cdot\|_1}(\hat{x}) &= \operatorname*{argmin}_z \frac{1}{2\tau}\|z - \hat{x}\|^2 + \|z\|_1 \\ &= \mathrm{sign}(\hat{x})(|\hat{x}| - \tau)_+ \end{aligned}$$

$$\begin{cases} \hat{x}_{k+1} &= x_k - \tau A^\top (Ax_k - y) \\ x_{k+1} &= \mathrm{Prox}_{\tau\|\cdot\|_1}(\hat{x}_{k+1}) \end{cases}$$

Iterative Soft Thresholding



$$\min_{x \in \mathbb{R}^n} \Phi(x) = \frac{1}{2\lambda}\|Ax - y\|^2 + \|x\|_1$$

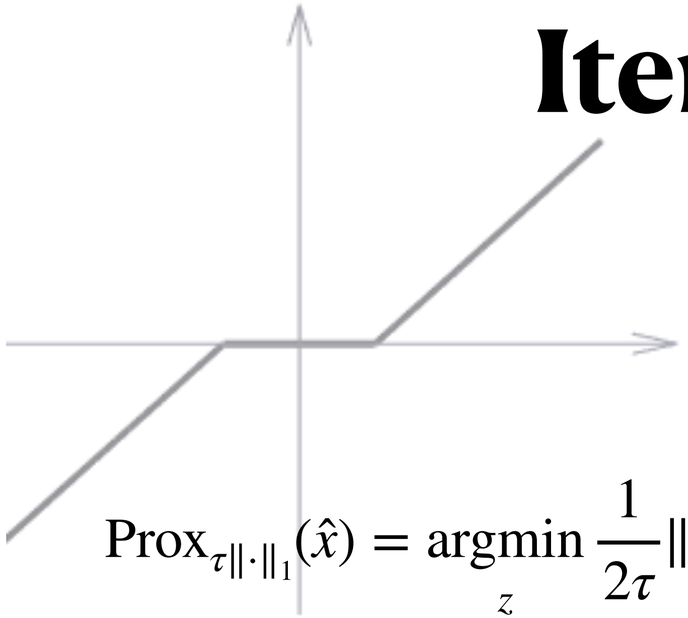
$$\begin{aligned}\mathrm{Prox}_{\tau\|\cdot\|_1}(\hat{x}) &= \operatorname*{argmin}_z \frac{1}{2\tau}\|z - \hat{x}\|^2 + \|z\|_1 \\ &= \mathrm{sign}(\hat{x})(|\hat{x}| - \tau)_+\end{aligned}$$

$$\begin{cases} \hat{x}_{k+1} &= x_k - \tau A^\top (Ax_k - y) \\ x_{k+1} &= \mathrm{Prox}_{\tau\|\cdot\|_1}(\hat{x}_{k+1}) \end{cases}$$

Convergence rates (Beck & Teboulle 2009):

$$\Phi(x_k) - \min_x \Phi(x) \leq \frac{C_n}{k}$$

Iterative Soft Thresholding



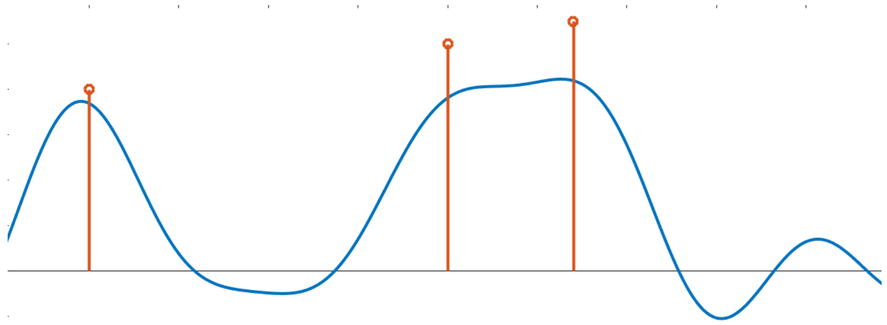
$$\min_{x \in \mathbb{R}^n} \Phi(x) = \frac{1}{2\lambda}\|Ax - y\|^2 + \|x\|_1$$

$$\begin{aligned}\mathrm{Prox}_{\tau\|\cdot\|_1}(\hat{x}) &= \operatorname*{argmin}_z \frac{1}{2\tau}\|z - \hat{x}\|^2 + \|z\|_1 \\ &= \mathrm{sign}(\hat{x})(|\hat{x}| - \tau)_+\end{aligned}$$

$$\begin{cases} \hat{x}_{k+1} &= x_k - \tau A^\top (Ax_k - y) \\ x_{k+1} &= \mathrm{Prox}_{\tau\|\cdot\|_1}(\hat{x}_{k+1}) \end{cases}$$

Convergence rates (Beck & Teboulle 2009):

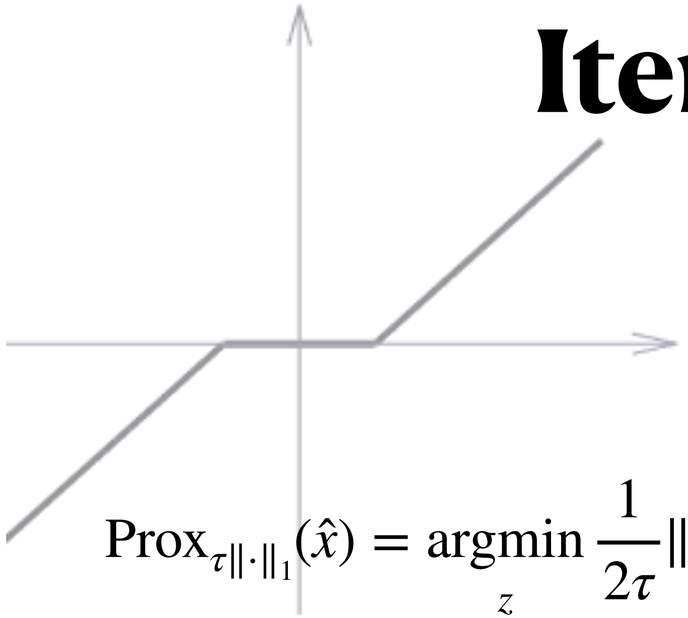
$$\Phi(x_k) - \min_x \Phi(x) \leq \frac{C_n}{k}$$



Column i : $A_i = \left(\exp(2\pi\sqrt{-1}x_i^\top \omega_k)\right)_k$

$C_n = \mathcal{O}(p^{-d})$ *if* $[x_i] \subseteq [0,1]^d$ *spaced* p *apart*

Iterative Soft Thresholding



$$\min_{x \in \mathbb{R}^n} \Phi(x) = \frac{1}{2\lambda}\|Ax - y\|^2 + \|x\|_1$$

$$\begin{aligned}\mathrm{Prox}_{\tau\|\cdot\|_1}(\hat{x}) &= \operatorname*{argmin}_z \frac{1}{2\tau}\|z - \hat{x}\|^2 + \|z\|_1 \\ &= \mathrm{sign}(\hat{x})(|\hat{x}| - \tau)_+\end{aligned}$$

$$\begin{cases} \hat{x}_{k+1} &= x_k - \tau A^\top (Ax_k - y) \\ x_{k+1} &= \mathrm{Prox}_{\tau\|\cdot\|_1}(\hat{x}_{k+1}) \end{cases}$$

Convergence rates (Beck & Teboulle 2009):

$$\Phi(x_k) - \min_x \Phi(x) \leq \frac{C_n}{k}$$

Grid-free convergence rates (Chizat 2021):

$$\Phi(x_k) - \min_x \Phi(x) \leq k^{-2/(d+1)}$$

NB: Result is independent of n

Proximal Mirror descent

Entropy function:

η strongly convex and smooth.

Bregman distance:

$$D_\eta(a,b) = \eta(a) - \eta(b) - \langle \nabla\eta(b), a-b\rangle$$

$$x_{k+1} \in \operatorname{argmin} F(x_k) + \langle \nabla F(x_k), x - x_k\rangle + \lambda\|x\|_1 + \frac{1}{\tau} D_\eta(x, x_k)$$

[Chizat 2021]:

- $\eta(x) = \frac{1}{2}\|x\|^2$ corresponds to ISTA with convergence rate $\mathcal{O}(k^{-2/(d+2)})$
- $\eta(x) = \gamma \operatorname{arsinh}(x/\gamma) - \sqrt{x^2+\gamma^2} + \gamma$ with $\gamma > 0$ with convergence rate $\mathcal{O}(k^{-1}\log(k))$

Proximal Mirror descent

Entropy function:

η strongly convex and smooth.

Mirror distance:

$$D_\eta(a,b) = \eta(a) - \eta(b) - \langle \nabla\eta(b), a-b\rangle$$

$$x_{k+1} \in \operatorname{argmin} F(x_k) + \langle \nabla F(x_k), x - x_k\rangle + \lambda\|x\|_1 + \frac{1}{\tau} D_\eta(x, x_k)$$

[Chizat 2021]:

- $\eta(x) = \frac{1}{2}\|x\|^2$ corresponds to ISTA with convergence rate $\mathcal{O}(k^{-2/(d+2)})$
- $\eta(x) = \gamma \operatorname{arsinh}(x/\gamma) - \sqrt{x^2+\gamma^2} + \gamma$ with $\gamma > 0$ with convergence rate $\mathcal{O}(k^{-1}\log(k))$

One can improve to $\mathcal{O}(1/k^2)$ rates, however, ***standard acceleration techniques such as BB step, Quasi-Newton, etc are non-trivial to apply****.*

Use of other proximal based methods (e.g. Primal Dual, ADMM) require ***tuning of several parameters.***

This Talk:

New class of algo for nonsmooth structured optimisation

(1) Overparametrize to obtain a **smooth** and **nonconvex** problem.

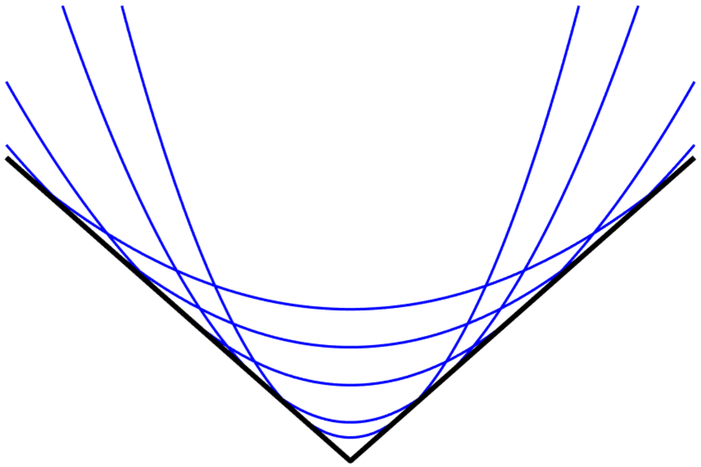
(2) **VarPro**: Place into a bilevel form for better problem conditioning.

(3) Over-parameterization lets us stay in Euclidean geometry with **dimension** and **grid independent** convergence analysis.

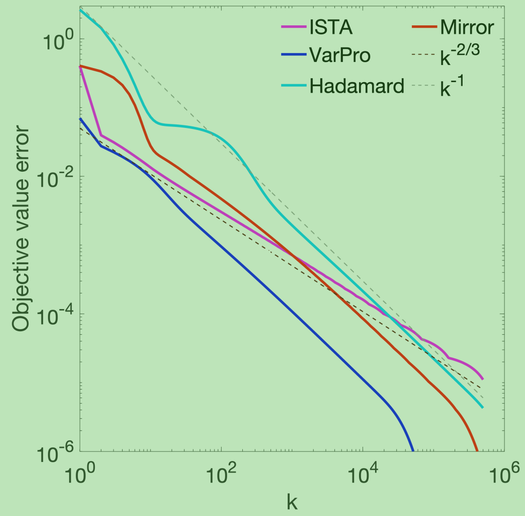
(4) Use BFGS to obtain fast convergence behaviour.

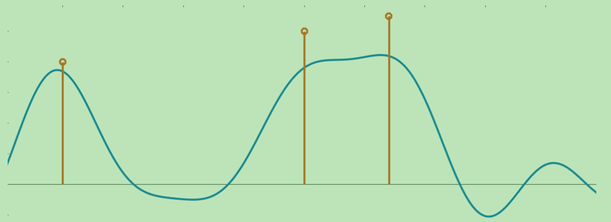
I will focus on the Lasso but our method can be applied in many settings, including total variation/group ℓ_1/ nuclear norm regularisation and nonsmooth loss functions.

Overparameterization

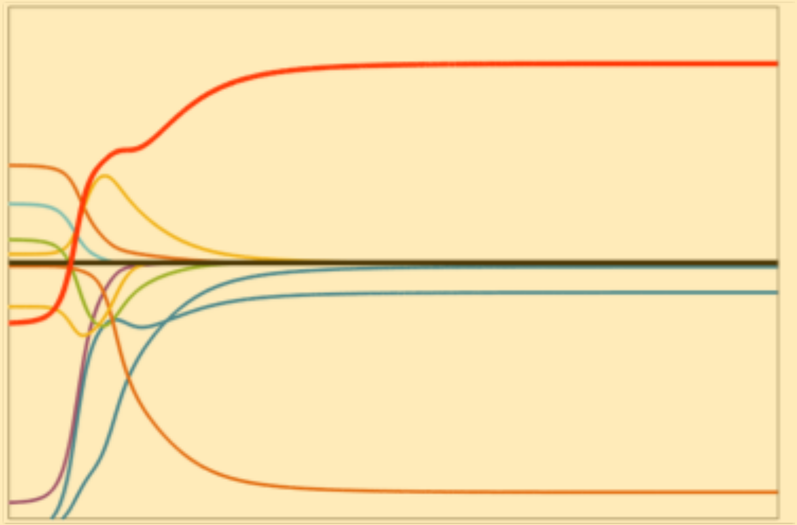


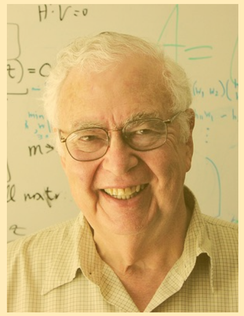
Convergence



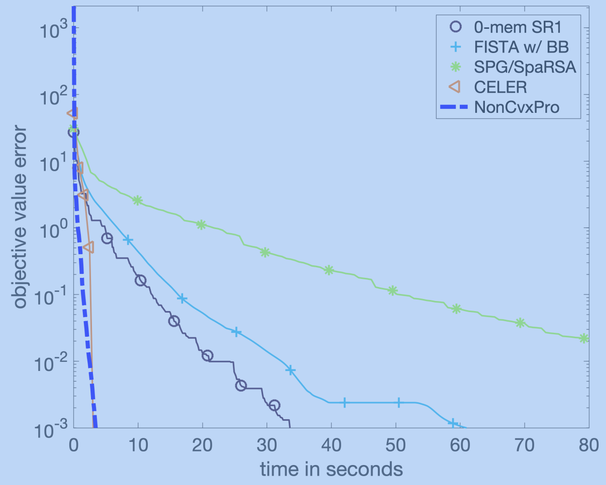


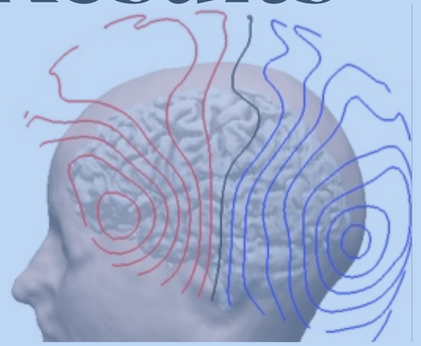
VarPro





Numerical Results





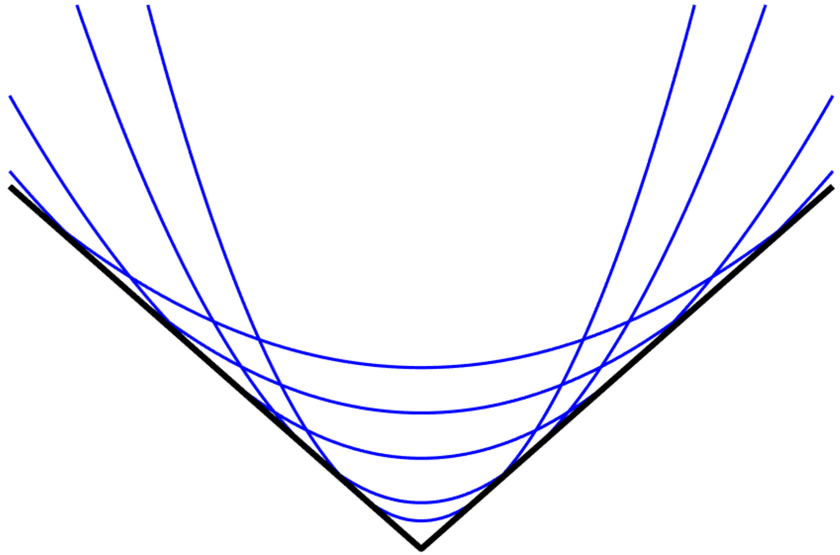
Iterative Reweighted Least Squares

The η-trick :

$$|x| = \inf_{\eta>0} \frac{1}{2}\frac{x^2}{\eta} + \frac{1}{2}\eta$$

Jointly convex in x and η

$$\min_{x\in\mathbb{R}^n} \Phi(x) = \frac{1}{2\lambda}\|Ax - y\|^2 + \|x\|_1$$



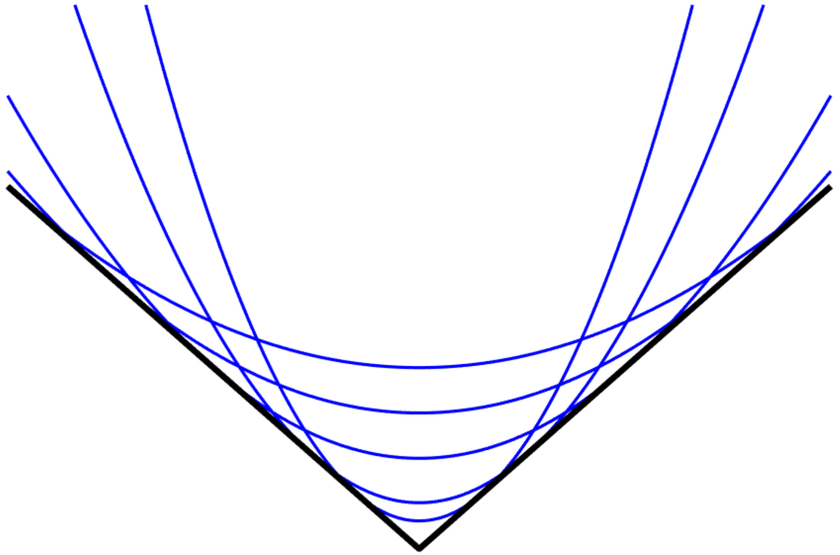
Iterative Reweighted Least Squares

The η-trick :

$$|x| = \inf_{\eta>0} \frac{1}{2}\frac{x^2}{\eta} + \frac{1}{2}\eta$$

Jointly convex in x and η

$$\min_{x\in\mathbb{R}^n} \Phi(x) = \frac{1}{2\lambda}\|Ax - y\|^2 + \|x\|_1$$



$$\min_{x\in\mathbb{R}^n \eta\in\mathbb{R}^n_+} \frac{1}{2}\sum_i \left(\frac{x_i^2}{\eta_i} + \eta_i\right) + \frac{1}{2\lambda}\|Ax - y\|^2 + \frac{\epsilon}{2}\sum_i \frac{1}{\eta_i}$$

Alternating minimisation

$$\begin{cases} x_{k+1} = (A^\top A + \lambda \mathrm{diag}(1/\eta_k))^{-1} A^\top y \\ \eta_{k+1} = \sqrt{\epsilon + x_k^2} \end{cases}$$

The Hadamard parametrization

$$|x| = \inf_{\eta>0} \frac{1}{2}\frac{x^2}{\eta} + \frac{1}{2}\eta$$

$$u = x/\sqrt{\eta}, \quad v = \sqrt{\eta}$$

$$|x| = \inf_{x=u\odot v} u^2/2 + v^2/2$$

The Hadamard parametrization

$$|x| = \inf_{\eta>0} \frac{1}{2}\frac{x^2}{\eta} + \frac{1}{2}\eta$$

$$u = x/\sqrt{\eta}, \quad v = \sqrt{\eta}$$

$$|x| = \inf_{x=u\odot v} u^2/2 + v^2/2$$

$$\min_{x\in\mathbb{R}^n} \frac{1}{2\lambda}\|Ax - y\|^2 + \|x\|_1$$

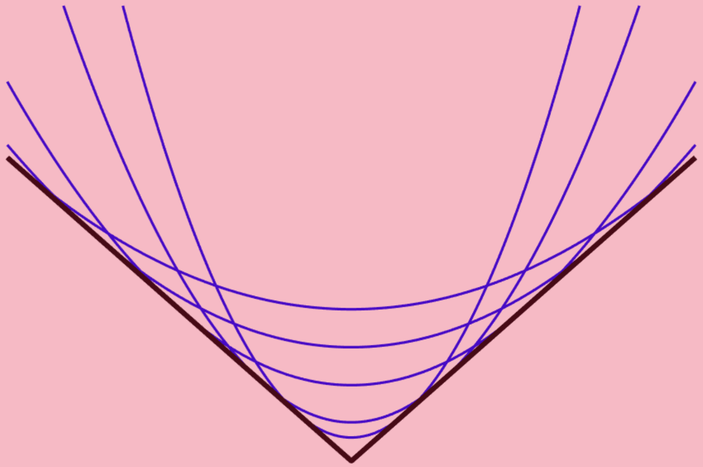
Convex and nonsmooth

$$\min_{u,v\in\mathbb{R}^n} \frac{1}{2\lambda}\|Au \odot v - y\|^2 + \frac{1}{2}\|u\|^2 + \frac{1}{2}\|v\|^2$$

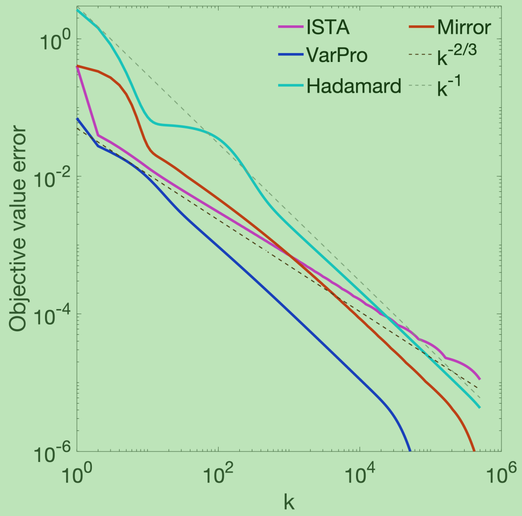
Non-Convex and smooth

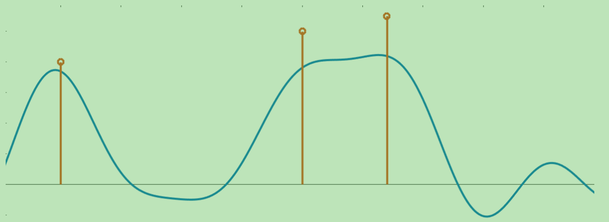
One can either do alternating minimisation or gradient descent on u, v

Overparameterization

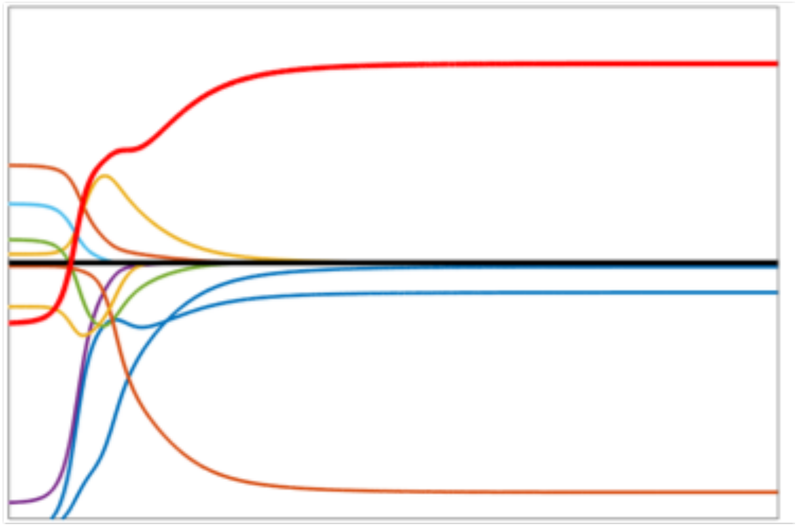


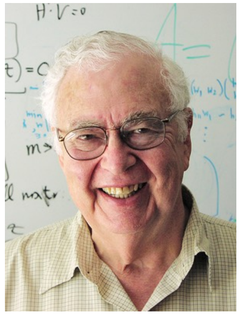
Convergence



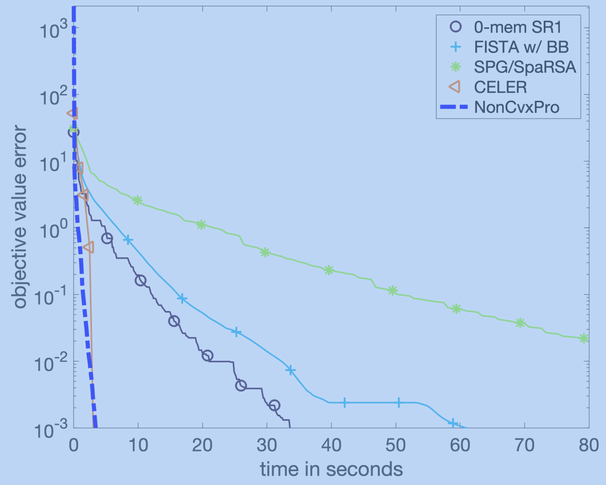


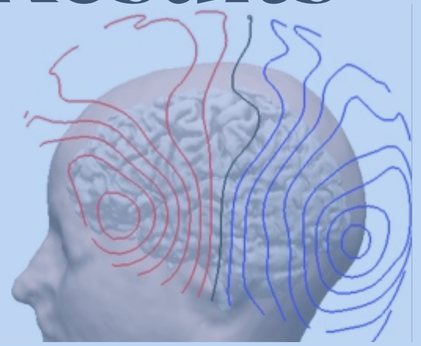
VarPro





Numerical Results



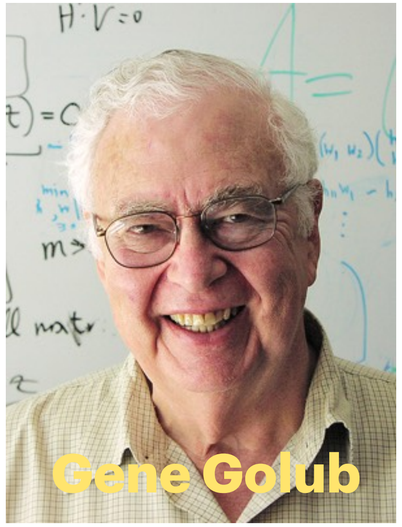


VarPro

VarPro: Instead of $\min_{u,v} F(u, v)$

Solve $\min_{v} f(v) = F(u(v), v)$ with

$$u(v) \in \operatorname{argmin}_{u} F(u, v)$$

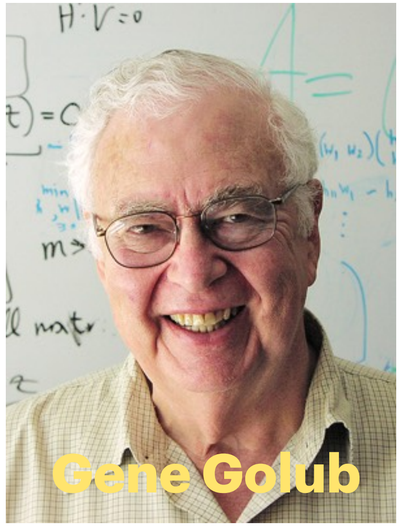


VarPro

VarPro: Instead of $\min_{u,v} F(u,v)$

Solve $\min_{v} f(v) = F(u(v), v)$ with

$$u(v) \in \operatorname*{argmin}_{u} F(u,v)$$



Easy to compute derivative:

Envelope theorem: $\nabla f(v) = \partial_v F(u(v), v)$

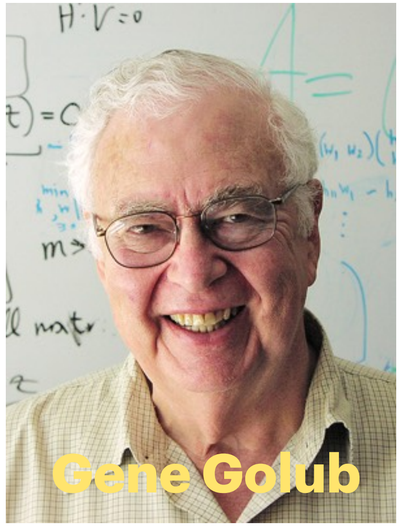
Chain rule: $\nabla f(v) = \partial_v F(u(v), v) + \partial_u F(u(v), v)\, \nabla u(v)$

VarPro

VarPro: Instead of $\min_{u,v} F(u,v)$

Solve $\min_{v} f(v) = F(u(v), v)$ with

$$u(v) \in \operatorname*{argmin}_{u} F(u,v)$$



Easy to compute derivative:

Envelope theorem: $\nabla f(v) = \partial_v F(u(v), v)$

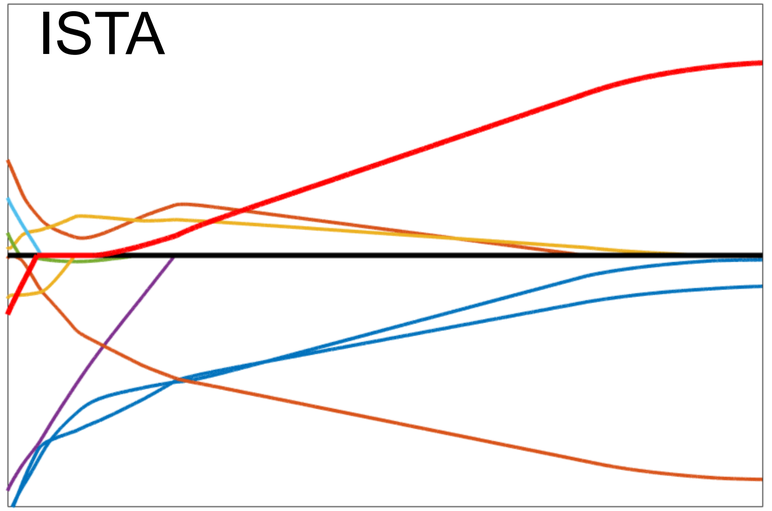
Chain rule: $\nabla f(v) = \partial_v F(u(v), v) + \partial_u F(u(v), v)\, \nabla u(v)$

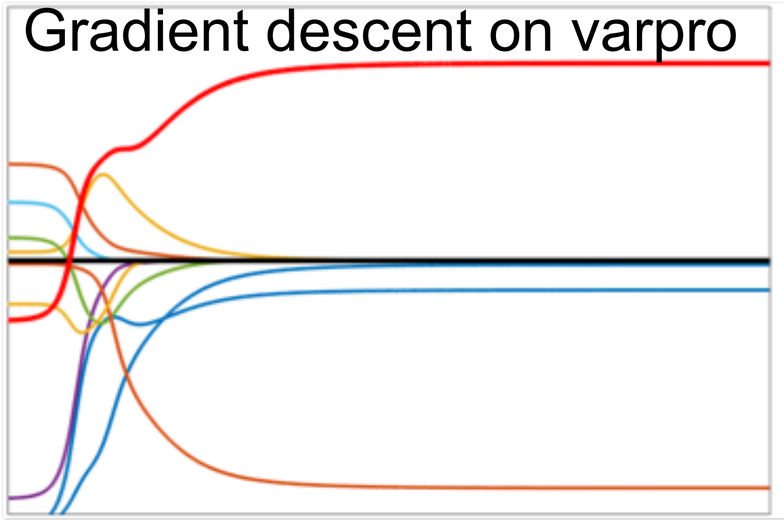
$\nabla^2 f$ is the **Schur complement** of $\nabla^2 F(u,v)$ and is typically better conditioned

Bilevel formulation

$$\min_v f(v)$$

$$f(v) := \min_u \frac{1}{2}\|u\|^2 + \frac{1}{2}\|v\|^2 + \frac{1}{2\lambda}\|Au \odot v - y\|^2$$





Bilevel formulation

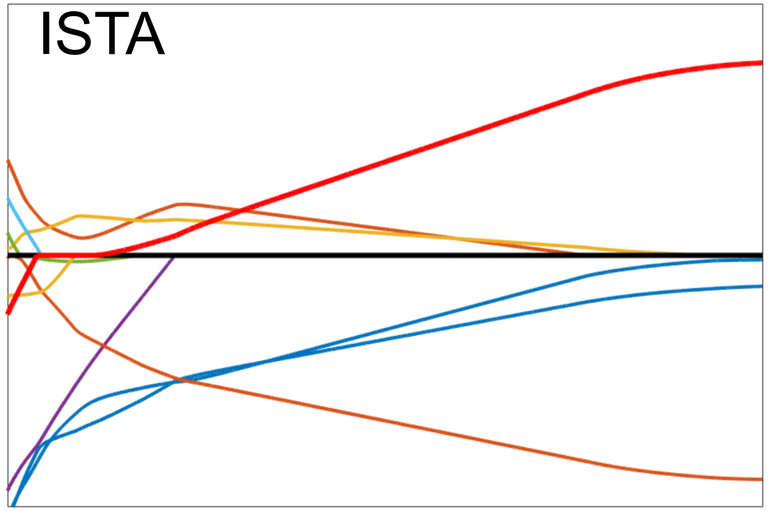
$$\min_v f(v)$$

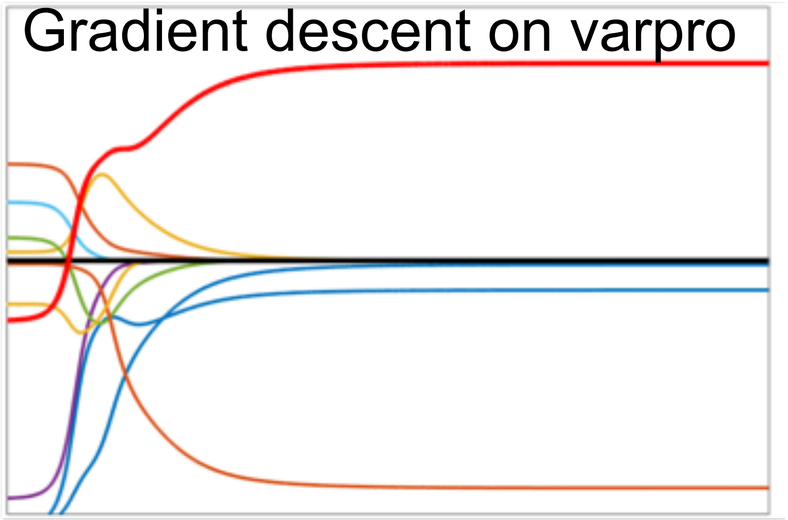
$$f(v) := \min_u \frac{1}{2}\|u\|^2 + \frac{1}{2}\|v\|^2 + \frac{1}{2\lambda}\|Au \odot v - y\|^2$$

Computing the gradient:

$$\nabla f(v) = v + \frac{1}{\lambda} u \odot A^\top (Au \odot v - y)$$

$$u = (\lambda I + \mathrm{diag}(v) A^\top A \mathrm{diag}(v))^{-1} (v \odot A^\top y)$$





Bilevel formulation

$$\min_v f(v)$$

$$f(v) := \min_u \frac{1}{2}\|u\|^2 + \frac{1}{2}\|v\|^2 + \frac{1}{2\lambda}\|Au \odot v - y\|^2$$

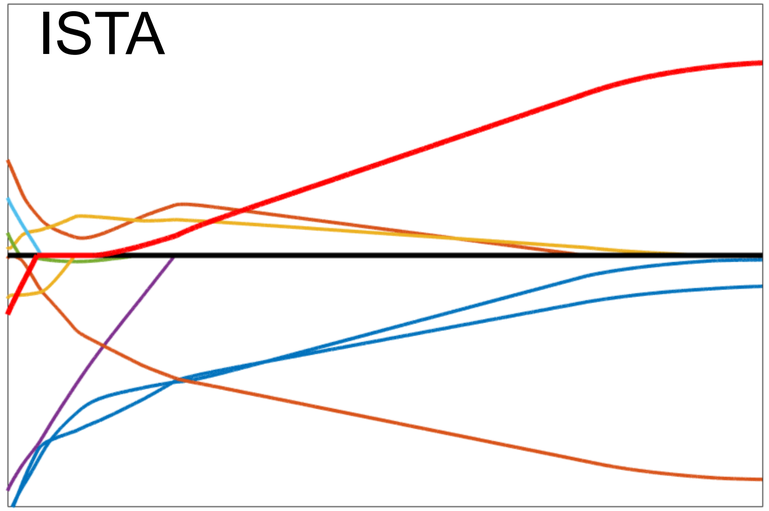
Computing the gradient:

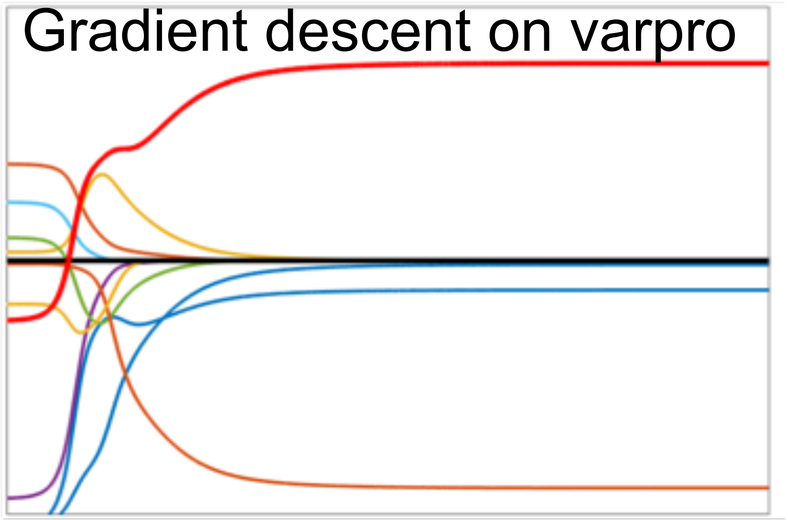
$$\nabla f(v) = v + \frac{1}{\lambda} u \odot A^\top (Au \odot v - y)$$

$$u = (\lambda I + \mathrm{diag}(v) A^\top A \mathrm{diag}(v))^{-1} (v \odot A^\top y)$$

Woodbury identity: $\nabla f(v) = v - v \odot (X^\top a)^2$

(allows $\lambda = 0$) $a = (A\mathrm{diag}(v^2)A^\top + \lambda I)^{-1} y$





Bilevel formulation

$$\min_v f(v)$$

$$f(v) := \min_u \frac{1}{2}\|u\|^2 + \frac{1}{2}\|v\|^2 + \frac{1}{2\lambda}\|Au \odot v - y\|^2$$

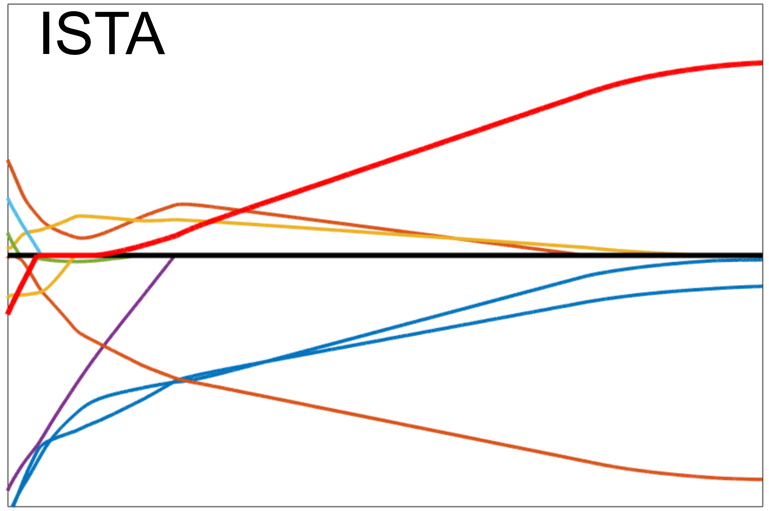
Computing the gradient:

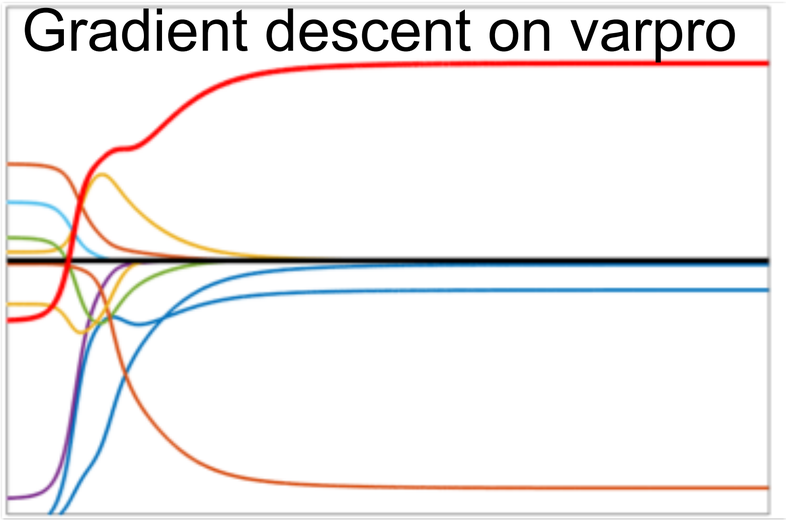
$$\nabla f(v) = v + \frac{1}{\lambda} u \odot A^\top (Au \odot v - y)$$

$$u = (\lambda I + \mathrm{diag}(v) A^\top A \mathrm{diag}(v))^{-1} (v \odot A^\top y)$$

Woodbury identity: $\nabla f(v) = v - v \odot (X^\top a)^2$

(allows $\lambda = 0$) $a = (A\mathrm{diag}(v^2)A^\top + \lambda I)^{-1} y$





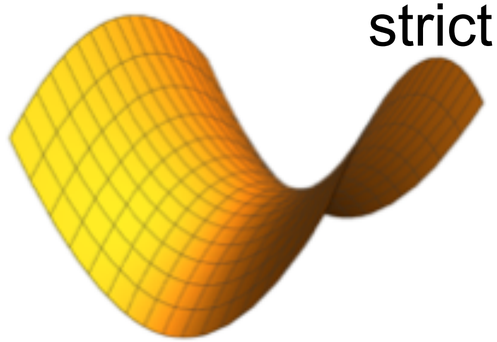
Need to handle $m \times m$ or $n \times n$ symmetric positive-definite linear system.

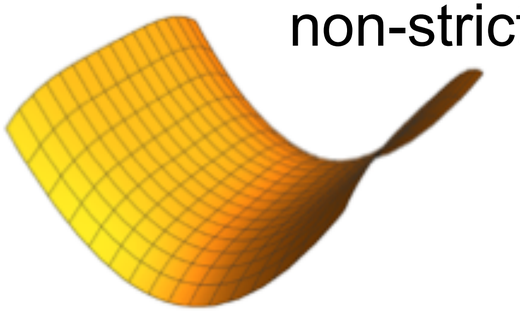
Mildly non-convex

Definition: v is a strict saddle point if

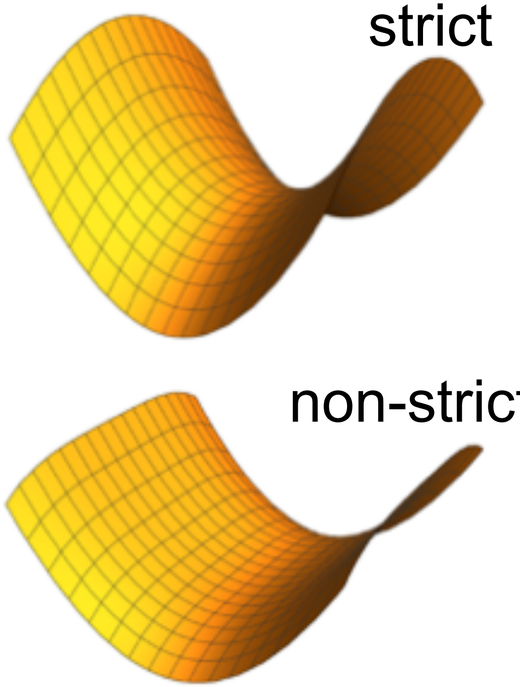
$\nabla f(v) = 0$ *but* $\nabla^2 f(v) \succeq 0$ *does not hold*

Lee et al (2017): Gradient descent almost always avoid strict saddles.





Mildly non-convex



Definition: v is a strict saddle point if

$\nabla f(v) = 0$ *but* $\nabla^2 f(v) \succeq 0$ *does not hold*

Lee et al (2017): Gradient descent almost always avoid strict saddles.

Primal: $x = u(v) \odot v$ Dual: $\xi = \frac{1}{\lambda} A^\top (Ax - y)$

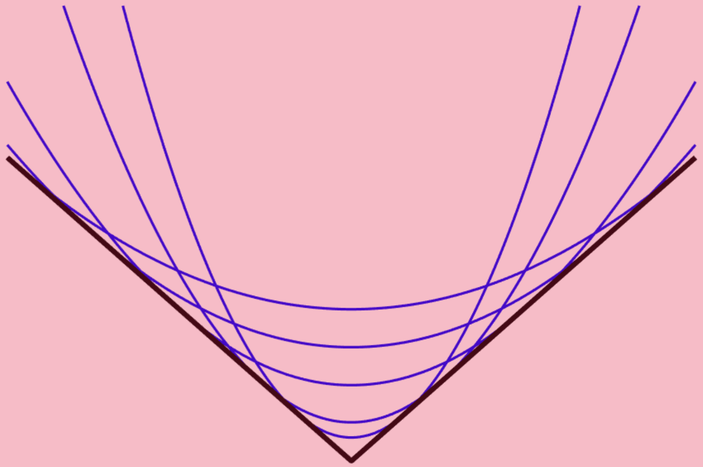
x global minimum $\Longleftrightarrow$ $\xi_I = \mathrm{sign}(x_I)$ and $|\xi|_\infty \leq 1,\ I = \mathrm{Supp}(x)$

Theorem: All stationary points of f are either global minima or strict saddles.

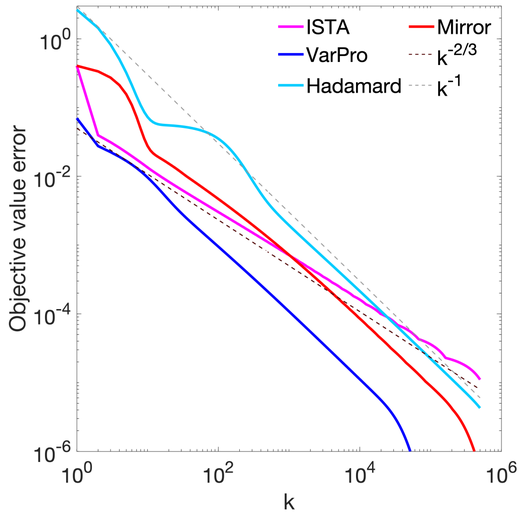
$$\nabla f(v) = 0 \text{ implies } \quad \mathrm{Eig}(\nabla^2 f(v)) \subseteq [1 - |\xi_{I_c}|_\infty, 4]$$

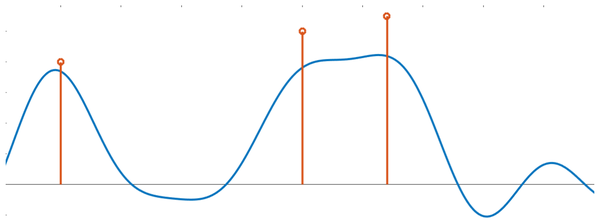
If the solution is nondegenerate $|\xi_{I_c}|_\infty < 1$, *then quasi-Newton methods achieve local super-linear convergence.*

Quadratic Variational Forms

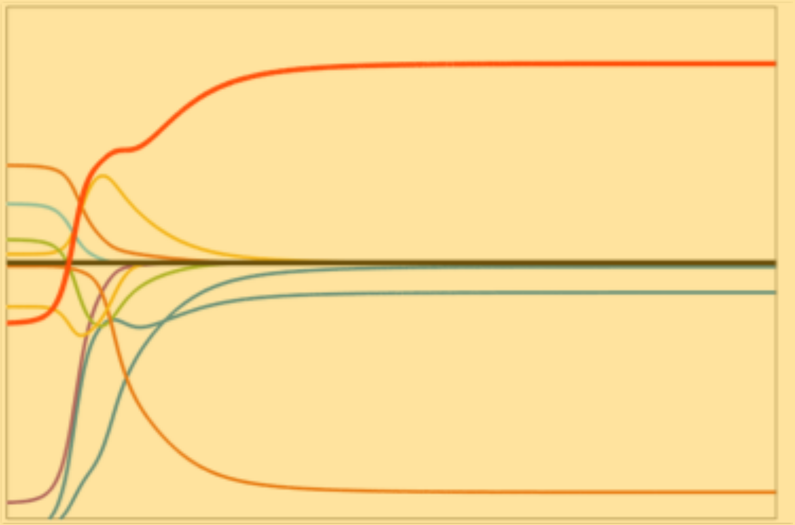


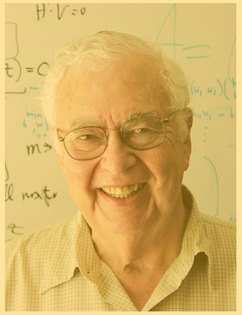
Convergence



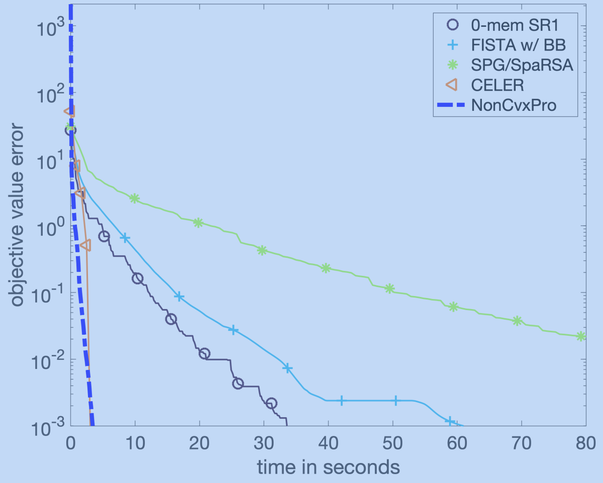


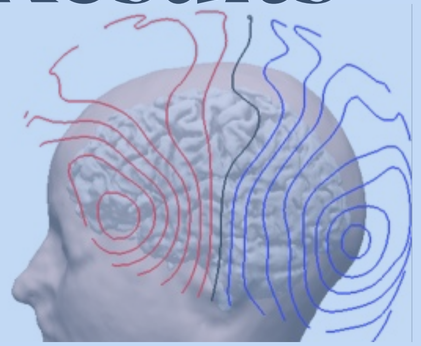
VarPro



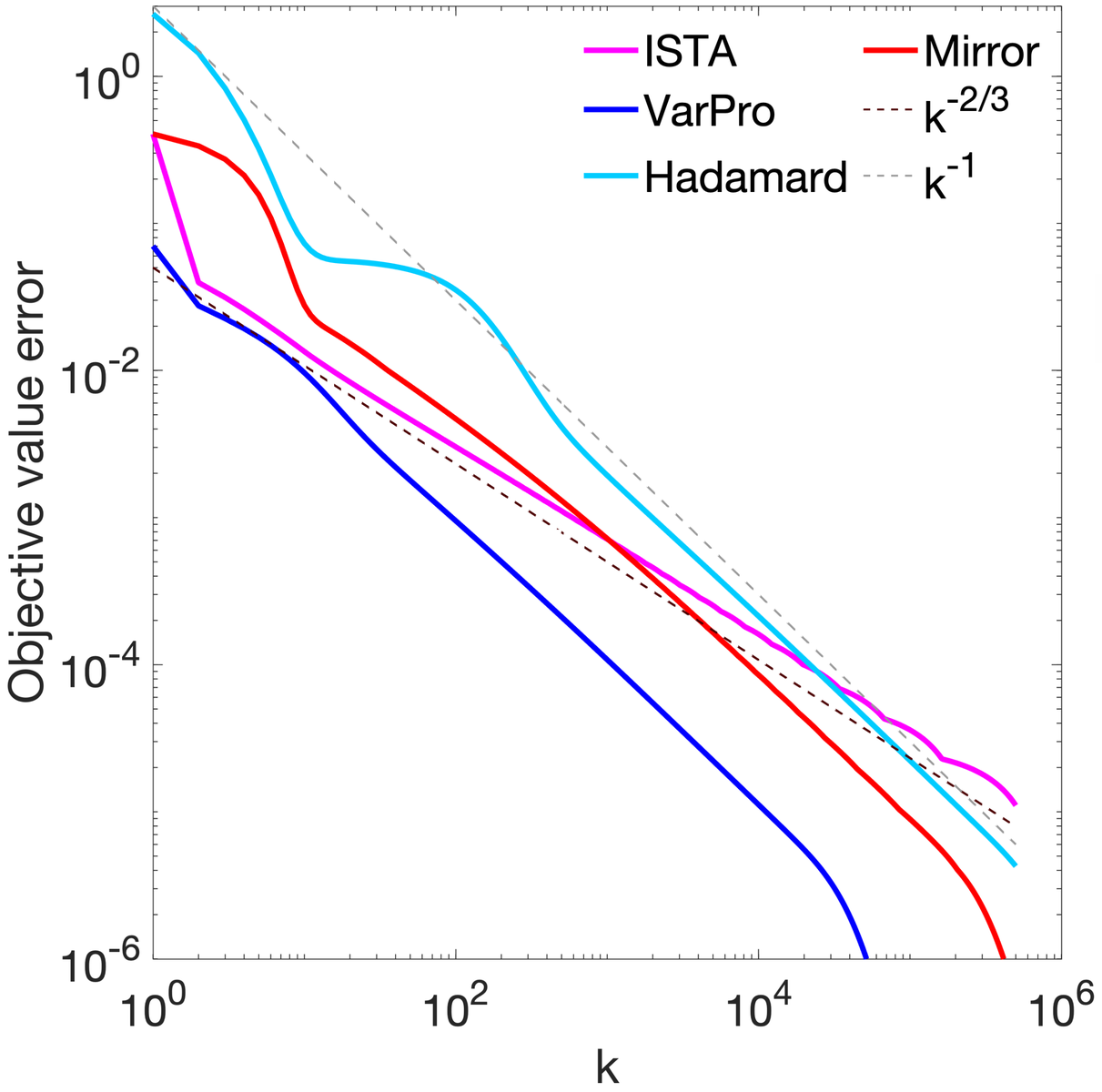


Numerical Results



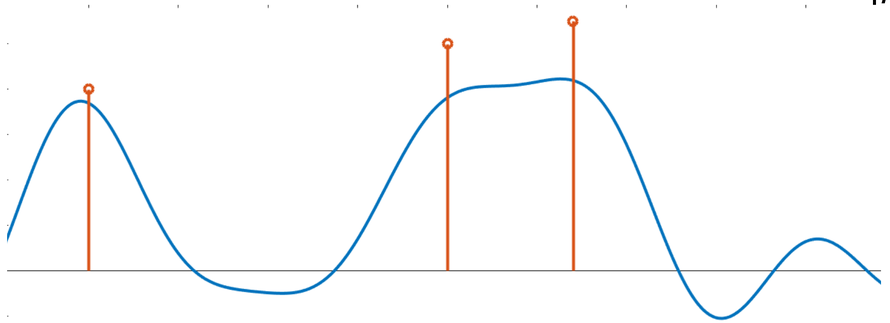


Observation

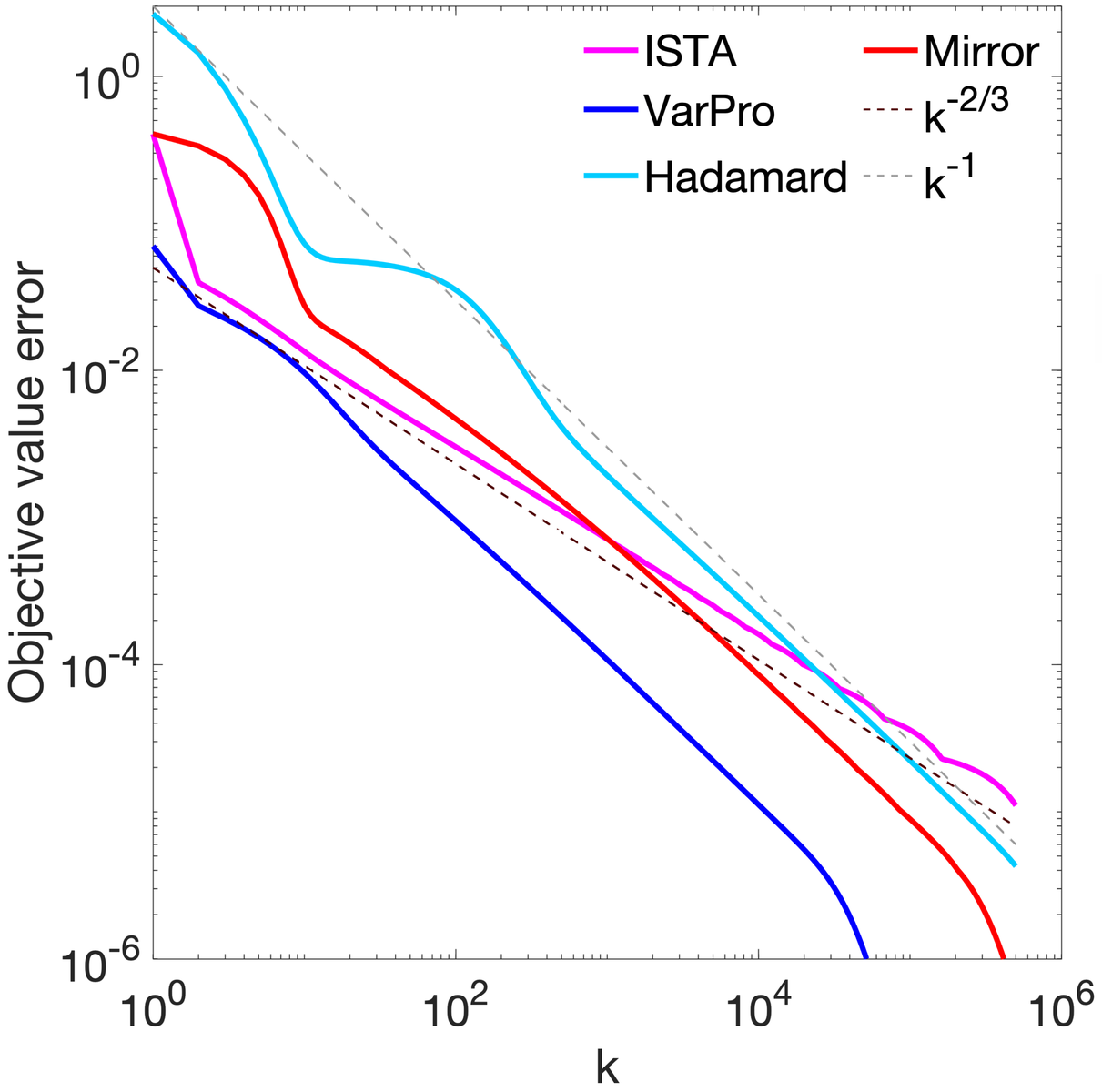


Solving the Lasso with a discretised Fourier operator ($n = 500$):

$$\text{Column i}: \quad A_i = \left(\exp(2\pi\sqrt{-1}ik/n)\right)_{|k|\le m}$$

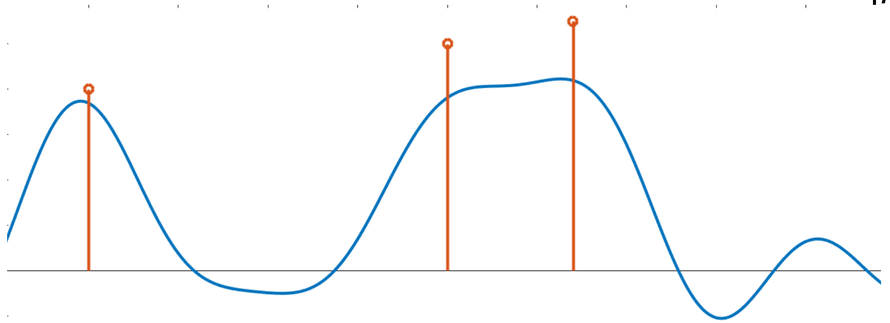


Observation



Solving the Lasso with a discretised Fourier operator ($n = 500$):

$$\text{Column i}: \quad A_i = \left(\exp(2\pi\sqrt{-1}ik/n)\right)_{|k|\le m}$$



Observations:

- ISTA converges at $\mathcal{O}(k^{-2/3})$ while proximal mirror descent converges at $\mathcal{O}(k^{-1})$ as shown by Chizat (2021).
- The Hadamard parameterisations also converge at $\mathcal{O}(k^{-1})$

Gradient Descent Convergence

$$G(u,v) = \frac{1}{2}\|u\|^2 + \frac{1}{2}\|v\|^2 + F(u \odot v)$$

$$\begin{cases} u_{k+1} = u_k - \tau(u_k + v_k \nabla F(u_k \odot v_k)) \\ v_{k+1} = v_k - \tau(v_k + u_k \nabla F(u_k \odot v_k)) \end{cases}$$

Assume:

- $\|\nabla F(x) - \nabla F(x')\|_\infty \leq L_1 \|x - x'\|_1$
- $\sup_{\|x\|_1 \leq B} \|\nabla F(x)\|_\infty \leq K$ where $B := G(u_0, v_0)$

If $F(x) = \|Ax - y\|^2$ and A has normalised columns, then $L_1, K = \mathcal{O}(1)$

Gradient Descent Convergence

$$G(u,v) = \frac{1}{2}\|u\|^2 + \frac{1}{2}\|v\|^2 + F(u \odot v)$$

$$\begin{cases} u_{k+1} = u_k - \tau(u_k + v_k \nabla F(u_k \odot v_k)) \\ v_{k+1} = v_k - \tau(v_k + u_k \nabla F(u_k \odot v_k)) \end{cases}$$

Assume:

- $\|\nabla F(x) - \nabla F(x')\|_\infty \le L_1 \|x - x'\|_1$
- $\sup_{\|x\|_1 \le B} \|\nabla F(x)\|_\infty \le K$ where $B := G(u_0, v_0)$

If $F(x) = \|Ax - y\|^2$ and A has normalised columns, then $L_1, K = \mathcal{O}(1)$

Proposition: $\nabla G(u,v)$ is Lipschitz with constant $L \sim K + L_1 B^2$,

$$\min_{j \le k} \|\nabla G(u_j, v_j)\|^2 \le \frac{2L}{k}(G(u_0, v_0) - \min_{u,v} G(u,v))$$

Corollary: If $u_k v_k \to x_*$ and $\|\nabla G(u_k, v_k)\| = \mathcal{O}(1/k)$, then $\Phi(u_k v_k) - \Phi(x_*) = \mathcal{O}(1/k)$

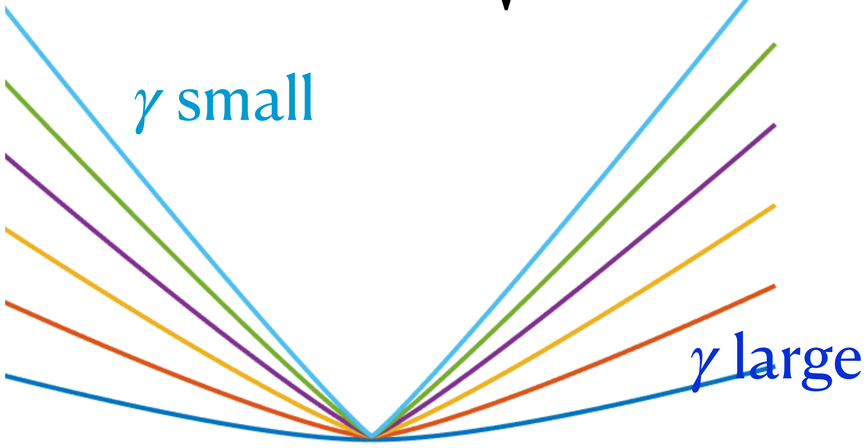
Mirror flow interpretation

Hadamard parametrised gradient flow as $\tau \to 0$

Let $x(t) := u(t) \odot v(t)$ and η_γ be the hyperbolic entropy with parameter $\gamma > 0$.

$$\frac{d}{dt} \nabla \eta_{\gamma(t)}(x(t)) = -2 \nabla F(x(t)) \qquad \gamma(t) = \frac{1}{2} e^{-2\lambda t} |u(0) - v(0)|$$

$$\eta_\gamma(x) = \gamma \operatorname{arsinh}(x/\gamma) - \sqrt{x^2 + \gamma^2} + \gamma$$



Recall Mirror Descent:

$$x_{k+1} \in \operatorname{argmin} F(x_k) + \frac{1}{n} \langle \nabla F(x_k), x - x_k \rangle + \frac{1}{\tau} D_\eta(x, x_k)$$

$$\nabla \eta(x_{k+1}) = \nabla \eta(x_k) - \tau \nabla F(x_k)$$

$$\frac{\mathrm{d}}{\mathrm{d}t} \nabla \eta(x(t)) = -\nabla F(x(t))$$

If $\lambda = 0$, then $\gamma(t) = \gamma(0)$ and links to works on implicit regularisation [e.g. Vaškevičius et al 2019]

From Gradient Descent to Quasi-Newton

Observations:

- $x_k = u_k \odot v_k$ is implicitly optimised in Hyperbolic geometry
- The parameters u_k, v_k operate in Euclidean geometry.

Robust acceleration tools are available (BB step, Quasi-Newton, BFGS, etc)!

$$\begin{cases} v_{k+1} & = v_k - \tau_k B_k \nabla f(v_k) \\ B_k & \approx [\nabla^2 f(v_k)]^{-1} \end{cases}$$

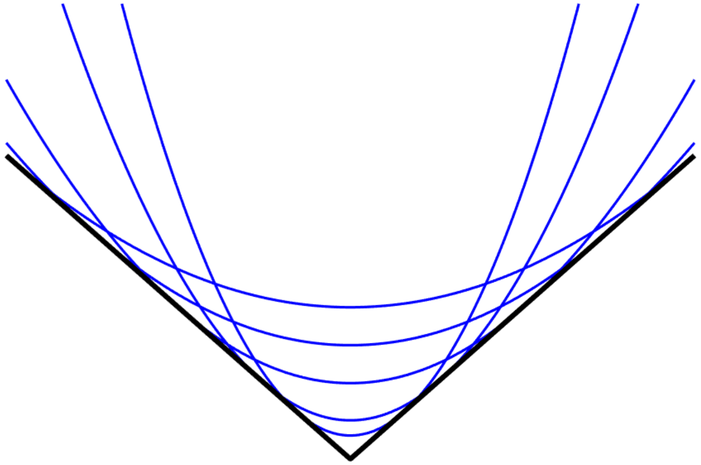
Secant condition:

$$\nabla f(v_{k+1}) - \nabla f(v_k) = \nabla^2 f(v_k)(v_{k+1} - v_k) + o(\|v_{k+1} - v_k\|)$$

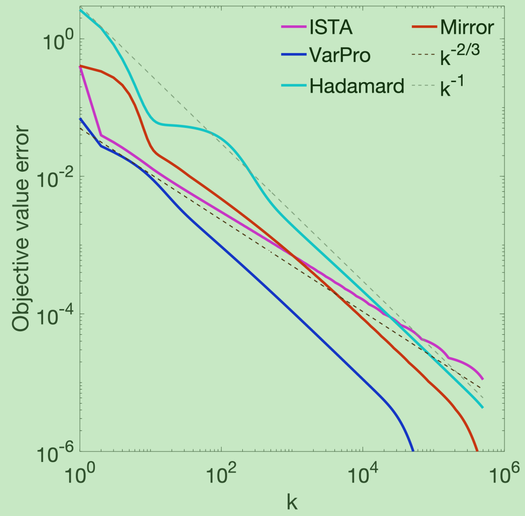
$$B_{k+1}(\nabla f(v_{k+1}) - \nabla f(v_k)) = v_{k+1} - v_k$$

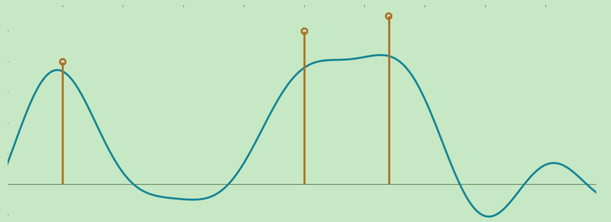
In practice, L-BFGS + Varpro leads to substantial performance gains over SOTA methods.

Quadratic Variational Forms

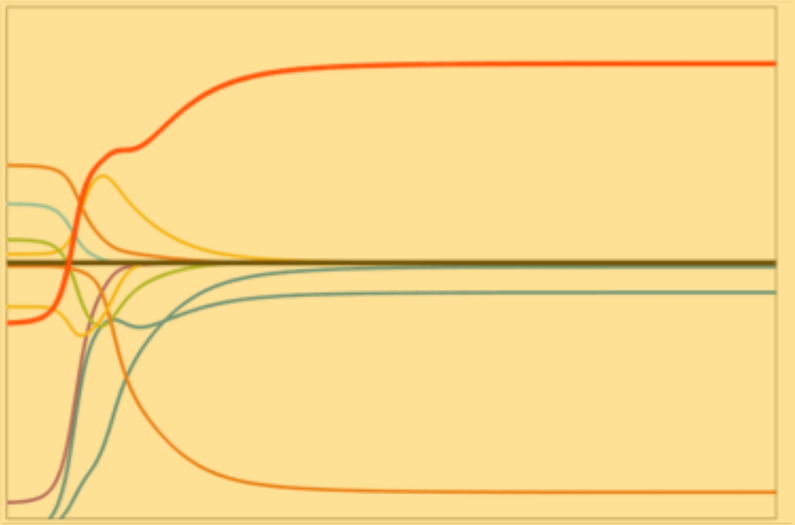


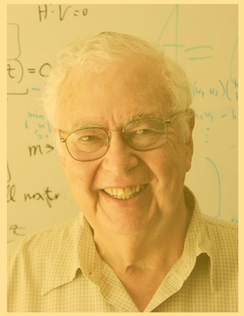
Convergence



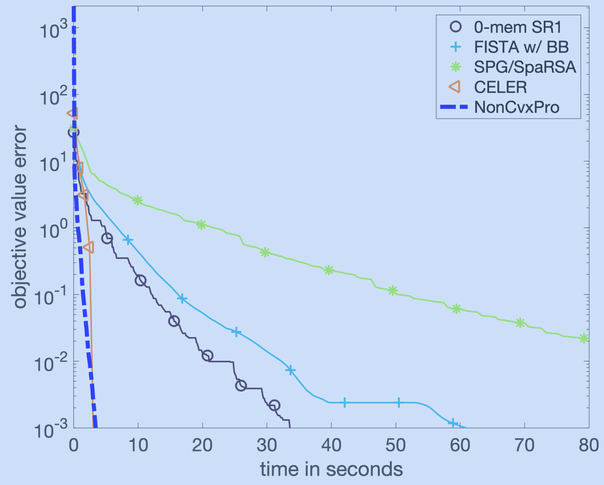


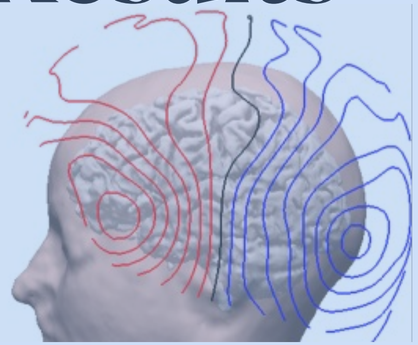
VarPro





Numerical Results





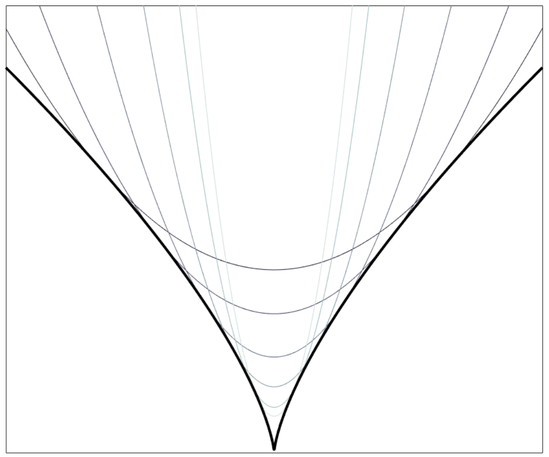
When do quadratic variational forms exist?

Group ℓ_1: Given disjoint $\cup_g g = \{1,\ldots,n\}$

$$\|x\|_{1,2} = \sum_g \|x_g\| = \min_{x_g = u_g v_g} \sum_g \frac{\|u_g\|^2}{2} + \frac{v_g^2}{2}$$

$$|x|^q \propto \min_{x=uv} u^2 + v^p, \quad q = \frac{2p}{p+2}$$

$$|X|_* = \min_{X=UV} \frac{\|U\|_F^2 + \|V\|_F^2}{2}$$



When do quadratic variational forms exist?

Group ℓ_1: Given disjoint $\cup_g g = \{1,\ldots,n\}$

$$\|x\|_{1,2} = \sum_g \|x_g\| = \min_{x_g = u_g v_g} \sum_g \frac{\|u_g\|^2}{2} + \frac{v_g^2}{2}$$

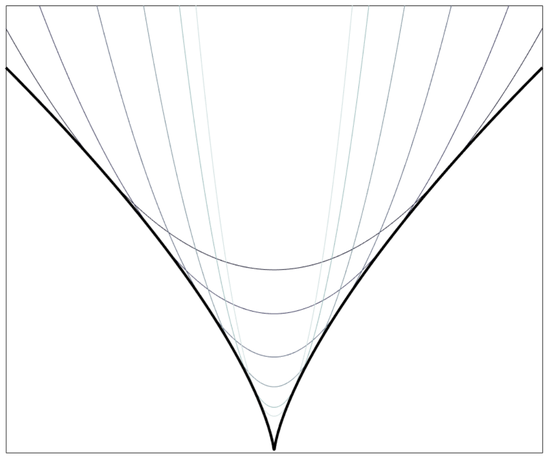
$$|x|^q \propto \min_{x=uv} u^2 + v^p, \quad q = \frac{2p}{p+2}$$

$$|X|_* = \min_{X=UV} \frac{\|U\|_F^2 + \|V\|_F^2}{2}$$

[Black & Rangarajan 1992] [Geman & Reynolds 1992]

Let $R : \mathbb{R}^n \to [0,\infty]$. The following are equivalent:

- $R(x) = \phi(x^2)$ where ϕ is concave.
- $R(x) = \inf_{\eta \in \mathbb{R}^n_+} \frac{1}{2} \sum_i \frac{x_i^2}{\eta_i} + h(\eta)$ where $h(\eta) = (-\phi)^*(-1/2\eta)$



General formulation

$$\min_{x \in \mathbb{R}^n} R(Lx) + F(Ax)$$

Assume:
$L \in \mathbb{R}^{p \times n}, A \in \mathbb{R}^{m \times n}$

$$R(x) = \inf_{\eta \in \mathbb{R}^n_+} \frac{1}{2} \sum_i \frac{|x_i|^2}{\eta_i} + h(\eta)$$

Over-parametrized form: $\min_{u,v} G(u, v)$

$$G(u, v) = h(v) + \min \left\{ \frac{1}{2}\|u\|^2 + F(Ax);\ Lx = u \odot v \right\}$$

General formulation

$$\min_{x \in \mathbb{R}^n} R(Lx) + F(Ax)$$

Assume:
$L \in \mathbb{R}^{p \times n}, A \in \mathbb{R}^{m \times n}$

$$R(x) = \inf_{\eta \in \mathbb{R}^n_+} \frac{1}{2} \sum_i \frac{|x_i|^2}{\eta_i} + h(\eta)$$

Over-parametrized form: $\min_{u,v} G(u, v)$

$$G(u, v) = h(v) + \min \left\{ \frac{1}{2}\|u\|^2 + F(Ax);\ Lx = u \odot v \right\}$$

Bilevel function.
$f(v) := \min_u G(u, v).$

General formulation

$$\min_{x\in\mathbb{R}^n} R(Lx) + F(Ax)$$

Assume:
$L \in \mathbb{R}^{p\times n}, A \in \mathbb{R}^{m\times n}$

$$R(x) = \inf_{\eta\in\mathbb{R}^n_+} \frac{1}{2}\sum_i \frac{|x_i|^2}{\eta_i} + h(\eta)$$

Over-parametrized form: $\min_{u,v} G(u,v)$

$$G(u,v) = h(v) + \min\left\{ \frac{1}{2}\|u\|^2 + F(Ax);\ Lx = u \odot v \right\}$$

Bilevel function.

$$f(v) := \min_u G(u,v).$$

Dual of inner function:

$$f(v) = h(v) + \max_{\xi\in\mathbb{R}^m,\alpha\in\mathbb{R}^p} -\frac{1}{2}\|v\alpha\|^2 - F^*(\xi) - \iota_{\{L^\top\alpha + A^\top\xi = 0\}}(\alpha,\xi)$$

General formulation

$$\min_{x\in\mathbb{R}^n} R(Lx) + F(Ax)$$

Assume:
$L \in \mathbb{R}^{p\times n}, A \in \mathbb{R}^{m\times n}$

$$R(x) = \inf_{\eta\in\mathbb{R}^n_+} \frac{1}{2}\sum_i \frac{|x_i|^2}{\eta_i} + h(\eta)$$

Over-parametrized form: $\min_{u,v} G(u,v)$

$$G(u,v) = h(v) + \min\left\{ \frac{1}{2}\|u\|^2 + F(Ax);\ Lx = u \odot v \right\}$$

Bilevel function.

$$f(v) := \min_u G(u,v).$$

Dual of inner function:

$$f(v) = h(v) + \max_{\xi\in\mathbb{R}^m,\alpha\in\mathbb{R}^p} -\frac{1}{2}\|v\alpha\|^2 - F^*(\xi) - \iota_{\{L^\top\alpha + A^\top\xi = 0\}}(\alpha,\xi)$$

Gradient formula:

$$\nabla f(v) = \nabla h(v) - v\alpha_v^2$$

- If $F \in C^{1,1}$, then f is differentiable.
- If $F = \iota_{\{y\}}$, f is differentiable at v if $Ax = y$ for some x and $\mathrm{Supp}(v) \supseteq \mathrm{Supp}(Lx)$

Nonsmooth Loss functions

$$\min_x R_1(Dx) + R_2(Ax - y)$$

$$R_i(z) = \min_{\eta \in \mathbb{R}^n_+} \frac{1}{2} \sum_i \frac{x_i^2}{\eta_i} + h_i(\eta) \text{ with } h_i \text{ smooth}$$

$$\min_x \lambda \|x\|_1 + \|Ax - y\|$$

Square root Lasso

$$\min_x \lambda \|Dx\|_1 + \|Ax - y\|_1$$

TV-L1

$$\min_X \lambda \|X\|_{1,2} + \|AX - Y\|_*$$

Multitask learning

Nonsmooth Loss functions

$$\min_x R_1(Dx) + R_2(Ax - y)$$

$$R_i(z) = \min_{\eta \in \mathbb{R}^n_+} \frac{1}{2} \sum_i \frac{x_i^2}{\eta_i} + h_i(\eta) \text{ with } h_i \text{ smooth}$$

$\Leftrightarrow$

$$\min_{v,w} f(v,w) := h_1(v) + h_2(w) + \phi(v,w)$$

$$\phi(v,w) = \max_{D^\top \alpha + A^\top \xi = 0} -\frac{1}{2}\|v\alpha\|^2 - \frac{1}{2}\|w\xi\|^2 + \langle \xi, y \rangle$$

Gradient formulas: $\partial_v f = \nabla h_1(v) - v\alpha^2$ $\partial_w f = \nabla h_2(w) + w\xi^2$

NB: inner problem is a quadratic problem

$$\min_x \lambda \|x\|_1 + \|Ax - y\|$$

Square root Lasso

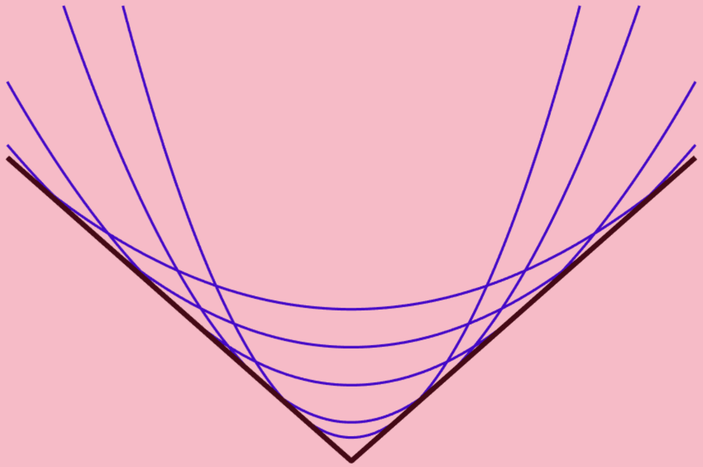
$$\min_x \lambda \|Dx\|_1 + \|Ax - y\|_1$$

TV-L1

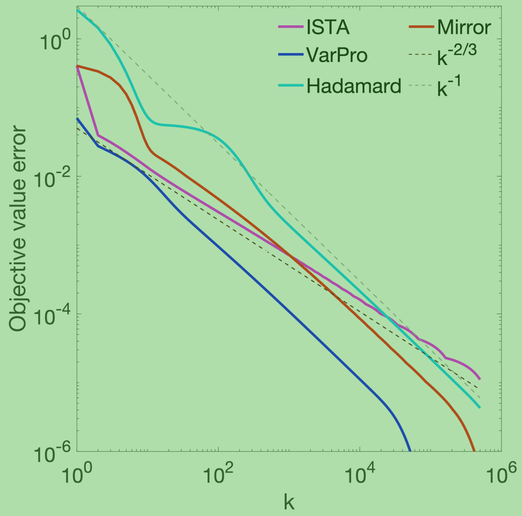
$$\min_X \lambda \|X\|_{1,2} + \|AX - Y\|_*$$

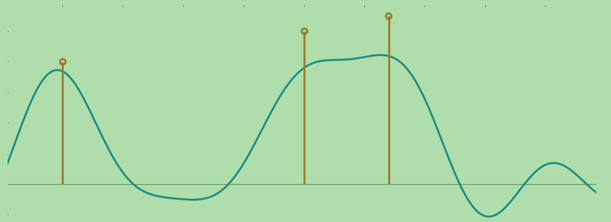
Multitask learning

Quadratic Variational Forms

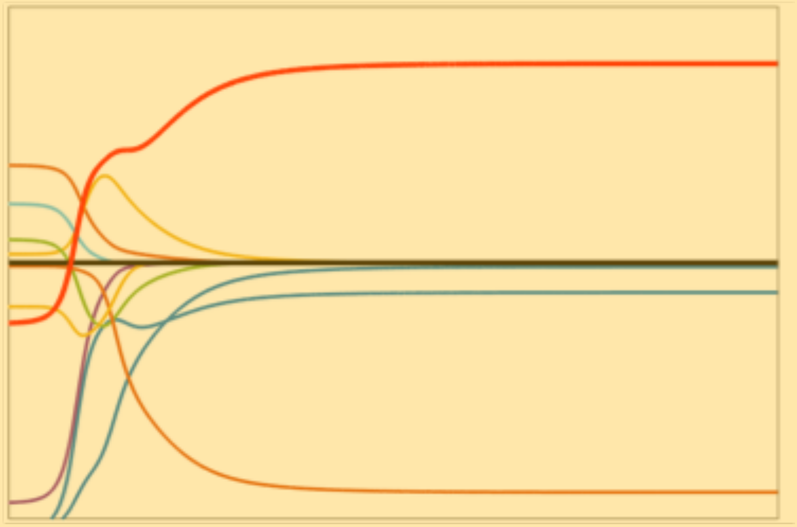


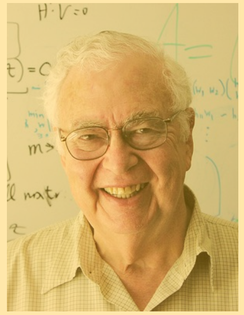
Convergence



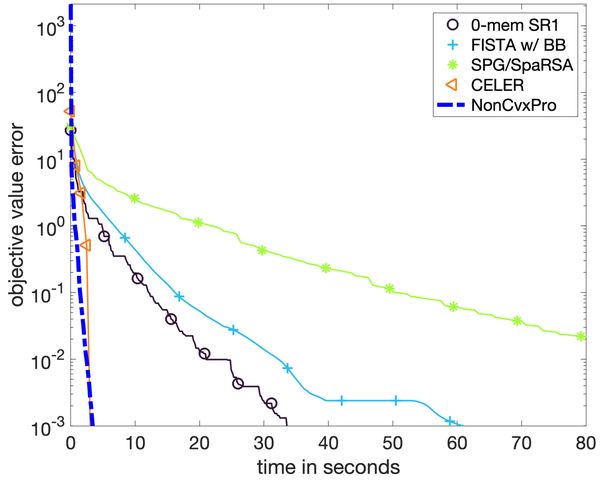


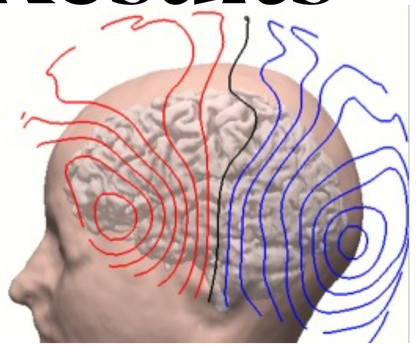
VarPro





Numerical Results



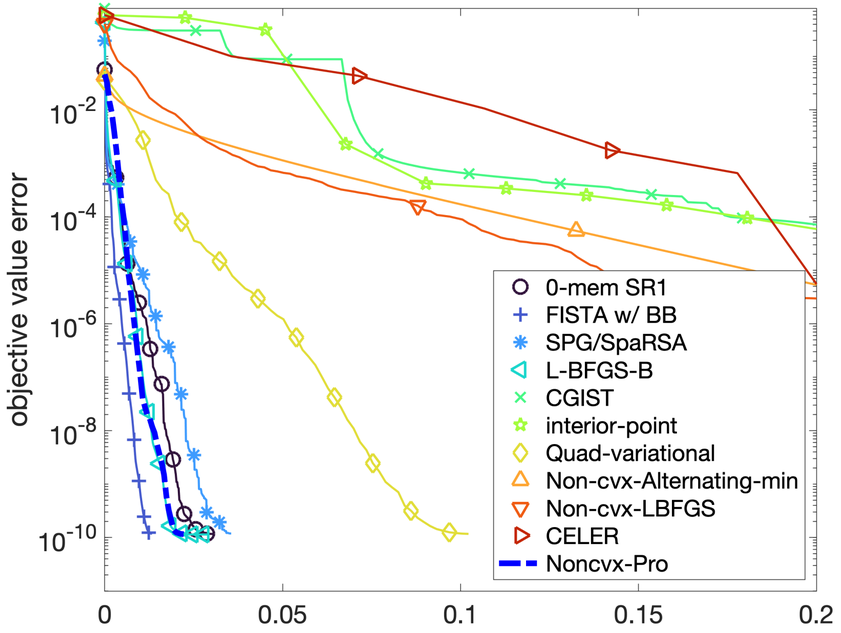


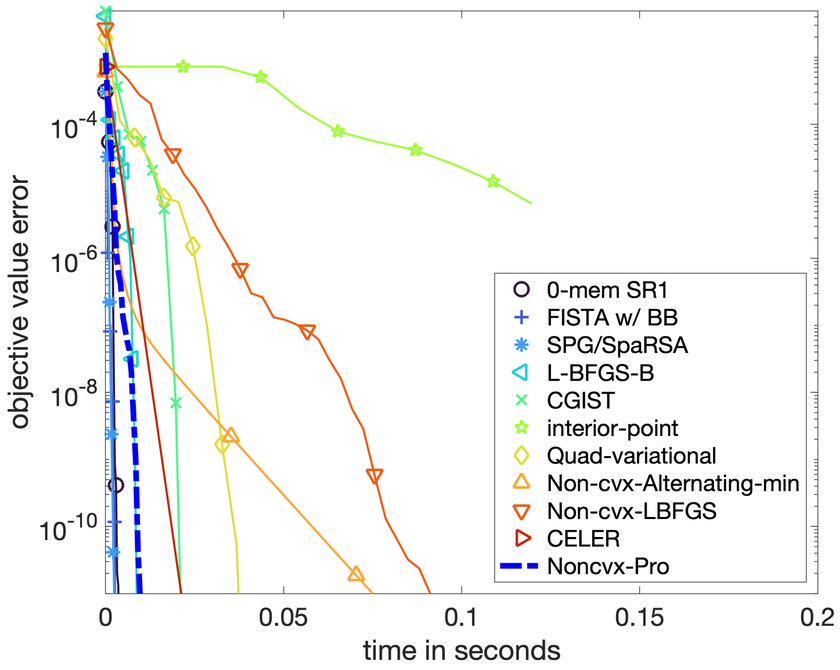
UCI/Adult income prediction

Lasso

a8a $(m, n) = (22696, 122)$

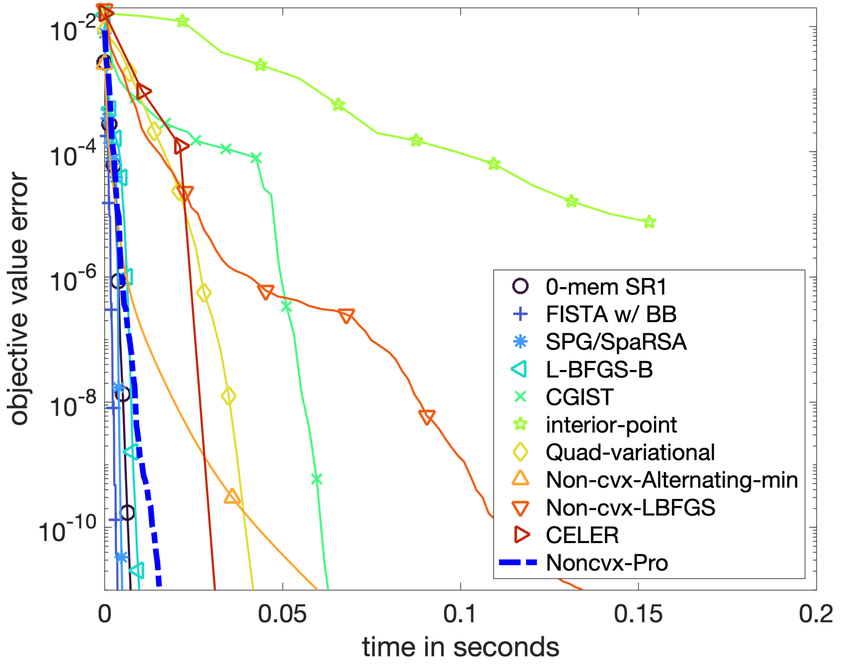
λ_*

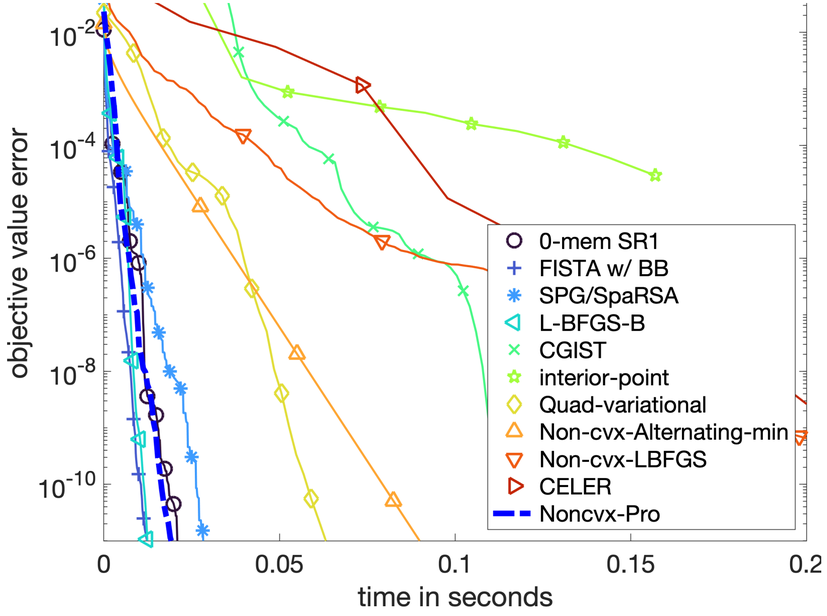




$\frac{1}{2}\lambda_{\max}$

$\frac{1}{10}\lambda_{\max}$





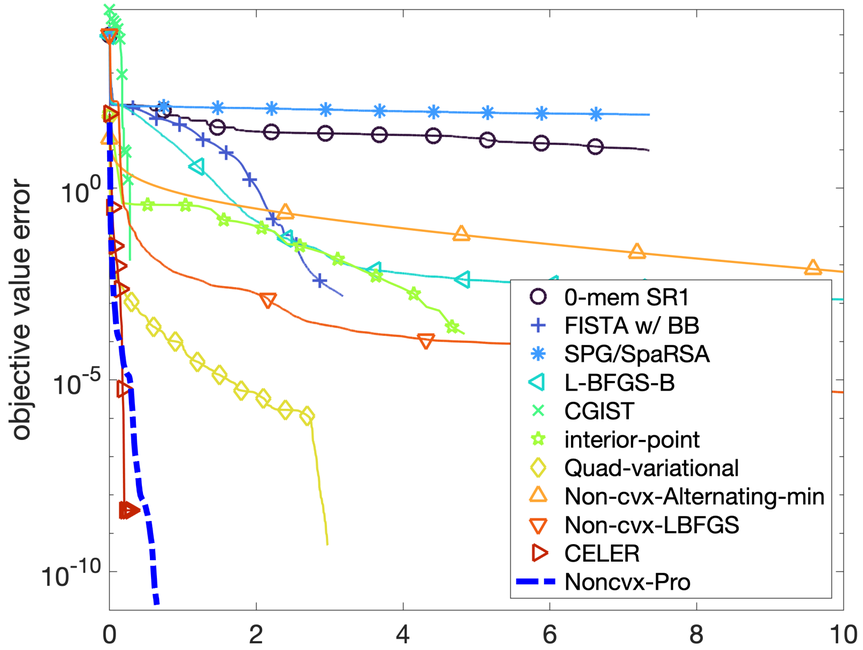
$\frac{1}{50}\lambda_{\max}$

Lasso

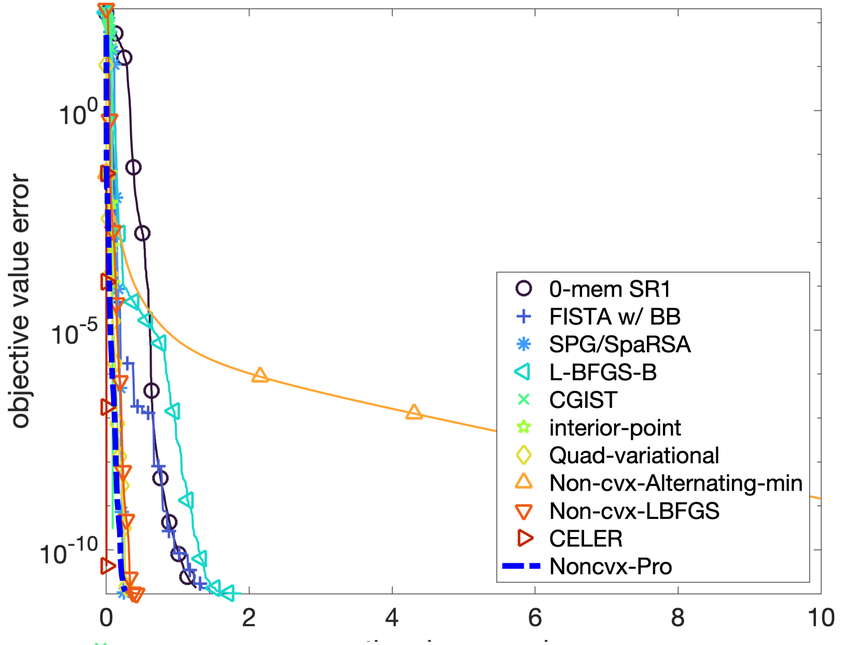
[Golub et al 1999] Molecular classification of cancer

Leukemia, $(m, n) = (38, 7129)$

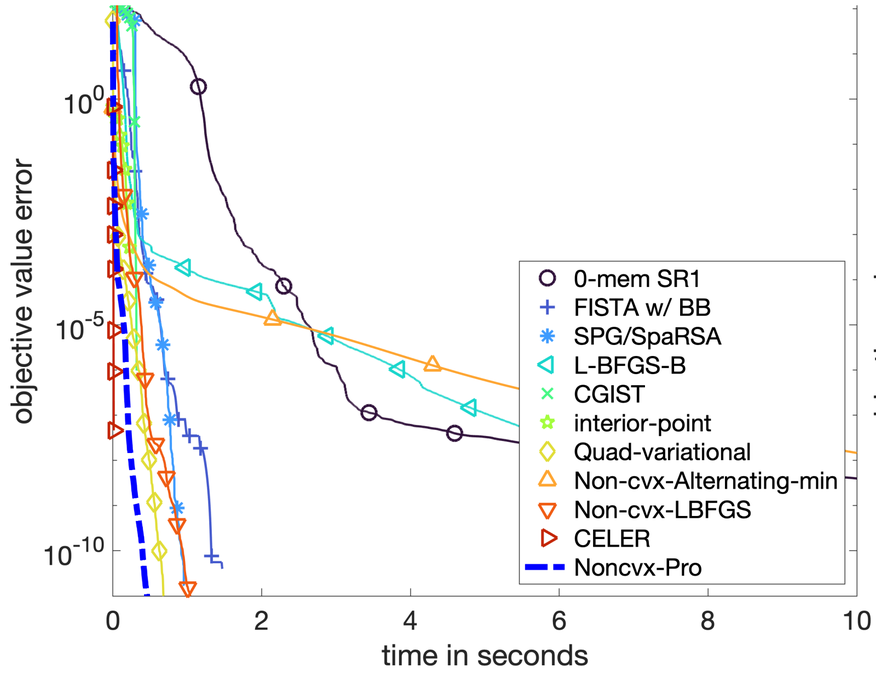
λ_*



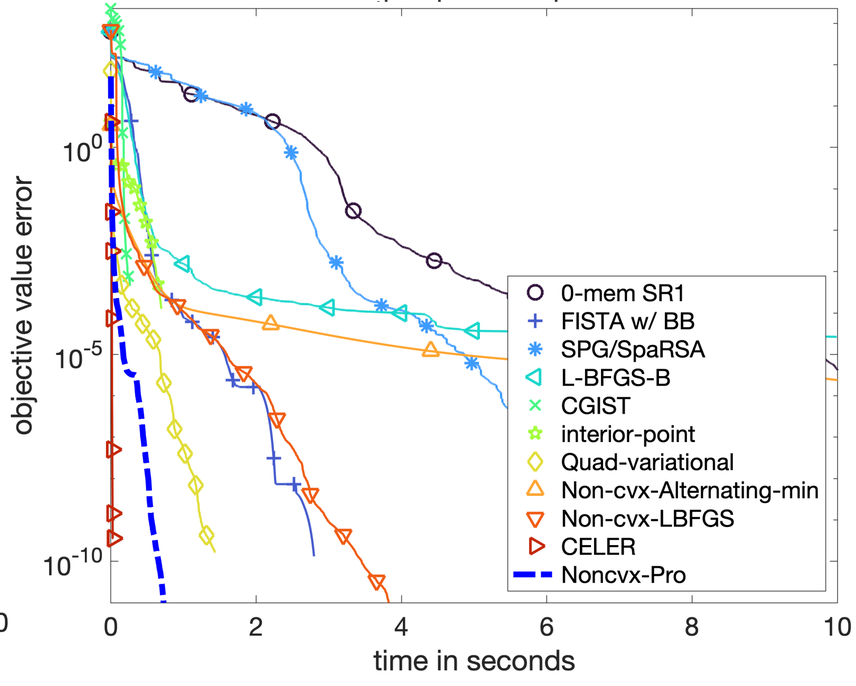
$\frac{1}{2}\lambda_{\max}$



$\frac{1}{10}\lambda_{\max}$

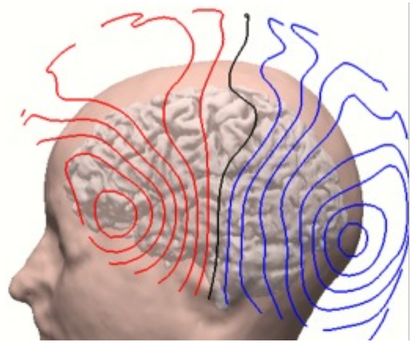


$\frac{1}{50}\lambda_{\max}$



[Ndiaye et al, 2015]

Group Lasso - MEG/EEG

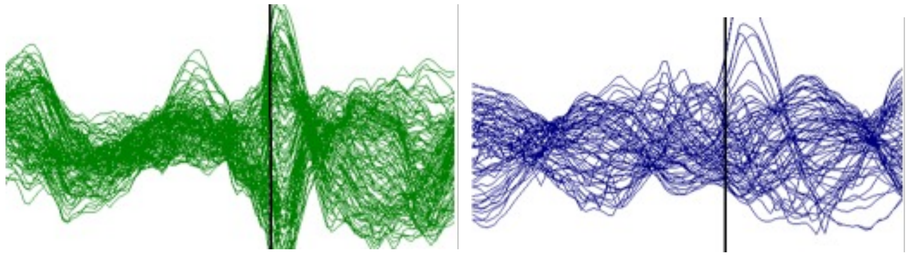


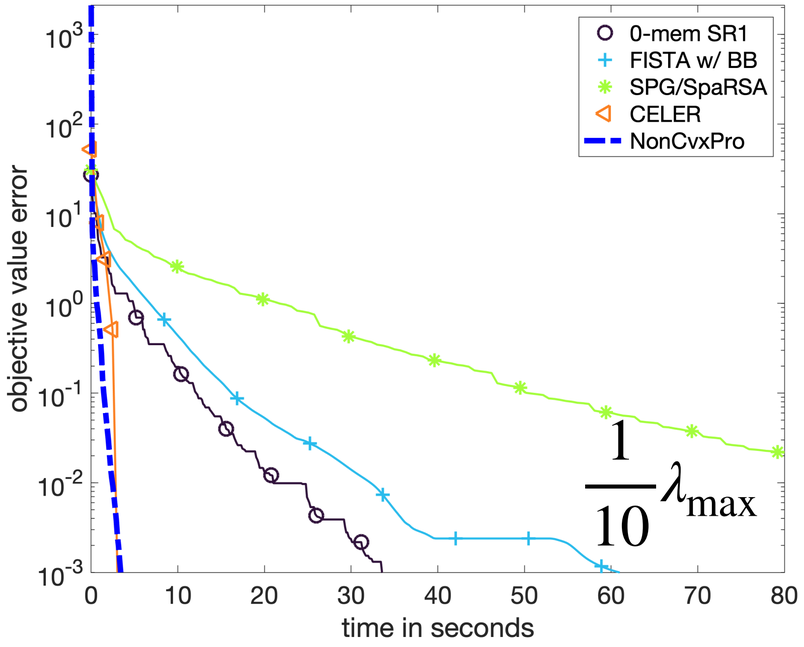
2294 source locations

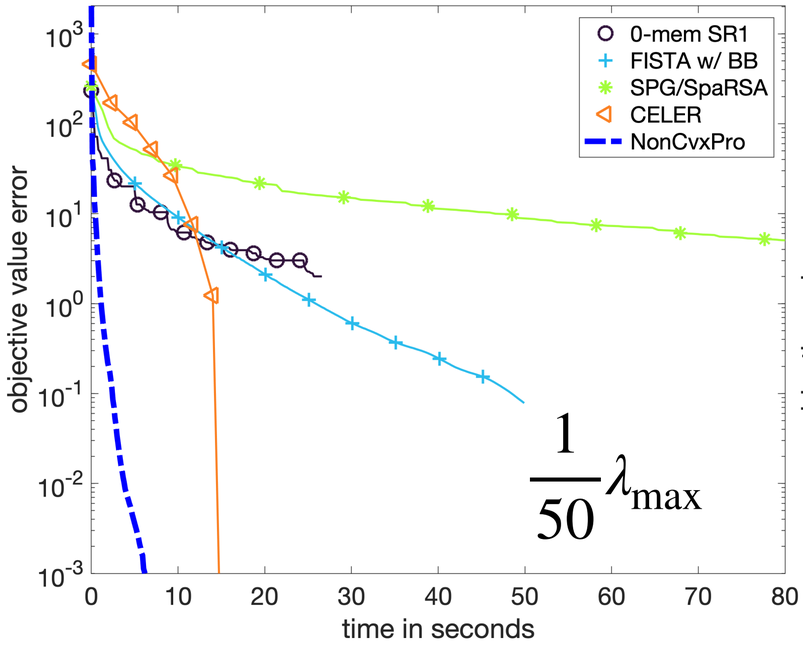
Group in time, $|g| = 181$

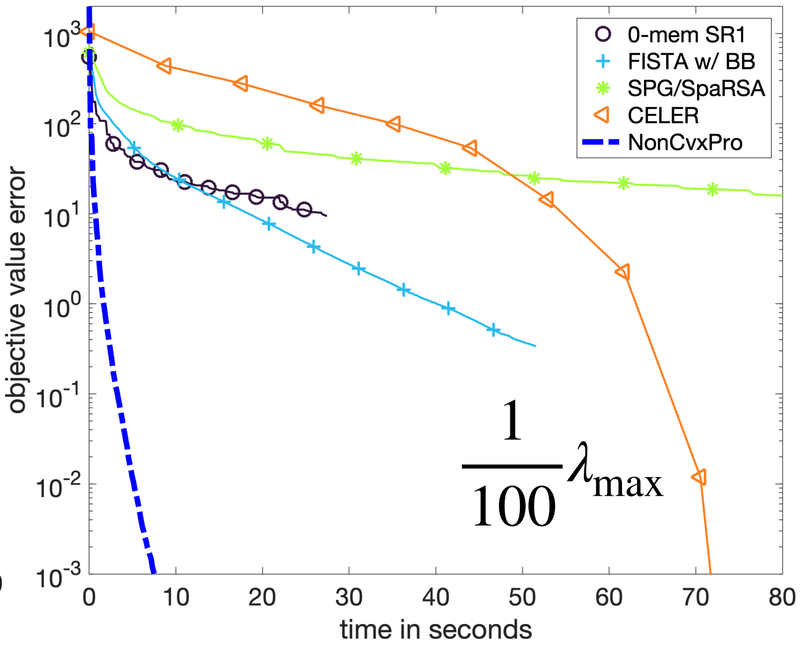
$\longrightarrow n = 2294 \times 181$

301 MEG + 59 EEG Sensors $\longrightarrow m = 360 \times 181$









Optimal Transport

$$W_1(\alpha,\beta) = \min_{\vec{w}} \left\{ \int \|\vec{w}(x)\|;\ \mathrm{div}(\vec{w}) = \alpha - \beta \right\}$$

Discretization:

$$A = \mathrm{div} \in \mathbb{R}^{n\times d}$$

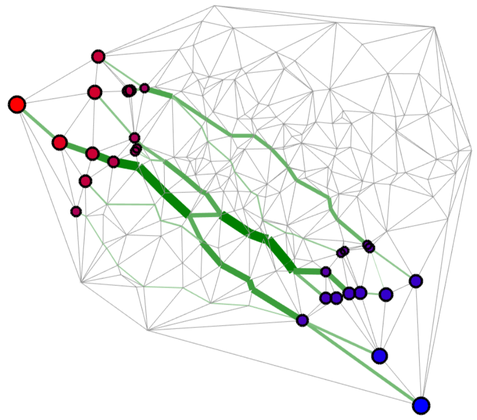
$$n = |\,\text{vertices}\,|\,,\ d = |\,\text{edges}\,|$$

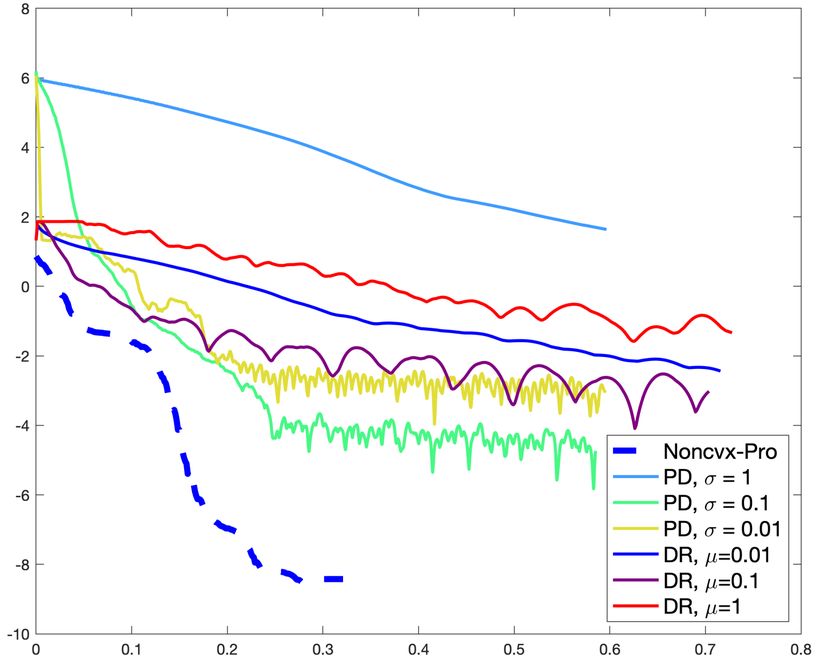
Graph:

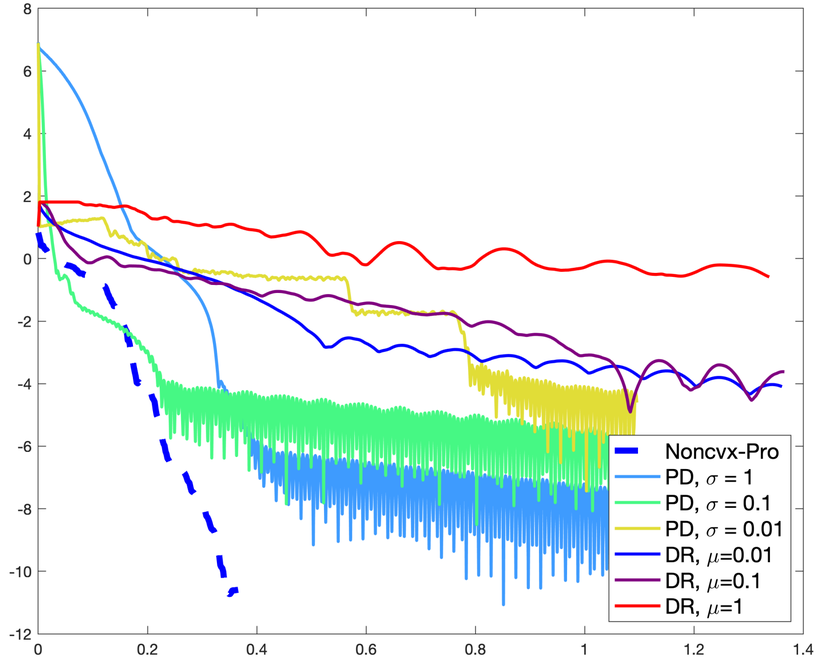
Lasso ($\lambda = 0$)

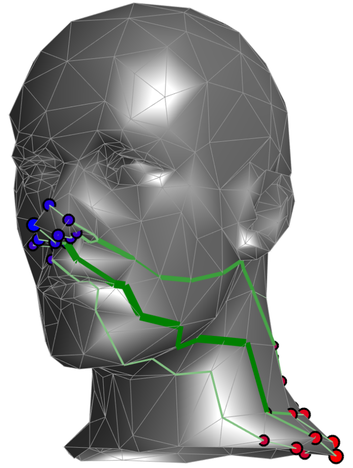
Finite elements:

Group Lasso ($\lambda = 0$)





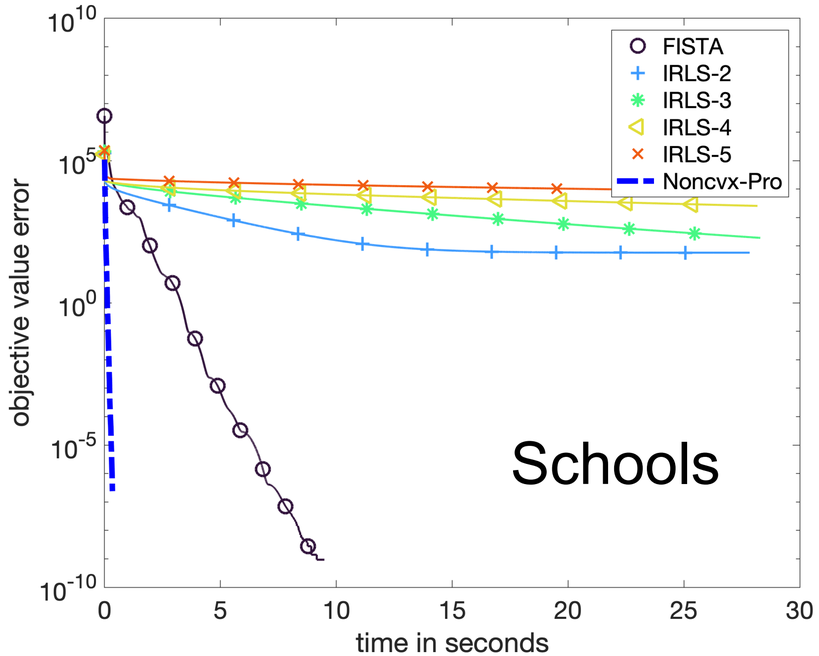




Multitask learning

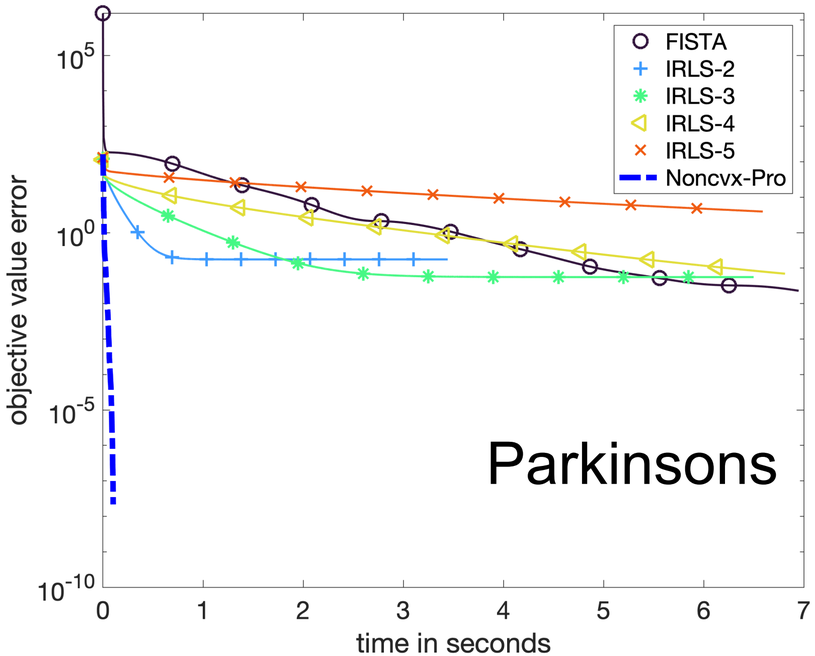
$$\min_{X=(x_t)_{t=1}^T} \frac{1}{2\lambda} \sum_{t=1}^{T} \|A_t x_t - y_t\|^2 + \|X\|_*$$

y = exam scores
T schools
n students



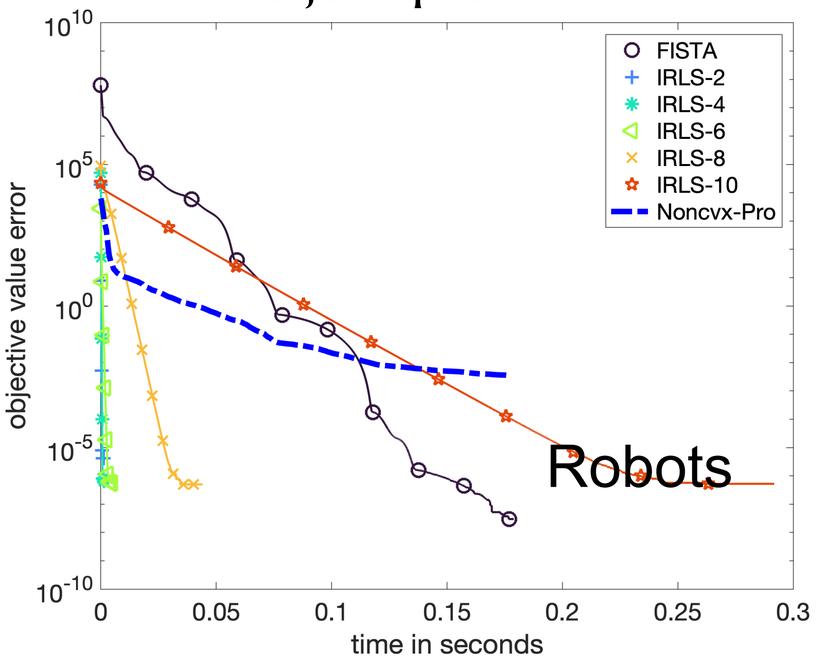
$(T, n, d) = (139, 15362, 27)$

y = symptoms
T patients
d bio-medical features



$(T, n, d) = (42, 5875, 19)$

y = motor
T robot arms
d joint positions



$(T, n, d) = (7, 48933, 21)$

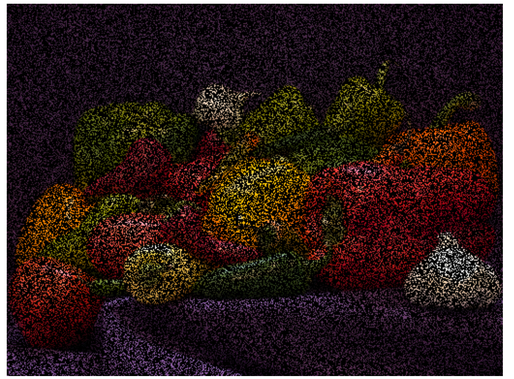
IRLS-d implies regularisation 10^{-d}

Image inpainting

A = masking operator

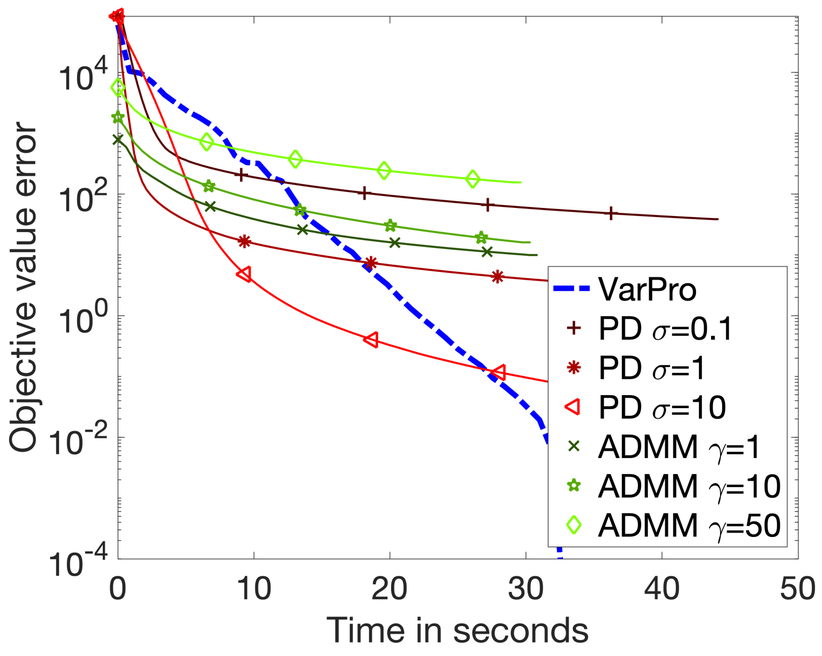
Loss: $\frac{1}{2\lambda}\|Ax - y\|^2$

$D : \mathbb{R}^{d\times d\times T} \to \mathbb{R}^{d\times d\times 2T}$ $\quad Dx = ((D^h x^t, D^h x^t))_{t=1}^T$

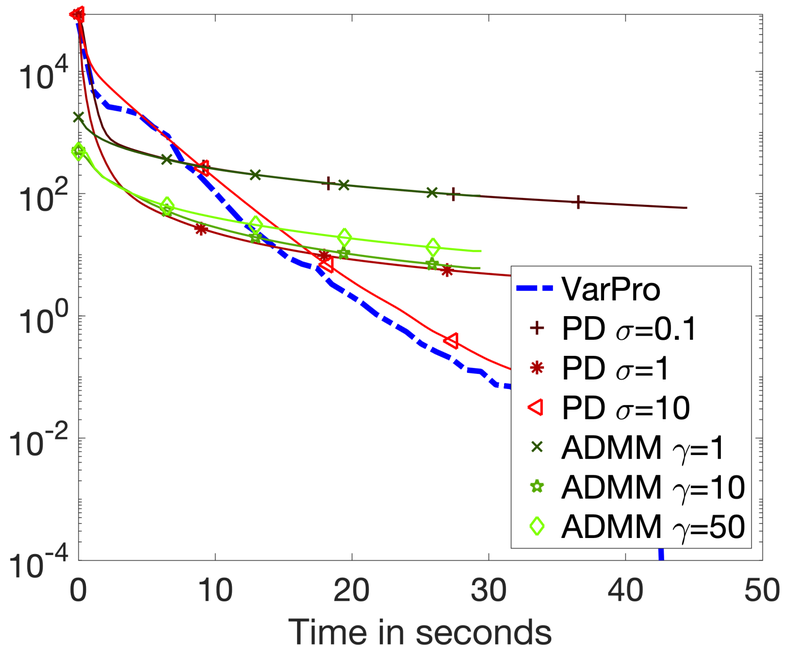


384×512

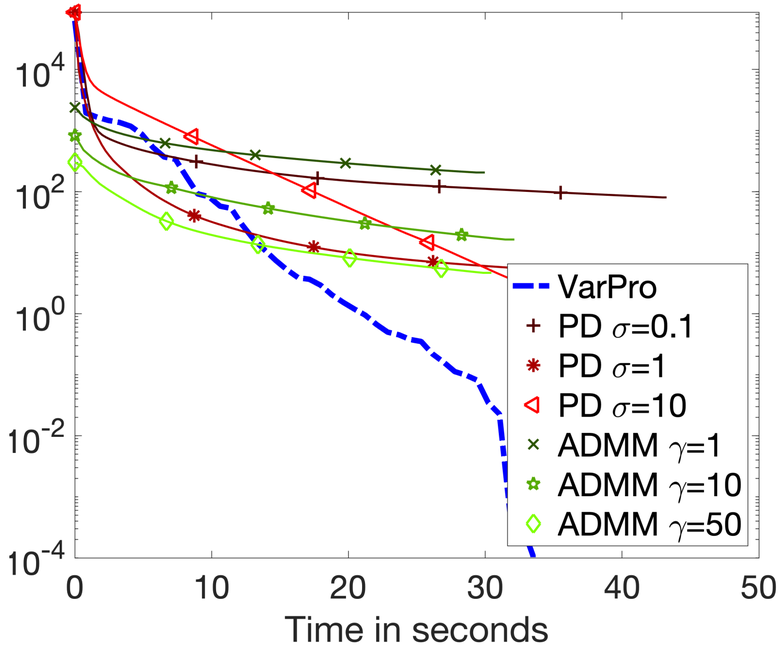
Group TV: $R(x) = \|Dx\|_{2,1} = \sum_{i,j=1}^{d} \sqrt{\sum_{t=1}^{T} (D^h x^t)_{i,j}^2 + (D^v x^t)_{i,j}^2}$



$\lambda = 0.1$



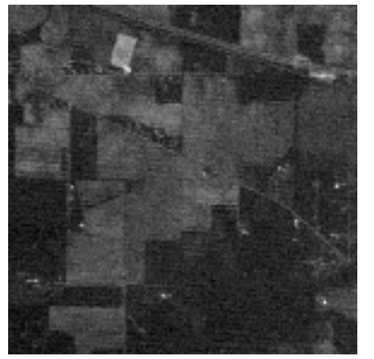
$\lambda = 0.5$



$\lambda = 1$

Hyperspectral imaging

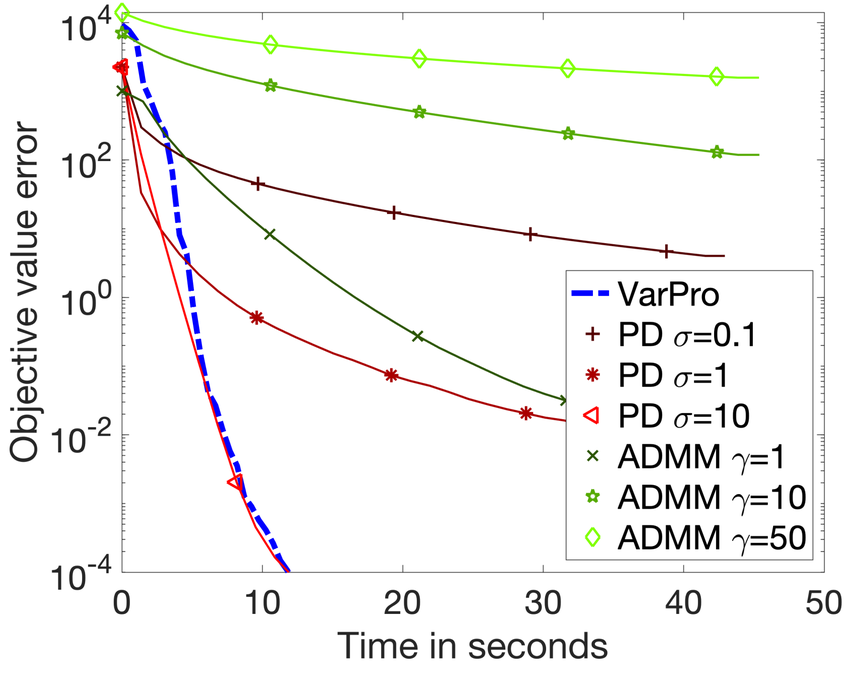
Group TV image denoising: $A = I$



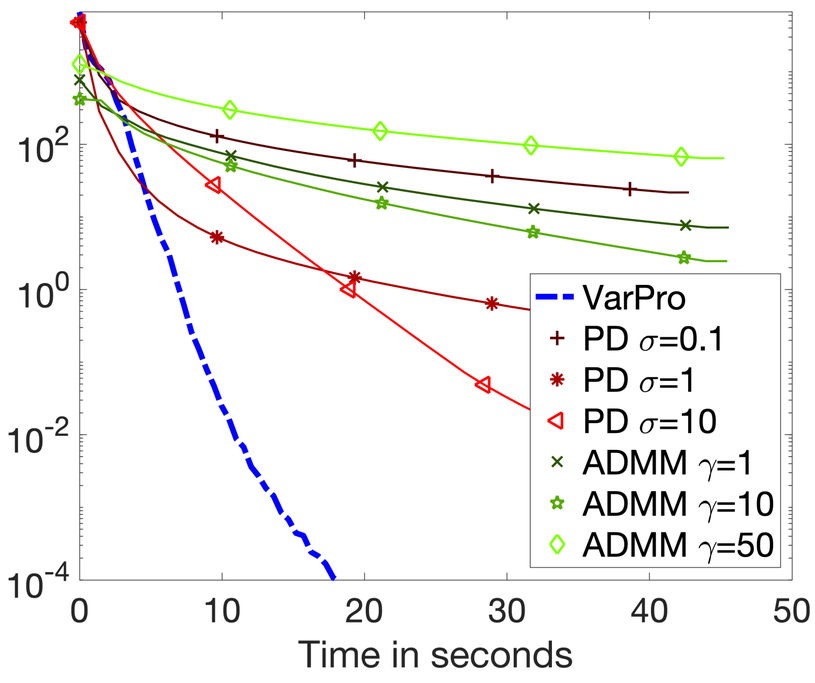
145×145

Indian Pines dataset

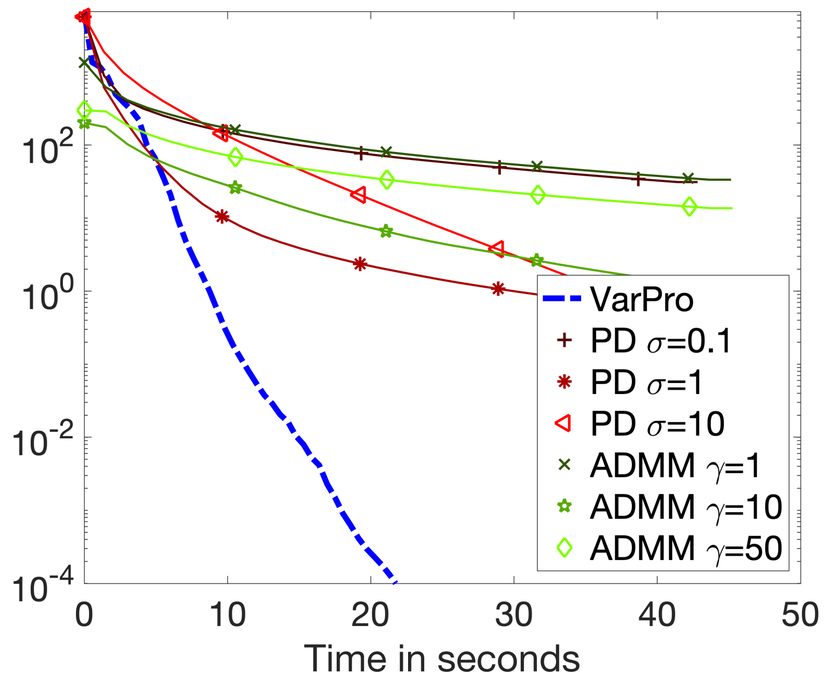
Input: noisy images of dimension $d = 145$ with $T = 224$ spectral bands



$\lambda = 0.1$



$\lambda = 0.5$



$\lambda = 1$

Conclusion

Practice

- *Trade*: Nonsmooth Convex $\Longleftrightarrow$ Smooth Nonconvex
- Simple algorithm and operates in Euclidean geometry. *Use L-BFGS*
- Handles a wide range of settings, including nonsmooth loss

Theory

- Reconditioning via Varpro
- Mild non-convexity.
- Hyperbolic geometry better than Euclidean in fine grids setting.