ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Winter Conference in Statistics 2013

Compressed Sensing

LECTURE #7-8

Algorithms for low-dimensional models

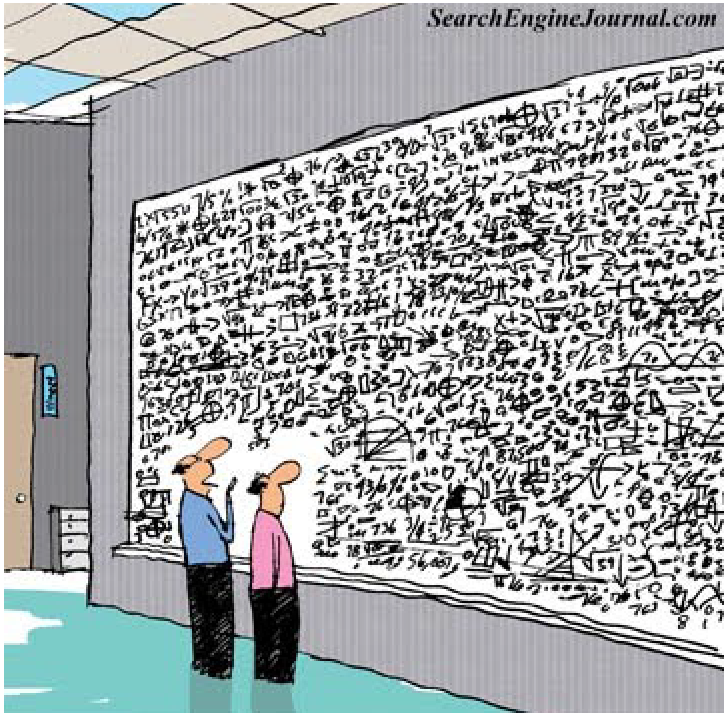
lions@epfl

Prof. Dr. Volkan Cevher

volkan.cevher@epfl.ch

LIONS/Laboratory for Information and Inference Systems

Convex Algorithms
for Low-Dimensional Models



...And, this is how you solve
huge-dimensional problems

The classical problem templates

$$\|x\|_1 = \sum_{i=1}^{N} \big|[x]_i\big|$$

Criteria seen above have the form

Basis pursuit (BP)
[Chen, Donoho, Saunders, 1998]

$$\min_x \|x\|_1 \quad \text{s.t.} \quad \Phi x = u$$

BP denoising (BPDN):
[Chen, Donoho, Saunders, 1998]

$$\min_x \|x\|_1 \quad \text{s.t.} \quad \|\Phi x - u\|_2^2 \leq \varepsilon$$

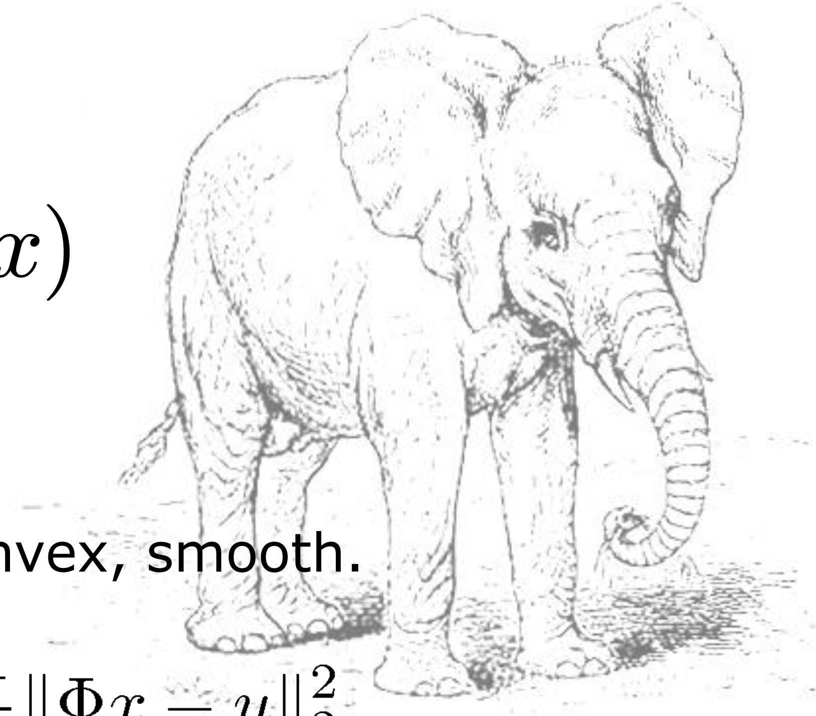
Also well known: LASSO (least absolute shrinkage/selection operator):
[Tibshirani, 1996]

$$\min_x \|\Phi x - u\|_2^2 \;\; \text{s.t.} \; \|x\|_1 \leq \tau$$

All can be written as

$$\widehat{x} \quad \in \quad \arg\min_{x \in \mathbb{R}^N} f_1(x) + f_2(x)$$

Convex optimization and proximal algorithms



$$\widehat{x} \in \arg\min_{x \in \mathbb{R}^N} f_1(x) + f_2(x)$$

$f_1 : \mathbb{R}^N \to \mathbb{R}$ data fidelity term; convex, smooth.

typically: $f_1(x) = \frac{1}{2}\|\Phi x - u\|_2^2$

$f_2 : \mathbb{R}^N \to \bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ Convex regularizer
(maybe non-smooth; e.g. ℓ_1)
(non-convex, later…).

Difficulties: non-smoothness and large dimension ($N \gg 1$)

Constrained vs unconstrained formulations

Constrained optimization formulations

$$\widehat{x} \in \arg\min_{x\in\mathbb{R}^N} f_1(x) \quad \text{s.t.} \; h(x) \leq \nu \qquad (*)$$

$$\widehat{x} \in \arg\min_{x\in\mathbb{R}^N} f_2(x) \quad \text{s.t.} \; g(x) \leq \tau$$

can be written as $$\widehat{x} \in \arg\min_{x\in\mathbb{R}^N} f_1(x) + f_2(x)$$

...using indicator functions: $$\iota_S(x) = \begin{cases} 0 & \Leftarrow \quad x \in S \\ +\infty & \Leftarrow \quad x \notin S \end{cases}$$

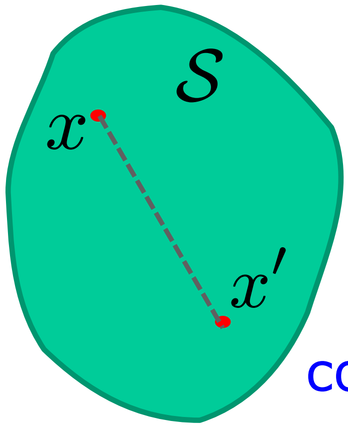
Example: $(*)$ same as

$$\widehat{x} \in \arg\min_{x\in\mathbb{R}^N} f_1(x) + \iota_{\{x: g(x)\leq\nu\}}(x)$$

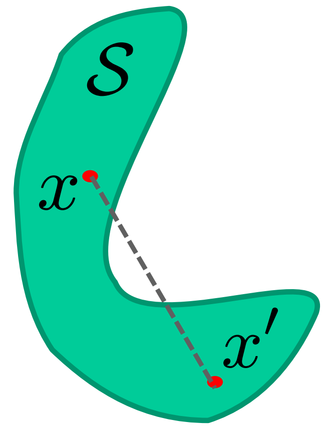
Classical example: the LASSO: $\min \|\Phi x - u\|_2^2 \; \text{s.t.} \; \|x\|_1 \leq \tau$

Convex and strictly convex sets

$\mathcal{S}$ is convex if $x, x' \in \mathcal{S} \Rightarrow \forall \lambda \in [0,1] \;\; \lambda x + (1-\lambda)x' \in \mathcal{S}$

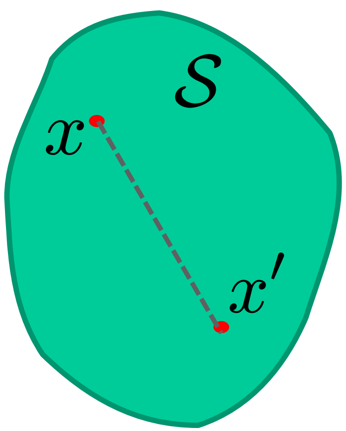


convex

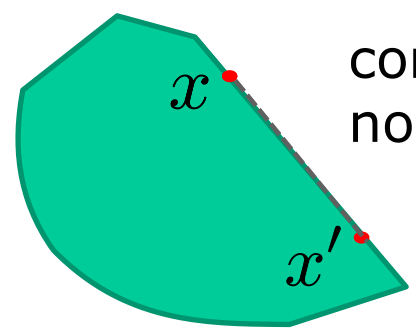


non-convex

$\mathcal{S}$ is strictly convex if $x, x' \in \mathcal{S} \Rightarrow \forall \lambda \in (0,1) \;\; \lambda x + (1-\lambda)x' \in \mathrm{int}(\mathcal{S})$



strictly convex



convex, but not strictly

Convex and strictly convex functions

Extended real valued function: $f : \mathbb{R}^N \to \bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$

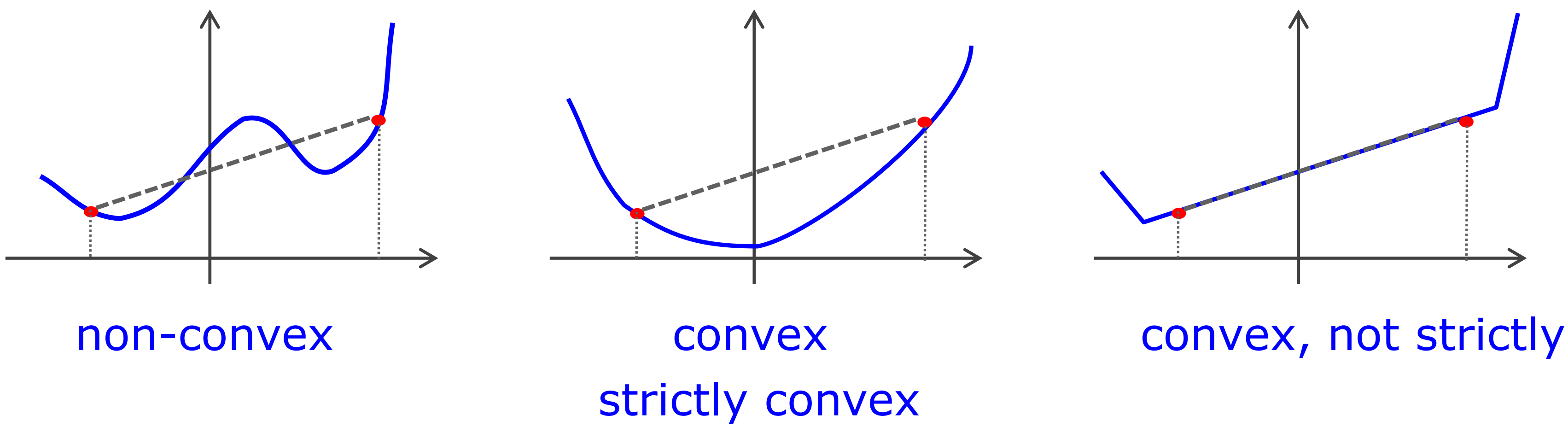
Domain of a function: $\mathrm{dom}(f) = \{x : f(x) \neq +\infty\}$

f is a convex function if

$$\forall \lambda \in [0,1], x, x' \in \mathrm{dom}(f) \quad f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x')$$

f is a strictly convex function if

$$\forall \lambda \in (0,1), x, x' \in \mathrm{dom}(f) \quad f(\lambda x + (1-\lambda)x') < \lambda f(x) + (1-\lambda) f(x')$$



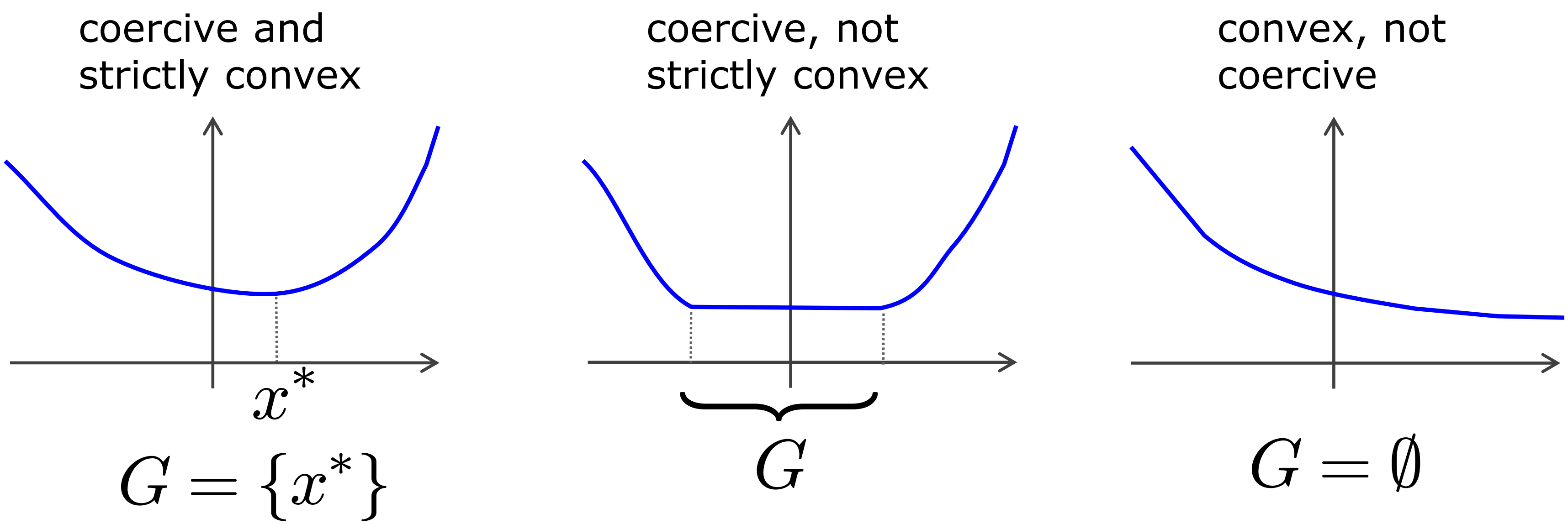
Convexity, coercivity, and minima

$$f : \mathbb{R}^N \to \bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$$

f is coercive if $\displaystyle\lim_{\|x\| \to +\infty} f(x) = +\infty$

if f is coercive, then $G \equiv \arg\min_x f(x)$ is a non-empty set

if f is strictly convex, then G has at most one element



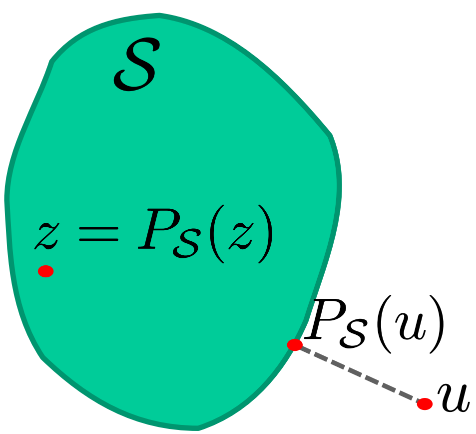
Euclidean projections on convex sets

Our problem: $\widehat{x} \in \arg\min_{x\in\mathbb{R}^n} f_1(x) + f_2(x)$

consider $f_2(x) = \iota_{\mathcal{S}}(x) = \begin{cases} 0 & \Leftarrow \quad x \in \mathcal{S} \\ +\infty & \Leftarrow \quad x \notin \mathcal{S} \end{cases}$

(convex if $\mathcal{S}$ is convex)

and $f_1(x) = \frac{1}{2}\|u - x\|_2^2$ (strictly convex)



$$\begin{aligned} \widehat{x} &= \arg\min_{x\in\mathbb{R}^n} f_1(x) + f_2(x) \\ &= \arg\min_{x\in\mathcal{S}} \|u - x\|_2^2 \\ &\equiv P_{\mathcal{S}}(u) \end{aligned}$$

(Euclidean projection)

Projected gradient algorithm

Our problem: $\widehat{x} \in \arg\min_{x \in \mathbb{R}^n} f_1(x) + f_2(x)$

with $f_2(x) = \iota_{\mathcal{S}}(x)$ ($\mathcal{S}$ is a convex set)

and f_1 some function, e.g., $f_1(x) = \frac{1}{2}\|\Phi x - u\|_2^2$

Projected gradient algorithm:

$$x_{k+1} = P_{\mathcal{S}}\Big(x_k - \beta_k \nabla f_1(x_k)\Big)$$

step size

if $f_1(x) = \frac{1}{2}\|\Phi x - u\|_2^2$

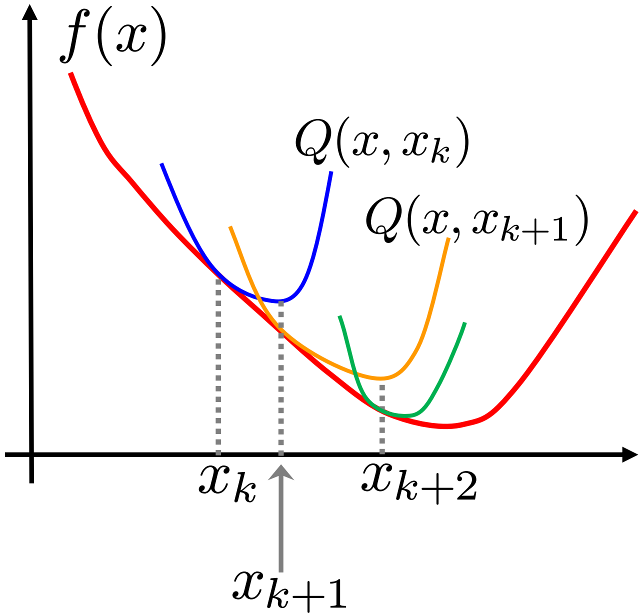
$$x_{k+1} = P_{\mathcal{S}}\Big(x_k - \beta_k \, \Phi^T(\Phi x_k - u)\Big)$$

Detour: majorization-minimization (MM)

Problem: $\widehat{x} \in \arg\min_{x \in \mathbb{R}^n} f(x)$

$Q(x, x_k)$ is a majorizer of f at x_k

$$Q(x, x_k) \geq f(x), \quad Q(x_k, x_k) = f(x_k)$$



MM algorithm:

$$x_{k+1} = \arg\min_{x} Q(x, x_k)$$

monotonicity:

$$\begin{aligned} f(x_{k+1}) &\leq Q(x_{k+1}, x_k) \\ &\leq Q(x_k, x_k) \\ &= f(x_k) \end{aligned}$$

Projected gradient from majorization-minimization

Our problem: $\widehat{x} \quad \in \quad \arg\min_{x\in\mathbb{R}^n} f_1(x) + f_2(x)$

with $f_2(x) = \iota_{\mathcal{S}}(x)$ ($\mathcal{S}$ is a convex set)

and f_1 has L -Lipschitz gradient

$$\|\nabla f_1(x) - \nabla f_1(x')\| \leq L\|x - x'\|$$

e.g. $f_1(x) = \dfrac{1}{2}\|\Phi x - u\|_2^2 \;\Rightarrow\; L = \lambda_{\max}(\Phi^T\Phi) = \|\Phi\|_2^2$

Hessian of f_1

...a separable approximation of f_1

$$Q(x, x_k) = f_1(x_k) + (x - x_k)^T \nabla f_1(x_k) + \frac{1}{2\,\beta_k}\|x - x_k\|_2^2$$

Projected gradient from majorization-minimization

Our problem: $\widehat{x} \in \arg\min_{x \in \mathbb{R}^n} f_1(x) + \iota_{\mathcal{S}}(x)$

Separable approximation of f_1

$$Q(x, x_k) = f_1(x_k) + (x - x_k)^T \nabla f_1(x_k) + \frac{1}{2\beta_k} \|x - x_k\|_2^2$$

$Q(x, x_k)$ is a majorizer of f_1, if $\boxed{\beta_k < \frac{1}{L}}$

$Q(x, x_k) + \iota_{\mathcal{S}}(x)$ is a majorizer $f_1(x) + \iota_{\mathcal{S}}(x)$

MM algorithm:

$$\begin{aligned} x_{k+1} &= \arg\min_x Q(x, x_k) + \iota_{\mathcal{S}}(x) \\ &= \arg\min_x \frac{1}{2\beta_k} \left\| x - \big(x_k - \beta_k \nabla f_1(x_k)\big) \right\|_2^2 + \iota_{\mathcal{S}}(x) \\ &= P_{\mathcal{S}}\Big(x_k - \beta_k \nabla f_1(x_k) \Big) \end{aligned}$$

...projected gradient.

Proximity operators

Our problem: $\widehat{x} \in \underset{x \in \mathbb{R}^n}{\arg\min}\, f_1(x) + f_2(x)$

with f_2 a convex function

and $f_1(x) = \frac{1}{2}\|u - x\|_2^2$ (strictly convex)

$$\widehat{x} = \underset{x \in \mathbb{R}^n}{\arg\min}\, \frac{1}{2}\|u - x\|_2^2 + f_2(x) \equiv \mathrm{prox}_{f_2}(u)$$

Proximity operator [Moreau 62], [Combettes 01].

Generalizes the notion of Euclidean projection.

Proximity operators (linear)

$$\mathrm{prox}_f(u) = \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|u-x\|_2^2 + f(x) \quad (\mathbb{R}^N \to \mathbb{R}^N)$$

Classical cases: squared ℓ_2 regulizer $f(x) = \frac{\tau}{2}\|x\|_2^2$

$$\mathrm{prox}_f(u) = \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|u-x\|_2^2 + \frac{\tau}{2}\|x\|_2^2 \;=\; \frac{u}{1+\tau}$$

squared ℓ_2 regularizer with "analysis" operator $f(x) = \frac{\tau}{2}\|Dx\|_2^2$

$$\begin{aligned}\mathrm{prox}_f(u) &= \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|u-x\|_2^2 + \frac{\tau}{2}\|Dx\|_2^2 \\ &= (I+\tau D^T D)^{-1}u\end{aligned}$$

if D is a circulant matrix, $O(N\log N)$ cost using the FFT

Proximity operator of the ℓ_1 norm

$$\text{prox}_{\tau\|\cdot\|_1}(u) = \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|u-x\|_2^2 + \tau\|x\|_1$$

Separable: solve w.r.t. each component: $\min_x \tau|x| + 0.5(x-u)^2$

Possible approach: write $|x| = \max_{|z|\leq 1} zx$

$$\min_x \max_{|z|\leq 1} \tau z x + 0.5(x-u)^2 = \max_{|z|\leq 1} \min_x \tau z x + 0.5(x-u)^2$$

$$= \max_{|z|\leq 1} -0.5\tau^2 z^2 + \tau z u \quad \text{(for } x = u - \tau z\text{)}$$

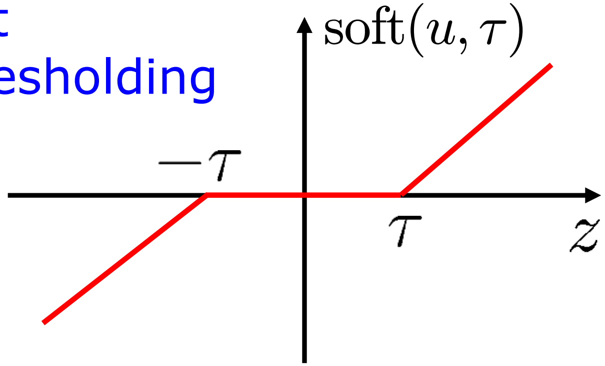
$$\arg\max_{|z|\leq 1} -0.5\tau^2 z^2 + \tau z u = \begin{cases} u/\tau & \Leftarrow \quad |u| \cdot \ \tau \\ 1 & \Leftarrow \quad u > \tau \\ -1 & \Leftarrow \quad u < -\tau \end{cases}$$

$$\arg\min_x \tau|x| + 0.5(x-u)^2 = \begin{cases} 0 & \Leftarrow \quad |u| \leq \tau \\ u-\tau & \Leftarrow \quad u > \tau \\ u+\tau & \Leftarrow \quad u < -\tau \end{cases}$$

Proximity operator of the ℓ_1 norm: the "soft"

$$\arg\min_x \tau|x| + 0.5(x-u)^2 = \begin{cases} 0 & \Leftarrow \quad |u| \leq \tau \\ u-\tau & \Leftarrow \quad u > \tau \\ u+\tau & \Leftarrow \quad u < -\tau \end{cases}$$

soft thresholding



$$= \mathrm{sign}(u)\max\{0, |u|-\tau\}$$

$$\equiv \mathrm{soft}(u,\tau) \;= \mathrm{prox}_{\tau|\cdot|}$$

(for vectors, $\mathrm{soft}(u,\tau)$ is applied component-wise)

p-th power of ℓ_p norms $\quad \|x\|_p^p = \sum_i |[x]_i|^p$

closed form prox for $\quad p \in \left\{1, 2, \frac{4}{3}, \frac{3}{2}, 3, 4\right\}$

[Combettes, Wajs, 2005]

Dual norms, proximity operators, and projections

Dual norm: some norm, $\|\cdot\|: \mathbb{R}^N \to \mathbb{R}_+$

its dual norm: $$\|x\|^* = \max_{\|z\|\leq 1} \langle x, z \rangle$$

Dual norm of $\|\cdot\|_p$ is $\|\cdot\|_q$, where $\dfrac{1}{p} + \dfrac{1}{q} = 1$ Hölder conjugates

... simple corollary of Hölder's inequality: $x^T z \leq \|x\|_p \|z\|_q$

Examples of Hölder conjugates: $(2,2),\ (1,+\infty),\ (3/2,3),...$

These concepts are related through:

$$\text{prox}_{\|\cdot\|}(u) = u - P_{\{x:\|x\|^*\leq 1\}}(u)$$

[Combettes, Wajs, 2005]

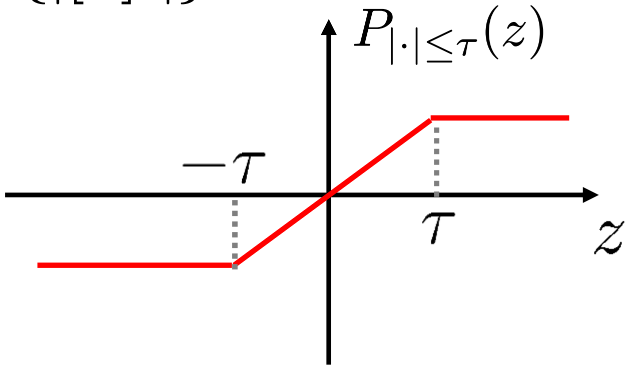
Dual norms, proximity operators, and projections

$$\mathrm{prox}_{\tau\|\cdot\|}(u) = u - P_{\{x:\|x\|^* \leq \tau\}}(u)$$

This relation underlies our earlier derivation of $\mathrm{prox}_{\|\cdot\|_1}$

$$\mathrm{prox}_{\tau\|\cdot\|_1}(u) = u - P_{\{x:\|x\|_\infty \leq \tau\}}(u)$$

$$\|x\|_\infty = \max\{|[x]_i|\}$$

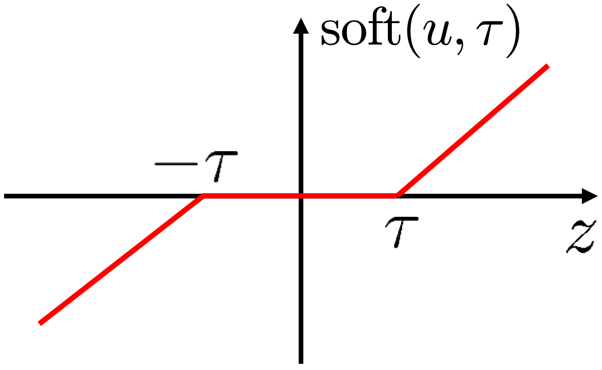


It's all separable,

$$\mathrm{prox}_{\tau|\cdot|}(u) = u - P_{\{x:|x|\leq\tau\}}(u)$$

$$= u - \begin{cases} u & \Leftarrow \quad |u| \leq \tau \\ -\tau & \Leftarrow \quad u < -\tau \\ \tau & \Leftarrow \quad u > \tau \end{cases}$$

$$= \mathrm{soft}(u, \tau)$$



Dual norms, proximity operators, and projections

$$\mathrm{prox}_{\tau\|\cdot\|}(u) = u - P_{\{x:\|x\|^* \leq \tau\}}(u)$$

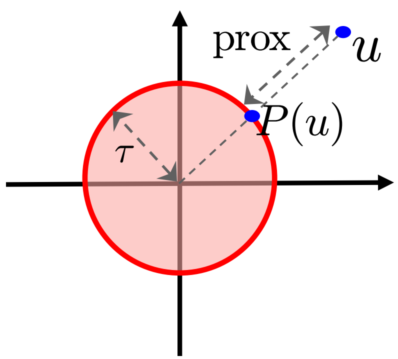
This relation allows deriving $\mathrm{prox}_{\|\cdot\|_\infty}$ and $\mathrm{prox}_{\|\cdot\|_2}$

$$\mathrm{prox}_{\|\cdot\|_\infty}(u) = u - P_{\{x:\|x\|_1 \leq \tau\}}(u)$$

projection on the ℓ_1 ball of radius τ

$O(n \log n)$

$$\mathrm{prox}_{\|\cdot\|_2}(u) = u - P_{\{x:\|x\|_2 \leq \tau\}}(u)$$



$$= u - \begin{cases} u & \Leftarrow \quad \|u\|_2 \leq \tau \\ \tau\, u/\|u\|_2 & \Leftarrow \quad \|u\|_2 > \tau \end{cases}$$

$$= \frac{u}{\|u\|_2} \max\{0, \|u\|_2 - \tau\}$$

vector soft thresholding

Proximity operators of atomic norms

$$\mathrm{prox}_{\tau\|\cdot\|}(u) = u - P_{\{x:\|x\|^*\leq\tau\}}(u)$$

These relation allows deriving prox operators of atomic norms:

$$\|x\|_{\mathcal{A}} = \inf\{t > 0 : x \in t\,\mathrm{conv}(\mathcal{A})\}$$

The dual of an atomic norm ball:

$$\begin{aligned}\|x\|_{\mathcal{A}}^* = \max_{\|z\|_{\mathcal{A}}\leq 1} \langle z, x\rangle &= \max_{z\in\mathrm{conv}(\mathcal{A})} \langle z, x\rangle \\ &= \max\{\langle a, x\rangle,\ a \in \mathcal{A}\}\end{aligned}$$

$$P_{\{x:\|x\|_{\mathcal{A}}^*\leq\tau\}}(u) = \arg\min_{\langle a,x\rangle\leq\tau,\,\forall_{a\in\mathcal{A}}} \|u - x\|_2^2$$

$$\mathrm{prox}_{\tau\|\cdot\|_{\mathcal{A}}}(u) = u - \arg\min_{\langle a,x\rangle\leq\tau,\,\forall_{a\in\mathcal{A}}} \|u - x\|_2^2$$

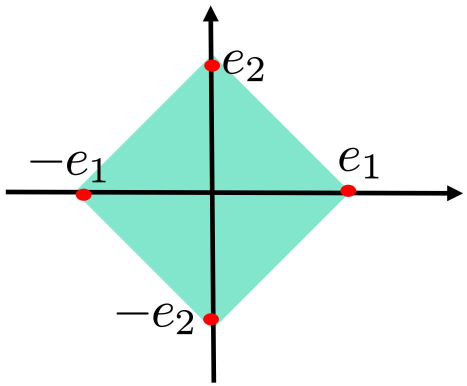
Proximity operators of atomic norms: ℓ_1

Deriving $\mathrm{prox}_{\tau\|\cdot\|_1}$ from the atomic norm view

$$\|x\|_1 = \|x\|_{\mathcal{A}}$$

$$\mathcal{A} = \left\{ \begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix}, \begin{bmatrix}0\\1\\\vdots\\0\end{bmatrix}, \ldots, \begin{bmatrix}0\\0\\\vdots\\1\end{bmatrix}, \begin{bmatrix}-1\\0\\\vdots\\0\end{bmatrix}, \ldots, \begin{bmatrix}0\\0\\\vdots\\-1\end{bmatrix} \right\}$$

$$= \{e_1, e_2, ..., e_N, -e_1, ..., -e_N\}$$



$$|\mathcal{A}| = 2\,N$$

$$\|x\|_{\mathcal{A}}^* = \max\{\langle a, x\rangle,\ a \in \mathcal{A}\} = \max\{|[x]_i|\} = \|x\|_\infty$$

$$\mathrm{prox}_{\tau\|\cdot\|_1}(u) = u - P_{\{x : \|x\|_\infty \le \tau\}}(u)$$

$$= \mathrm{soft}(x, \tau)$$

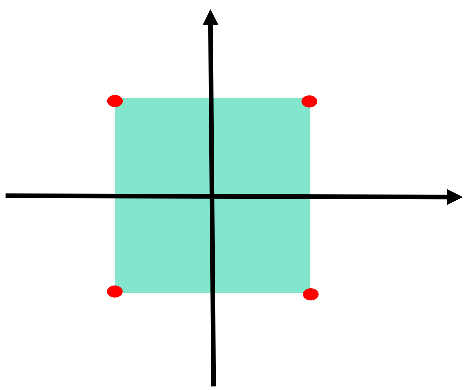
Proximity operators of atomic norms: ℓ_∞

Deriving $\text{prox}_{\tau\|\cdot\|_\infty}$ from the atomic norm view

$$\|x\|_\infty = \|x\|_{\mathcal{A}}$$

$$\mathcal{A} = \left\{ \begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}, \begin{bmatrix}-1\\1\\\vdots\\1\end{bmatrix}, \begin{bmatrix}1\\-1\\\vdots\\1\end{bmatrix}, \begin{bmatrix}-1\\-1\\\vdots\\1\end{bmatrix}, \ldots, \begin{bmatrix}-1\\-1\\\vdots\\-1\end{bmatrix} \right\}$$

$$= \{-1, +1\}^N$$



$$|\mathcal{A}| = 2^N$$

$$\|x\|_{\mathcal{A}}^* = \max\{\langle a, x\rangle,\ a \in \mathcal{A}\} = \sum_{i=1}^N |[x]_i| = \|x\|_1$$

$$\text{prox}_{\tau\|\cdot\|_\infty}(u) = u - P_{\{x:\|x\|_1\leq\tau\}}(u)$$

Proximity of atomic norms: matrix nuclear norm

Matrix nuclear norm: $\|X\|_* = \sum_i \sigma_i(X) = \sum_i \sqrt{\lambda_i(X^T X)}$

$$\|X\|_* = \|X\|_{\mathcal{A}} \qquad \mathcal{A} = \{Z : \mathrm{rank}(Z) = 1, \|Z\|_F = 1\}$$

$$\mathrm{rank}(Z) = |\{\sigma_i(Z) \neq 0\}|$$

Frobenius norm $\|Z\|_F^2 = \sum_{ij} [Z]_{ij}^2 = \sum_i \sigma_i^2(Z)$

$$\|X\|_{\mathcal{A}}^* = \max\{\langle Z, X\rangle,\ Z \in \mathcal{A}\}$$

$$= \max\left\{\sum_i \sigma_i(Z)\sigma_i(X),\ \mathrm{rank}(Z) = 1, \sum_i \sigma_i^2(Z) = 1\right\}$$

$$= \sigma_{\max}(X) \ = \ \|X\|_2$$ spectral norm

Proximity of atomic norms: matrix nuclear norm

Euclidean matrix projection: $P_{\mathcal{S}}(X) = \arg\min_{Z \in \mathcal{S}} \|Z - X\|_F^2$

Note: for any unitary matrix U $(U^T U = I, UU^T = I)$

$$\|UM\|_F^2 = \mathrm{trace}\left(M^T U^T U M\right) = \mathrm{trace}\left(M^T A\right) = \|M\|_F^2$$

$$\mathrm{prox}_{\tau\,\|\cdot\|_*}(X) = X - P_{\{Z : \|Z\|_2 \leq \tau\}}(X)$$

$$= U\Lambda V^T - P_{\{Z : \sigma_{\max}(Z) \leq \tau\}}(U\Lambda V^T)$$

singular value diagonal matrix

[Lewis, Malick, 2009]

$$= U \mathrm{diag}\Big(\mathrm{diag}(\Lambda) - P_{\{x : \|x\|_\infty \leq \tau\}}(\mathrm{diag}(\Lambda))\Big) V^T$$

$$= U \mathrm{soft}(\Lambda, \tau) V^T$$

singular value thresholding (svt)

Proximity of atomic norms: matrix spectral norm

Matrix spectral norm: $$\|X\|_2 = \sigma_{\max}(X)$$

$$\|X\|_2 = \|X\|_{\mathcal{A}} \quad \mathcal{A} = \{Z : Z^T Z = I\} = \{Z : \sigma_i(Z) = 1, \forall_i\}$$

orthogonal matrices

$$\|X\|_{\mathcal{A}}^* = \max\{\langle Z, X\rangle,\ Z \in \mathcal{A}\}$$

$$= \max\left\{\sum_i \sigma_i(Z)\sigma_i(X),\ \sigma_i(Z) = 1, \forall_i\right\}$$

$$= \sum_i \sigma_i(X) \ = \|X\|_*$$ nuclear norm

Proximity of atomic norms: matrix spectral norm

$$\mathrm{prox}_{\tau\,\|\cdot\|_2}(X) = X - P_{\{Z:\|Z\|_* \leq \tau\}}(X)$$

$$= U\Lambda V^T - P_{\{Z:\|Z\|_* \leq \tau\}}(U\Lambda V^T)$$

singular value diagonal matrix

$$= U\left(\Lambda - P_{\{Z:\sum_i \sigma_i(Z) \leq \tau\}}(\Lambda)\right)V^T$$

$$= U\mathrm{diag}\Big(\mathrm{diag}(\Lambda) - P_{\{x:\|x\|_1 \leq \tau\}}(\mathrm{diag}(\Lambda))\Big)V^T$$

residual of projection of the singular values on an ℓ_1 ball of radius τ

Proximity and atomic sets: vectors vs matrices

vectors			matrices		
norm	prox	atomic set	norm	prox	atomic set
ℓ_1 $\|\|x\|\|_1$	component soft thresholding	$\mathcal{A} = \{\pm e_i\}$ $\|\mathcal{A}\| = 2N$	nuclear $\|\|X\|\|_*$	singular value thresholding	$\mathcal{A} =$ set of all rank 1, norm 1 matrices
ℓ_∞ $\|\|x\|\|_\infty$	residual of projection on ℓ_1 ball	$\mathcal{A} = \{\pm 1\}^N$ $\|\mathcal{A}\| = 2^N$	spectral $\|\|X\|\|_2$	residual of s.v. proj. on ℓ_1 ball	$\mathcal{A} =$ set of all orthogonal matrices
ℓ_2 $\|\|x\|\|_2$	vector soft thresholding	$\mathcal{A} =$ set of all vectors with norm 1 $\|\mathcal{A}\| = \infty$	Frobenius $\|\|X\|\|_F$	matrix soft threshold.	$\mathcal{A} =$ all matrices of unit Frobenius norm.

Proximal algorithms

Back to the problem: $\widehat{x} \quad \in \quad \arg\min_{x\in\mathbb{R}^n} f_1(x) + f_2(x)$

with f_2 a proper convex function

and f_1 has a L -Lipschitz gradient; e.g. $f_1(x) = \frac{1}{2}\|\Phi x - u\|_2^2$

with $L = \lambda_{\max}(\Phi^*\Phi)$

separable majorizer ($\beta_k < 1/L$)

$$Q(x, x_k) = f_1(x_k) + (x - x_k)^T \nabla f_1(x_k) + \frac{1}{2\beta_k}\|x - x_k\|_2^2$$

majorization-minimization algorithm

$$\begin{aligned} x_{k+1} &= \arg\min_x Q(x, x_k) + f_2(x) \\ &= \arg\min_x \ \frac{1}{2\beta_k}\left\|x - \left(x_k - \beta_k \nabla f_1(x_k)\right)\right\|_2^2 + f_2(x) \end{aligned}$$

$$x_{k+1} = \operatorname{prox}_{\beta_k\, f_2}\Big(x_k - \beta_k \nabla f_1(x_k)\Big)$$

Proximal algorithms: convergence

$f(x)$

Problem: $\widehat{x} \in \arg\min_{x \in \mathbb{R}^n} f_1(x) + f_2(x)$

f_1 has a L-Lipschitz gradient; e.g. $f_1(x) = \frac{1}{2}\|\Phi x - u\|_2^2$

$L = \lambda_{\max}(\Phi^* \Phi)$

Iterative shrinkage/thresholding (IST)
(or forward-backward)

$$x_{k+1} = \text{prox}_{\beta_k f_2}\Big(x_k - \beta_k \nabla f_1(x_k)\Big)$$

if $\beta_k < \frac{1}{L}$, IST is a majorization-minimization algorithm, thus

$$f(x_{k+1}) \leq f(x_k)$$

$f(x) \geq 0$, thus $(f(x_1), f(x_2), ..., f(x_k), ...)$ converges.

Attention: this does **not** imply convergence of $(x_1, ..., x_k, ...)$

Proximal algorithms: convergence

$$\widehat{x} \in G = \arg\min_{x \in \mathbb{R}^n} f_1(x) + f_2(x)$$

IST algorithm: $x_{k+1} = \text{prox}_{\beta_k f_2}\Big(x_k - \beta_k \nabla f_1(x_k)\Big)$

if $0 < \beta_k < \frac{2}{L}$, then $(x_1, x_2, \ldots, x_k, \ldots)$ converges to a point in G

Inexact version:

errors

$$x_{k+1} = \text{prox}_{\beta_k f_2}\Big(x_k - (\beta_k \nabla f_1(x_k) + b_k)\Big) + a_k$$

convergence still guaranteed if

$$\sum_{k=1}^{\infty} \|a_k\| < \infty \qquad \sum_{k=1}^{\infty} \|b_k\| < \infty$$

Results and proofs in [Combettes and Wajs, 2005]

Proximal algorithms: convergence

Convergence of function values $\big(f(x_1),\, \ldots,\, f(x_k),\, \ldots\big) \to f(\widehat{x})$

$\not\Downarrow \qquad \Uparrow$

Convergence of iterates $\big(x_1,\, x_2,\, \ldots,\, x_k,\, \ldots\big) \to \widehat{x}$

Convergence rates (for function values) [Beck, Teboulle, 2009]:

$$f(x_k) - f(\widehat{x}) \leq \frac{L\|x_0 - \widehat{x}\|_2^2}{2\,k}$$

Convergence rate for the iterates require further assumptions on f

Proximal algorithms: convergence of iterates

$$\widehat{x} = \arg\min_x \frac{1}{2}\|\Phi x - u\|_2^2 + f_2(x)$$

With $L = \lambda_{\max}(\Phi^*\Phi)$ $\quad l = \lambda_{\min}(\Phi^*\Phi) > 0 \quad \Rightarrow G = \{\widehat{x}\}$

(unique minimizer)

$\kappa = l/L$ (condition number)

Under- ($\gamma < 1$) or over-relaxed ($\gamma > 1$) IST

$$x_{k+1} = (1-\gamma)x_k + \gamma\, \mathrm{prox}_{f_2}\Big(x_k - \beta\Phi^T(\Phi x_k - u)\Big)$$

Optimal choice $\gamma = \dfrac{2}{L+l}$ $\qquad \rho = \dfrac{1-\kappa}{1+\kappa}$

Q-linear convergence $\quad \|x_{k+1} - \widehat{x}\| \leq \rho \|x_k - \widehat{x}\|$

Small $l \Rightarrow \rho \lesssim 1 \Rightarrow$ slow convergence!

[F, Bioucas-Dias, 2007]

Proximal algorithms: convergence of iterates

$$\widehat{x} \in G = \arg\min_x \frac{1}{2}\|\Phi x - u\|_2^2 + \tau\|x\|_1$$

With $L = \lambda_{\max}(\Phi^*\Phi)$; using a step-size $\beta < 2/L$,

$$x_{k+1} = \text{soft}\Big(x_k - \beta\,\Phi^T(\Phi x_k - u), \beta\tau\Big)$$

$\mathcal{Z} \subseteq \{1, 2, ..., n\}$ such that $\widehat{x} \in G \Rightarrow [\widehat{x}]_{\mathcal{Z}} = 0$

Then, after a finite number of iterations: $[x_k]_{\mathcal{Z}} = [\widehat{x}]_{\mathcal{Z}} = 0$

After this, Q-linear convergence: $l = \lambda_{\min}(\Phi^*_{\bar{\mathcal{Z}}}\Phi_{\bar{\mathcal{Z}}}) > 0$

Optimal choice $\beta = \dfrac{2}{L + l}$, $\rho = \dfrac{1-\kappa}{1+\kappa}$

$$\|x_{k+1} - \widehat{x}\| \leq \rho\|x_k - \widehat{x}\|$$

[Hale, Yin, Zhang, 2008]

Slowness and acceleration of IST

Problem: $$\widehat{x} \in G = \arg\min_x \frac{1}{2}\|\Phi x - u\|_2^2 + \tau \|x\|_1$$

IST algorithm: $$x_{k+1} = \text{soft}\Big(x_k - \beta\, \Phi^T(\Phi x_k - u), \beta\tau\Big)$$

IST is **slow**, if Φ is very ill-conditioned and/or τ is very small!

Several proposals for accelerated variants of IST

Methods with memory (TwIST, FISTA)

Quasi-Newton methods (SpaRSA)

Continuation, i.e., use a varying τ (FPC, SpaRSA)

Memory-based variants of IST: FISTA

Fast IST algortihm (FISTA); based on Nesterov's work (1980's)

[Beck, Teboulle, 2009]

FISTA

$$t_{k+1} = \frac{1+\sqrt{1+4\,t_k^2}}{2}$$

$$z_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}\,(x_k - x_{k-1})$$

$$x_{k+1} = \mathrm{soft}\Big(z_k - \beta\,\Phi^T(\Phi z_k - u), \beta\tau\Big)$$

IST: $f(x_k) - f(\widehat{x}) = O\left(\frac{1}{k}\right) \qquad \left(\leq \frac{L\|x_0 - \widehat{x}\|_2^2}{2\,k}\right)$

FISTA: $f(x_k) - f(\widehat{x}) = O\left(\frac{1}{k^2}\right)$

Memory-based variants of IST: twist

Inspired by 2-step methods for linear systems

[Frankel, 1950], [Axelsson, 1996]

TwIST (two-step IST):

[Bioucas-Dias, F, 2007]

$$x_{k+1} = (\alpha - \beta) x_k + (1-\alpha) x_{k-1} + \beta \, \mathrm{prox}_{f_2}\left(x_k - \Phi^T (\Phi x_k - u)\right)$$

$$\kappa = \frac{\lambda_{\min}(\Phi^T \Phi)}{\lambda_{\max}(\Phi^T \Phi)}$$

$$\rho = \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}$$ TwIST

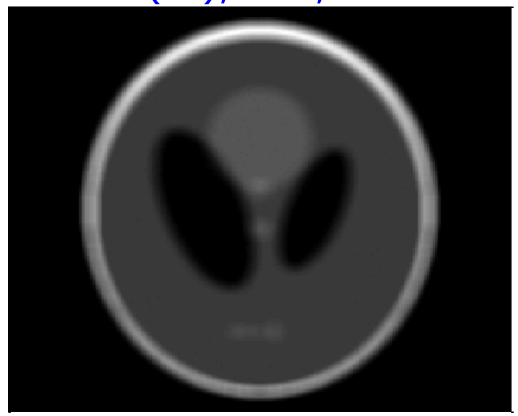
Q-linear convergence $\|x_{k+1} - \widehat{x}\| \leq \rho \|x_k - \widehat{x}\|$

$$\rho = \frac{1-\kappa}{1+\kappa}$$ IST

Memory-based variants of IST: twist

original

Blurred (B), 9x9, 40db noise

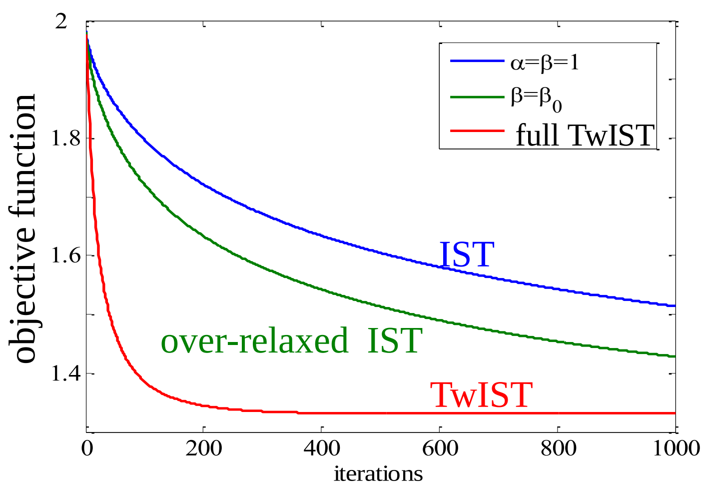


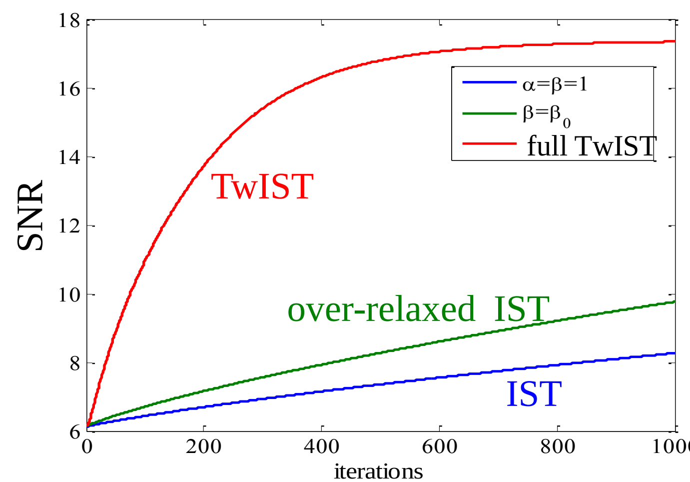
restored

$$\widehat{x} \quad \in \quad \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|B\Psi x - u\|_2^2 + \tau\|x\|_1$$

representation coefficients

dictionary (e.g, wavelet basis, frame, ...)





Quasi-newton acceleration of IST: SpaRSA

IST: $x_{k+1} = \text{prox}_{\beta_k f_2}\Big(x_k - \beta_k \nabla f_1(x_k)\Big)$

A Newton step (instead of gradient descent) would be:

$$x_{k+1} = \text{prox}_{\beta_k f_2}\Big(x_k - [H(x_k)]^{-1} \nabla f_1(x_k)\Big)$$

Hessian
(matrix of second derivatives)

...computationally too expensive!

Barzilai-Borwein approach:

[Barzilai-Borwein, 1988], [Wright, Nowak, F, 2009]

$$\frac{1}{\beta_k} I \simeq H(x_k)$$

$$\frac{1}{\beta_k} = \arg\min_{\alpha} \|\alpha(x_k - x_{k-1}) - (\nabla f(x_k) - \nabla f(x_{k-1})\|_2^2$$

If $f_1(x) = \frac{1}{2}\|\Phi x - u\|_2^2$, then $\beta_k = \dfrac{\|x_k - x_{k-1}\|_2^2}{\|\Phi(x_k - x_{k-1})\|_2^2}$

Acceleration via continuation

IST: $$x_{k+1} = \mathrm{soft}\Big(x_k - \beta\,\Phi^T(\Phi x_k - u), \beta\tau\Big)$$

Slow, if τ is small.

Observation: IST (as SpaRSA) benefits from

"warm-starting" (being initialized *close* to the minimizer)

Continuation: start with large τ

slowly decrease τ while tracking the solution.

[F, Nowak, Wright, 2007], [Hale, Yin, Zhang, 2007]

IST + continuation = fixed point continuation (FPC)

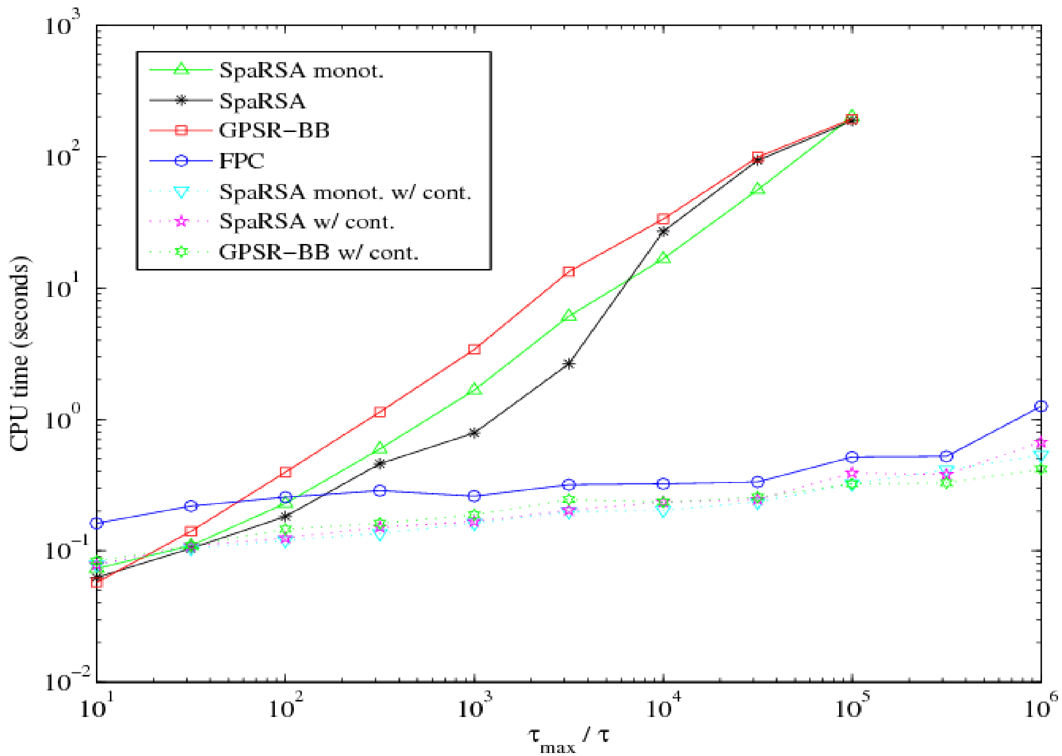
[Hale, Yin, Zhang, 2007]

Acceleration via continuation

$$\widehat{x} \in G = \arg\min_x \frac{1}{2}\|\Phi x - u\|_2^2 + \tau\|x\|_1$$

1024×4096

$$u = \Phi x^* + n$$



$$\tau_{\max} = \|\Phi^T \mathbf{y}\|_\infty \quad (\ \tau \geq \tau_{\max} \ \Rightarrow \ \widehat{x} = 0\ )$$

Some speed comparisons

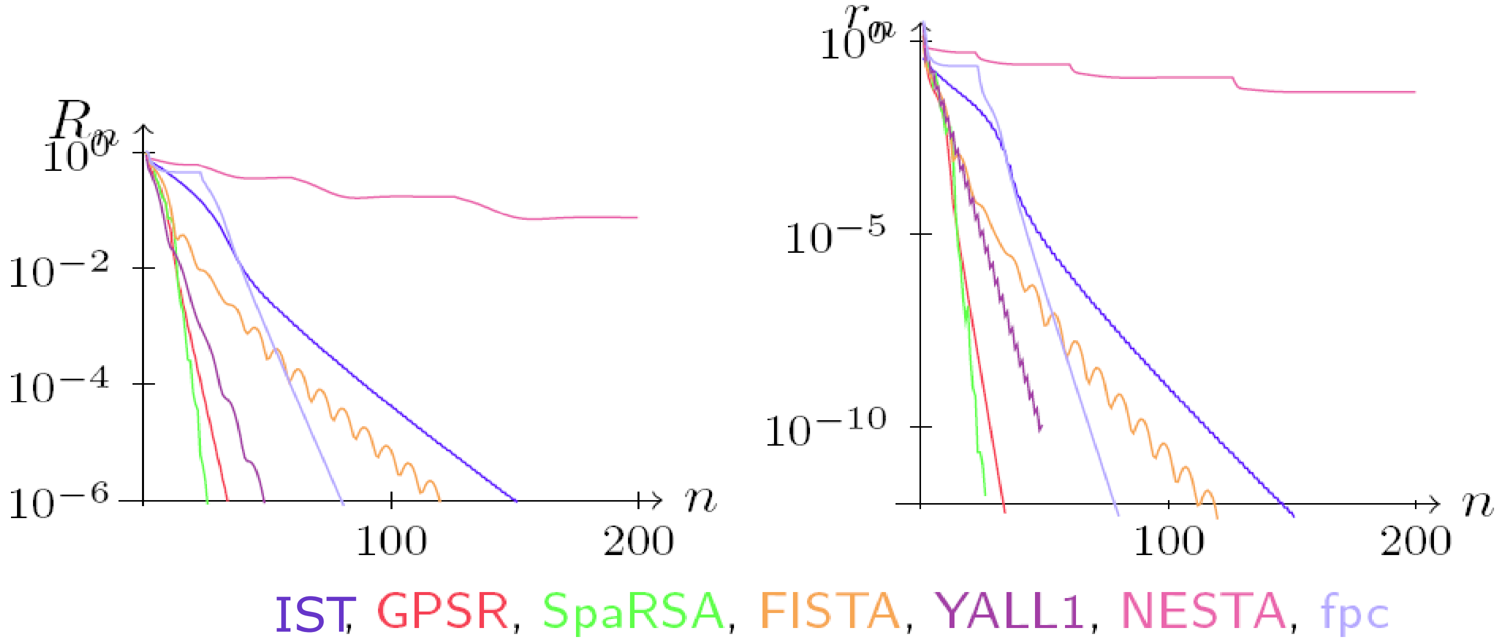
from [Lorenz, 2011]

$$\widehat{x} = \arg\min_x \frac{1}{2}\|\Phi x - u\|_2^2 + \tau\|x\|_1$$

$\Phi = [I\ U\ R]$ $\quad \tau = 0.1$

(512×1536) $\quad \widehat{x}$ with 120 non-zeros

$$R_n = \frac{\|\mathbf{x}^n - \widehat{\mathbf{x}}\|_2}{\|\widehat{\mathbf{x}}\|_2} \qquad r_n = \frac{\phi(\mathbf{x}^n) - \phi(\widehat{\mathbf{x}})}{\phi(\widehat{\mathbf{x}})}$$



Proximal algorithms for matrices

$$\widehat{M} \in \arg\min_{M\in\mathbb{R}^{n\times n}} \frac{1}{2}\|\Phi(M) - U\|_F^2 + \mu\|M\|_*$$

linear operator

The proximal algorithm (IST) is as before:

...its adjoint

$$X_{k+1} = \mathrm{svt}_{\mu\,\beta_k}\Big(X_k - \beta_k\,\Phi^*(\Phi(X_k) - U)\Big)$$

Matrix completion: $\Phi(X) = X$ (subset of entries) $|\Omega| = p$

Unknown M				IST			APG (FISTA)		
n/r	p	p/d_r	μ	iter	#sv	error	iter	#sv	error
100/10	5666	3	8.21e-03	7723	61	1.88e-01	655	13	1.06e-03
200/10	15665	4	1.05e-02	12180	96	2.45e-01	812	12	1.02e-03
500/10	49471	5	1.21e-02	10900	203	5.91e-01	1132	16	7.63e-04

Unknown M				FPC (continuation)			APG + continuation		
n/r	p	p/d_r	μ	iter	#sv	error	iter	#sv	error
100/10	5666	3	8.21e-03	429	32	1.06e-03	74	10	1.46e-04
200/10	15665	4	1.05e-02	278	49	4.38e-04	73	10	1.02e-04
500/10	49471	5	1.21e-02	484	125	5.50e-04	72	10	8.06e-05

from [Toh, Yun, 2009]

...the importance of acceleration!

Another class of methods: augmented Lagrangian

The problem:
$$\min_x \quad f(x)$$
$$\text{s.t.} \quad \Phi x = u$$

Penalty parameter

The augmented Lagrangian (AL)

$$L_\mu(x,\lambda) = f(x) + \lambda^T(\Phi x - u) + \frac{\mu}{2}\|\Phi x - u\|_2^2$$

The "AL method" (ALM)
(a.k.a. method of multipliers)
[Hestenes, Powell, 1969]

$$x_{k+1} = \arg\min_x L_\mu(x,\lambda_k)$$
$$\lambda_{k+1} = \lambda_k + \mu(\Phi\, x_{k+1} - u)$$

Can be written as:

$$x_{k+1} = \arg\min_x f(x) + \frac{\mu}{2}\|\Phi\, x - u - d_k\|_2^2$$
$$d_{k+1} = d_k - (\Phi\, x_{k+1} - u)$$

Similar to Bregman method [Osher, Burger, Goldfarb, Xu, Yin, 2005]
[Yin, Osher, Goldfarb, Darbon, 2008]

Augmented Lagrangian for variable splitting

The problem: $$\min_x f_1(\Phi x) + f_2(x)$$

Equivalent constrained formulation
$$\min_x \quad f_1(z) + f_2(x)$$
$$\text{s.t.} \quad \Phi x - z = 0$$

Can be written as
$$\min_y \quad f(y)$$
$$\text{s.t.} \quad \Psi y = 0$$

with $y = \begin{bmatrix} x \\ z \end{bmatrix}$

$\Psi = [\Phi \;\; -I]$

ALM:

$$(x_{k+1}, z_{k+1}) = \arg\min_{x,z} f_1(z) + f_2(x) + \frac{\mu}{2}\|\Phi\, x - z - d_k\|_2^2$$
$$d_{k+1} = d_k - (\Phi x_{k+1} - z_{k+1})$$

Augmented Lagrangian for variable splitting

It may be hard to solve

$$(x_{k+1}, z_{k+1}) = \arg\min_{x,z} f_1(z) + f_2(x) + \frac{\mu}{2}\|\Phi\, x - z - d_k\|_2^2$$

Alternative:

$$x_{k+1} = \arg\min_{x} f_2(x) + \frac{\mu}{2}\|\Phi\, x - z_k - d_k\|_2^2$$

$$z_{k+1} = \arg\min_{z} f_1(z) + \frac{\mu}{2}\|\Phi\, x_{k+1} - z - d_k\|_2^2$$

$$d_{k+1} = d_k - \left(\Phi x_{k+1} - z_{k+1}\right)$$

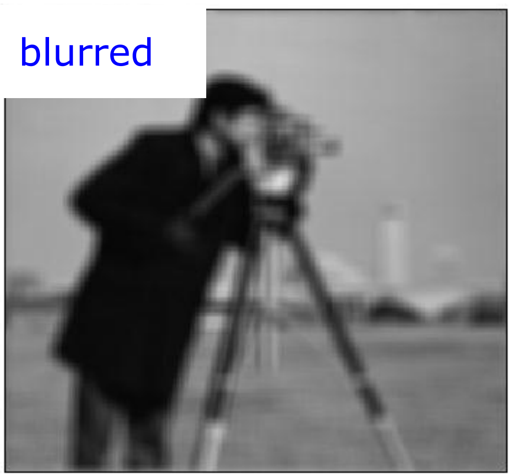
Alternating directions method of multipliers (ADMM)
[Glowinsky, Marrocco, 1975], [Gabay, Mercier, 1976], [Eckstein, Bertsekas, 1992]

When applied to $\widehat{x} = \arg\min_{x} \frac{1}{2}\|\Phi x - u\|_2^2 + \tau\|x\|_1$

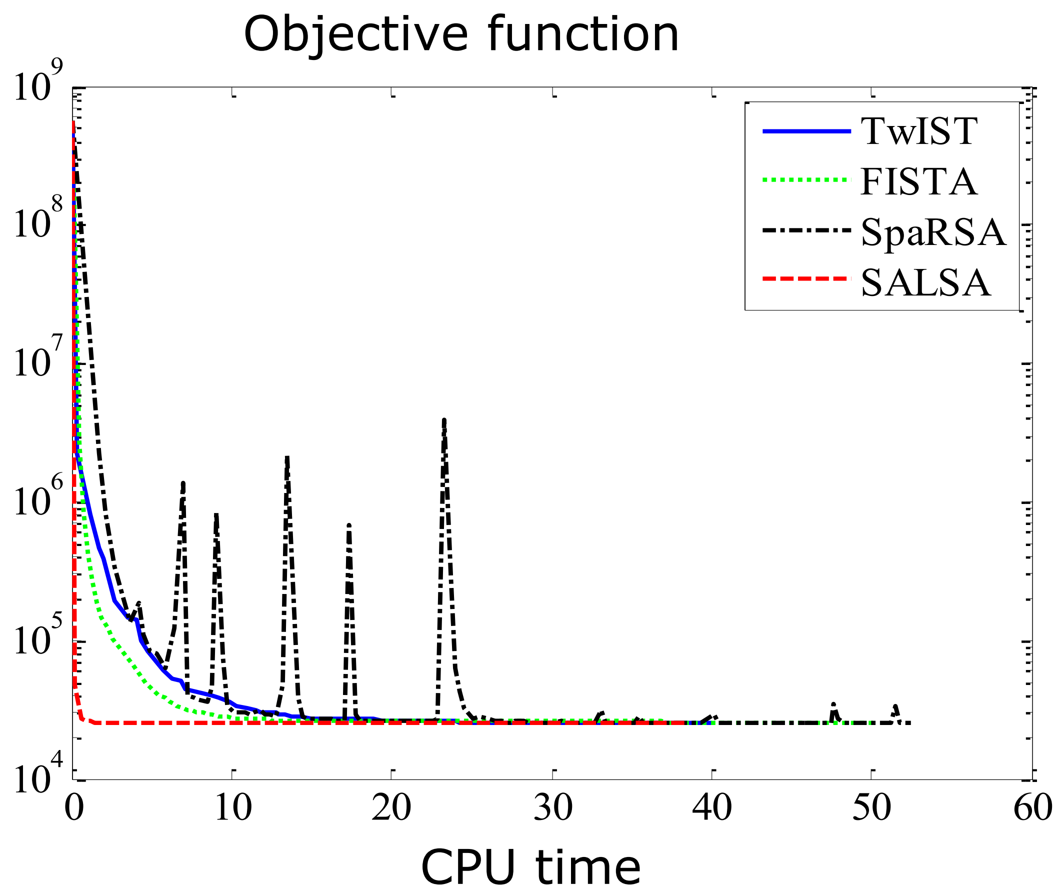
split augmented Lagrangian shrinkage algorithm (SALSA)
[F, Bioucas-Dias, Afonso, 2009]

Augmented Lagrangian for variable splitting

Testing ADMM/SALSA on a typical image deblurring problem



$$\widehat{x} \in \arg\min_{x\in\mathbb{R}^n} \frac{1}{2}\|B\Psi x - u\|_2^2 + \tau\|x\|_1$$



Handling more than two functions

$$\widehat{x} \in \arg\min_{x \in \mathbb{R}^n} f_0(x) + f_1(x) + \cdots + f_n(x)$$

f_0 has a L-Lipschitz gradient

$f_1, \ldots, f_n$ are convex

Possible uses: multiple regularizers, positivity constraints, ...

Generalized forward-backward algorithm [Raguet, Fadili, Peyré, 2011]

Parameters: $\omega_1, \ldots, \omega_n \in (0,1), \text{ s.t. } \sum_j \omega_j = 1$

Initialization: $k = 0,\ z_0^1, \ldots, z_0^n,\ x_0 = \sum_{j=1}^n \omega_j\, z_0^j$

repeat until convergence

for $i = 1 : n$

$$z_{k+1}^i = z_k^i + \mathrm{prox}_{\beta_k f_i/\omega_i}\Big(2\,x_k - z_k^i - \beta_k \nabla f_1(x_k)\Big) - x_k$$

$$x_{k+1} = \sum_{i=1}^n \omega_i\, z_{k+1}^i$$

$$k \leftarrow k+1$$

Handling more than two functions

$$\widehat{x} \in \arg\min_{x \in \mathbb{R}^n} f_1(x) + \cdots + f_n(x)$$

$f_1, ..., f_n$ arbitrary convex functions

ADMM-based method [F and Bioucas-Dias, 2009], [Setzer, Steidl, Teuber, 2009]

Parameter: γ

Initialization: $k = 0,\ z_0^1, ..., z_0^n, y_0^1, ..., y_0^n$

repeat until convergence

$x_{k+1} = (1/n) \sum_{i=1}^{n} (y_k^i - z_k^i)$

for $i = 1 : n$

$y_{k+1}^i = \text{prox}_{\gamma f_i}(x_k - z_k^i)$

$z_{k+1}^i = z_k^i + x_k - y_{k+1}^i$

$k \leftarrow k + 1$

Non-Convex Algorithms for Low-Dimensional Models

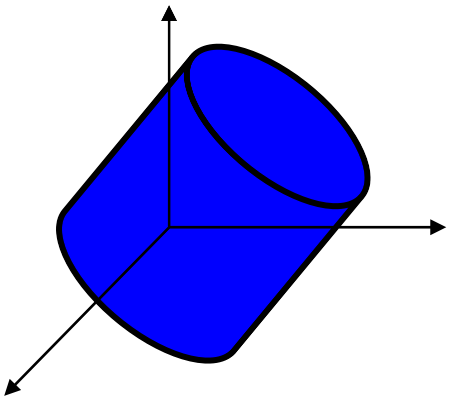
Motivation

Discrete descriptions of low-dimensional models

$$x = \sum_{i=1}^{|\mathcal{A}|} a_i c_i \qquad a_i \in \mathcal{A}, \|c_i\|_0 \le K$$

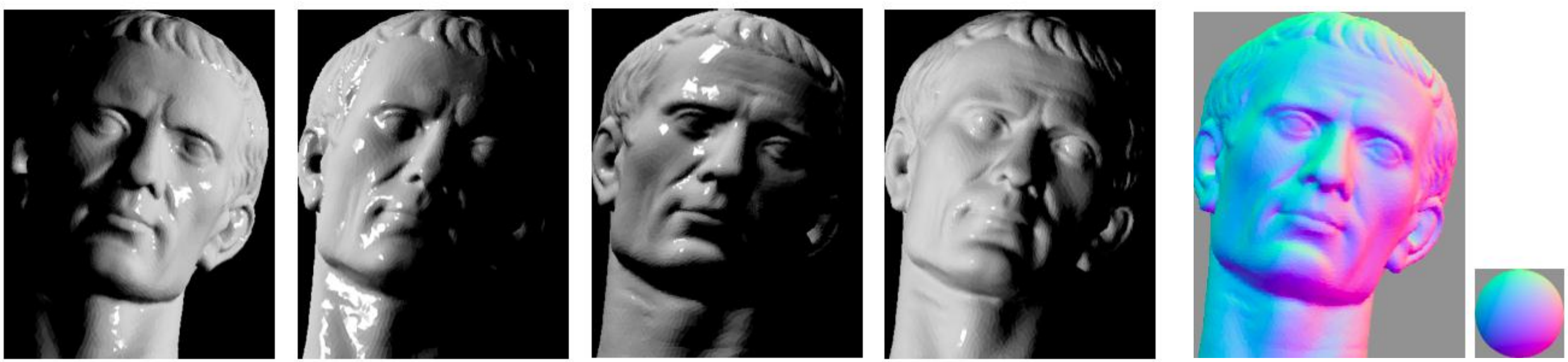
a_i: atoms

$\mathcal{A}$: atomic set



$$\mathcal{A} = \{A : \mathrm{rank}(A) = 1, \|A\|_F = 1\}$$

Example: reflectivity of Lambertian surfaces [Basri and Jacobs 2001]



$$K \le 9$$

$$\text{Intensity} = \rho \max\{\langle n, l \rangle, 0\}$$

Motivation

Discrete descriptions of low-dimensional models

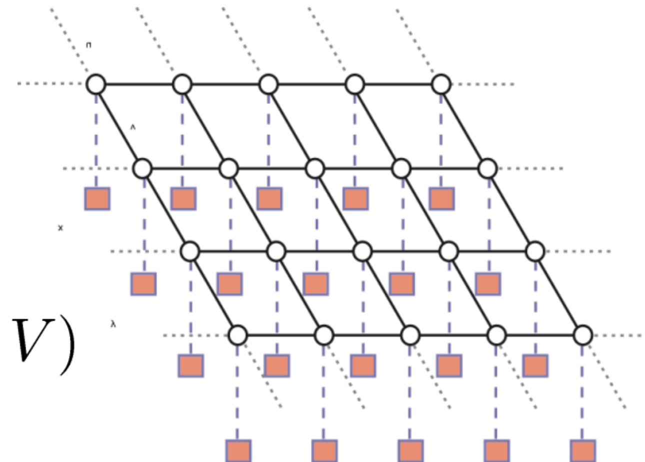
$$x = \sum_{i=1}^{|\mathcal{A}|} a_i c_i \qquad a_i \in \mathcal{A},\ \|c_i\|_0 \leq K$$

a_i: atoms

$\mathcal{A}$: atomic set

$$\mathcal{A} = \{\pm e_i\}_{i=1}^{N}$$

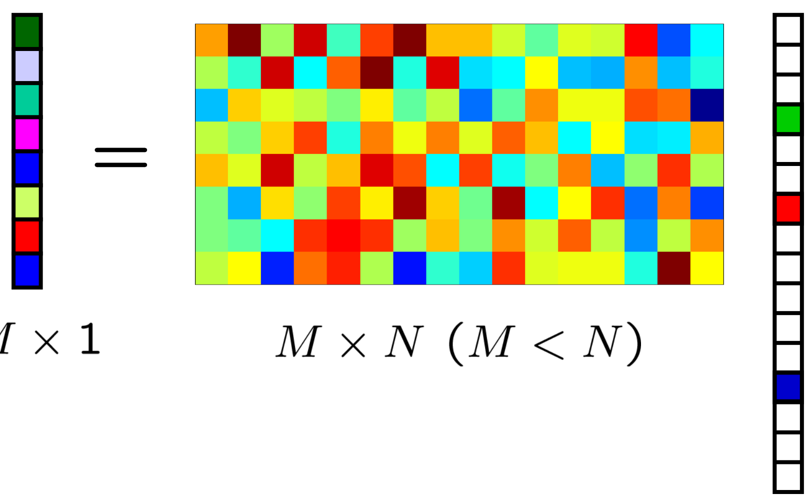
$$\mathcal{G} = (E, V)$$



Example: graphical model selection

vertex weights of the i-th edge

$$u = E_i \qquad \Phi = E_{\setminus i} \qquad \alpha = V_i$$



$M \times 1$ $\quad$ $M \times N \ (M < N)$ $\quad$ $N \times 1$

Gauss-Markov graph
<>
linear regression

$$K \leq \text{node degree of } \mathcal{G}$$

[Meinshausen and Buhlman 2006]

Motivation

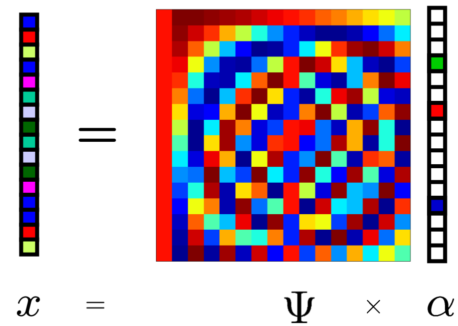
Discrete descriptions of ***structure*** in low-dimensional models

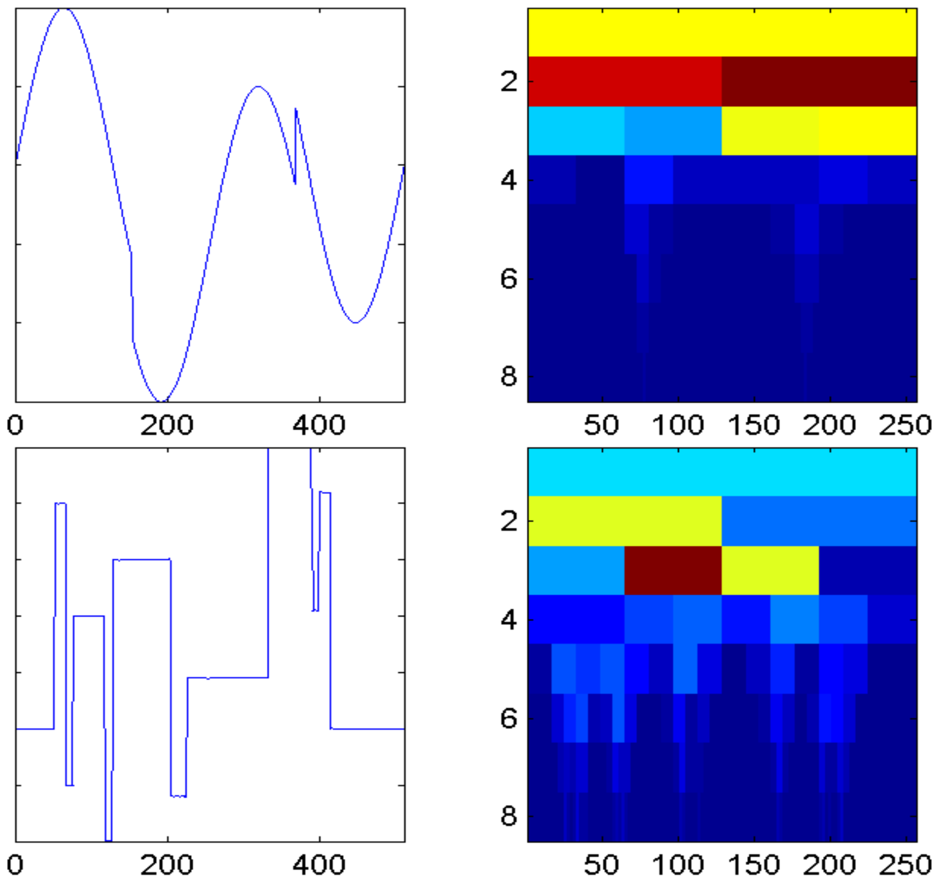
$$x = \sum_{i=1}^{|\mathcal{A}|} a_i c_i \qquad a_i \in \mathcal{A}, \|c_i\|_0 \leq K$$

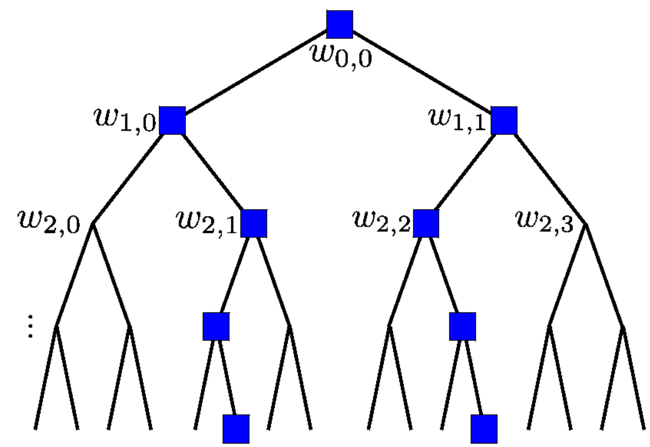
a_i: atoms

$\mathcal{A}$: atomic set

+







Typical of wavelet transforms of natural signals and images (piecewise smooth)

[Baraniuk, C, Duarte, Hegde 2010]

Motivation

Non-convex criteria beyond atomic norms

- 1-bit compressive sensing

$$u = \text{sign}\left(\Phi x\right)$$

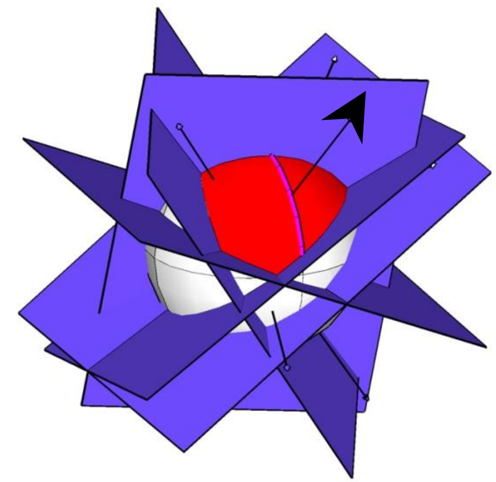
  - optimization criteria $\arg \min_{x:\|x\|_0 \leq K} f(x)$

$$f(x) = -\langle u, \text{sign}(\Phi x)\rangle$$

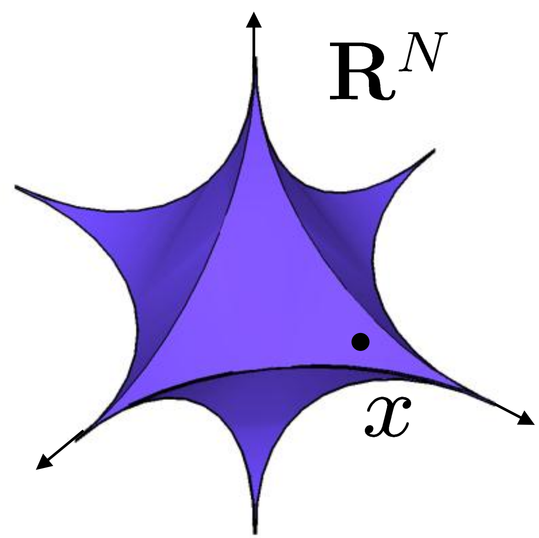
- Compressible signals in weak $\ell_q, q < 1$

$$|x|_{(i)} \leq R i^{-1/q}$$

  - optimization criteria $\arg \min_{x:u=\Phi x} \|x\|_q$



[Boufounos and Baraniuk 2008]



[Chartrand and Yin 2008]

Non-convexity in this tutorial

- Anything ***not*** convex <> too big to cover

  convexity is in general a rare condition

- Active research topic with great depth

  [Attouch et al. 2010]

  Key lesson:
  <u>convergence of the projected gradient-descent algorithm</u>

- This tutorial <> *a special subset*

$$\widehat{x} \in \arg\min_{x \in \mathbb{R}^N} f_1(x) + f_2(x)$$

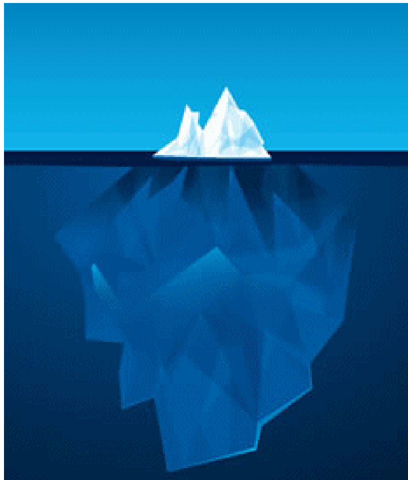
($\mathcal{S}$ is non-convex)

with

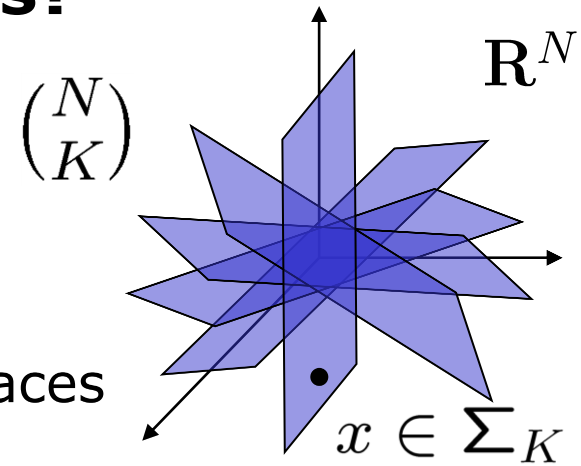
$$f_2(x) = \begin{cases} g(x) & \Leftarrow \quad x \in \mathcal{S} \\ +\infty & \Leftarrow \quad x \notin \mathcal{S} \end{cases}$$

Assumptions:

1. ***access to the gradient of convex*** f_1
2. ***tractable/approximate prox of non-convex*** f_2



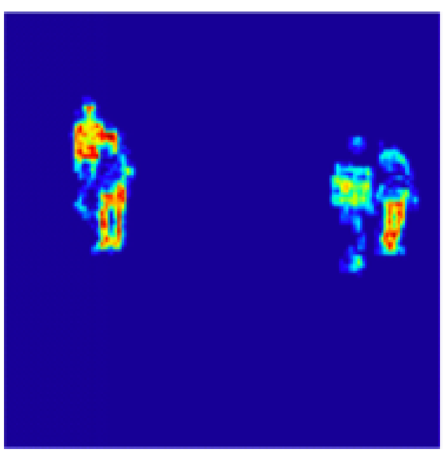
Can we project onto non-convex sets?

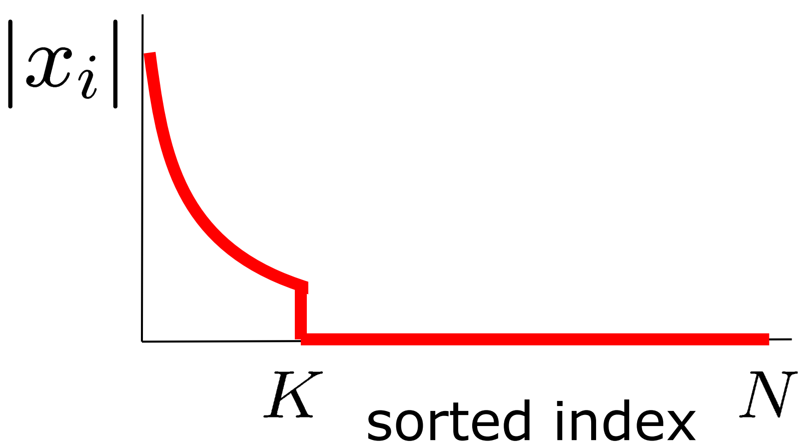


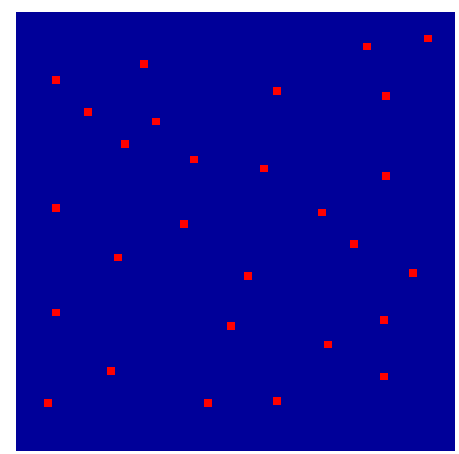
Running examples

- Sparse signal: only K out of N coordinates nonzero
  - model: union of all K-dimensional subspaces aligned w/ coordinate axes

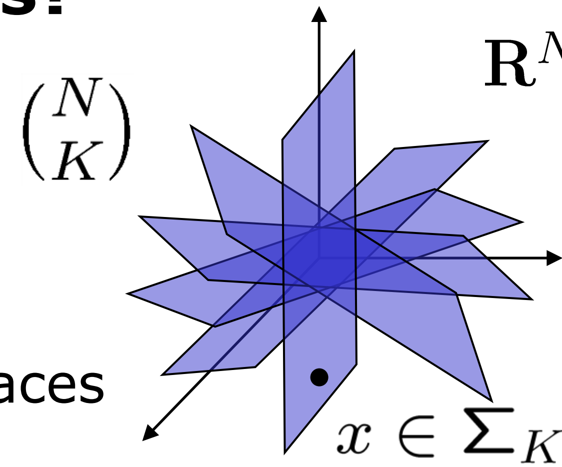
- **Structured** sparse signal: reduced set of subspaces (or model-sparse)
  - model: a particular union of subspaces ex: clustered or dispersed sparse patterns







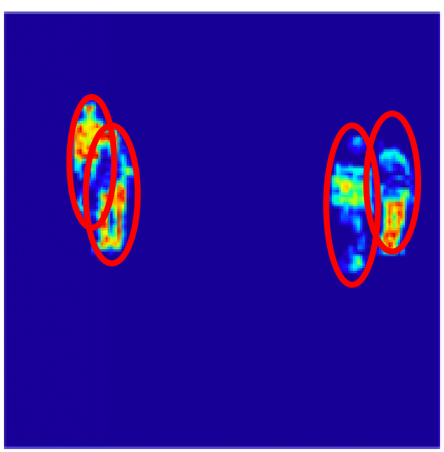
Can we project onto non-convex sets?

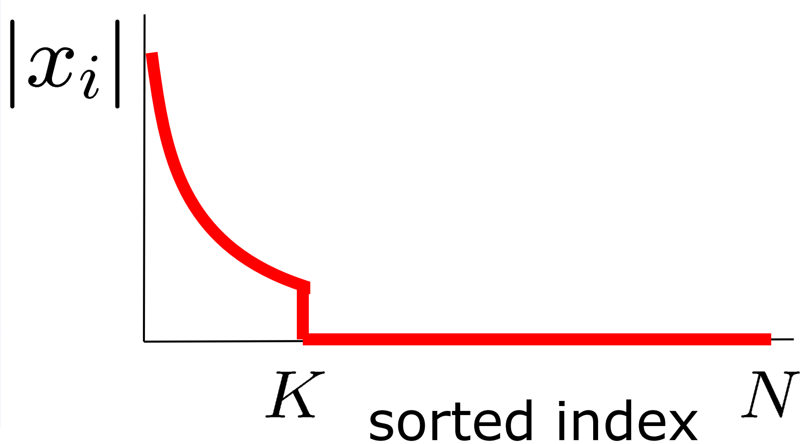


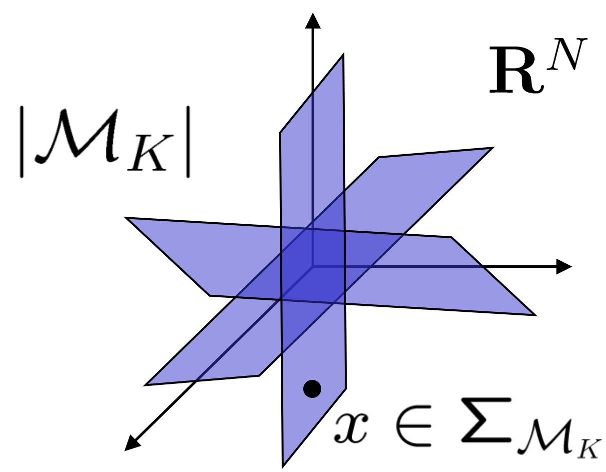
Running examples

- Sparse signal: only K out of N coordinates nonzero
  - model: union of all K-dimensional subspaces aligned w/ coordinate axes

- **Structured** sparse signal: reduced set of subspaces (or model-sparse)
  - model: a particular union of subspaces ex: clustered or dispersed sparse patterns







Can we project onto non-convex sets?

$$\widehat{x} = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|y - x\|_2^2 + f_2(x) \equiv \text{prox}_{f_2}(y)$$

- Analysis of the prox for *structured* sparse sets $g(x) = 0$

$$\text{prox}_{f_2}(y) = \arg\min_{x : x \in \Sigma_{\mathcal{M}_K}} \|x - y\|$$

support of the solution <> modular approximation problem

$$\text{supp}\left(\arg\min_{x:\text{supp}(x) \in \mathcal{M}_K} \|x - y\|_2^2\right) = \arg\min_{\mathcal{S}:\mathcal{S} \in \bar{\mathcal{M}}_K} \|(y)_{\mathcal{S}} - y\|_2^2$$

indexing set

Can we project onto non-convex sets?

$$\widehat{x} = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|y - x\|_2^2 + f_2(x) \equiv \operatorname{prox}_{f_2}(y)$$

- Analysis of the prox for *structured* sparse sets $g(x) = 0$

$$\operatorname{prox}_{f_2}(y) = \arg\min_{x : x \in \Sigma_{\mathcal{M}_K}} \|x - y\|$$

support of the solution <> modular approximation problem

$$\operatorname{supp}\left(\arg\min_{x:\operatorname{supp}(x)\in\mathcal{M}_K} \|x - y\|_2^2\right) = \arg\min_{\mathcal{S}:\mathcal{S}\in\bar{\mathcal{M}}_K} \|(y)_{\mathcal{S}} - y\|_2^2$$

$$= \arg\max_{\mathcal{S}:\mathcal{S}\in\bar{\mathcal{M}}_K} \left\{\|y\|^2 - \|(y)_{\mathcal{S}} - y\|_2^2\right\}$$

Can we project onto non-convex sets?

$$\widehat{x} = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|y - x\|_2^2 + f_2(x) \equiv \mathrm{prox}_{f_2}(y)$$

- Analysis of the prox for *structured* sparse sets $g(x) = 0$

$$\mathrm{prox}_{f_2}(y) = \arg\min_{x : x \in \Sigma_{\mathcal{M}_K}} \|x - y\|$$

support of the solution <> modular approximation problem

$$\begin{aligned}\mathrm{supp}\left(\arg\min_{x:\mathrm{supp}(x)\in\mathcal{M}_K} \|x - y\|_2^2\right) &= \arg\min_{\mathcal{S}:\mathcal{S}\in\bar{\mathcal{M}}_K} \|(y)_{\mathcal{S}} - y\|_2^2 \\ &= \arg\max_{\mathcal{S}:\mathcal{S}\in\bar{\mathcal{M}}_K} \left\{\|y\|^2 - \|(y)_{\mathcal{S}} - y\|_2^2\right\} \\ &= \arg\max_{\mathcal{S}:\mathcal{S}\in\bar{\mathcal{M}}_K} \|(y)_{\mathcal{S}}\|^2\end{aligned}$$

Can we project onto non-convex sets?

$$\widehat{x} = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|y - x\|_2^2 + f_2(x) \equiv \text{prox}_{f_2}(y)$$

- Analysis of the prox for *structured* sparse sets $g(x) = 0$

$$\text{prox}_{f_2}(y) = \arg\min_{x : x \in \Sigma_{\mathcal{M}_K}} \|x - y\|$$

support of the solution <> modular approximation problem

$$\text{supp}\left(\arg\min_{x:\text{supp}(x) \in \mathcal{M}_K} \|x - y\|_2^2\right) = \arg\max_{\mathcal{S}:\mathcal{S} \in \bar{\mathcal{M}}_K} F(S; y)$$

where $F(S; y) = \sum_{i \in \mathcal{S}} |y_i|^2$.

Can we project onto non-convex sets?

$$\widehat{x} = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|y - x\|_2^2 + f_2(x) \equiv \text{prox}_{f_2}(y)$$

- Analysis of the prox for *structured* sparse sets $g(x) = 0$

$$\text{prox}_{f_2}(y) = \arg\min_{x : x \in \Sigma_{\mathcal{M}_K}} \|x - y\|$$

support of the solution <> modular approximation problem

$$\text{supp}\left(\arg\min_{x:\text{supp}(x)\in\mathcal{M}_K} \|x - y\|_2^2\right) = \arg\max_{\mathcal{S}:\mathcal{S}\in\bar{\mathcal{M}}_K} F(S; y)$$

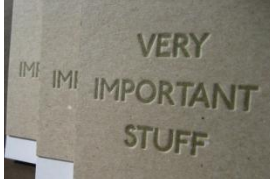
underlying optimization problem <> integer linear program

$$\text{supp}\left(\arg\min_z \left\{\rho^T z : z \in \Sigma_{\mathcal{M}_K}\right\}\right)$$

$z_i \in \{0, 1\}$: support indicator variables $\rho_i = -|y_i|^2$

[Kyrillidis and C, 2011]

Can we project onto non-convex sets?



$$\widehat{x} = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|y - x\|_2^2 + f_2(x) \equiv \mathrm{prox}_{f_2}(y)$$

- Analysis of the prox for *structured* sparse sets $g(x) = 0$

$$\mathrm{prox}_{f_2}(y) = \arg\min_{x : x \in \Sigma_{\mathcal{M}_K}} \|x - y\|$$

support of the solution <> modular approximation problem

$$\mathrm{supp}\left(\arg\min_{x:\mathrm{supp}(x)\in\mathcal{M}_K} \|x - y\|_2^2\right) = \arg\max_{\mathcal{S}:\mathcal{S}\in\bar{\mathcal{M}}_K} F(\mathcal{S}; y)$$

underlying optimization problem <> integer linear program

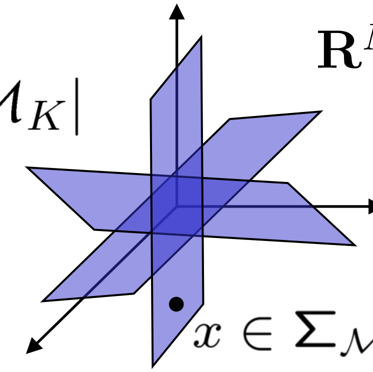
Class of problems we can tractably solve: ***PMAP***

- **Polynomial time modular epsilon-approximation property**

$$F(\widehat{\mathcal{S}}_\epsilon; y) \geq (1 - \epsilon)\max_{\mathcal{S}\in\bar{\mathcal{M}}_K} F(\mathcal{S}; y)$$

[Kyrillidis and C, 2011]

Can we project onto non-convex sets?



PMAP-0:

- Matroid structured sparse models:

$$\mathcal{M} = (\mathcal{N}, \mathcal{I} \subseteq 2^{\mathcal{N}}), \mathcal{N} = \{1, \ldots, N\}$$

$\mathcal{N}$: ground set

$\mathcal{I}$: base set

Definition:

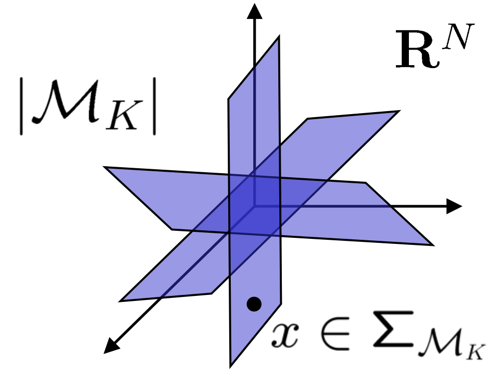
non-emptiness 1. $\emptyset \in \mathcal{I}$

heredity 2. $A \in \mathcal{I}$ and $B \subseteq A \Rightarrow B \in \mathcal{I}$

exchange 3. $A, B \in \mathcal{I}$ and $|A| > |B| \Rightarrow \exists e \in A \setminus B$ such that $B \cup \{e\} \in \mathcal{I}$

[Nemhauser and Wolsey, 1999]

Can we project onto non-convex sets?



PMAP-0:

- Matroid structured sparse models:

$$\mathcal{M} = (\mathcal{N}, \mathcal{I} \subseteq 2^{\mathcal{N}}),\, \mathcal{N} = \{1, \ldots, N\}$$

$\mathcal{N}$: ground set

$\mathcal{I}$: base set

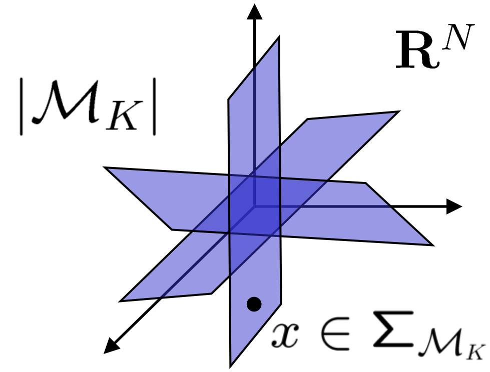
Definition:

non-emptiness	1. $\emptyset \in \mathcal{I}$
heredity	2. $A \in \mathcal{I}$ and $B \subseteq A \Rightarrow B \in \mathcal{I}$
exchange	3. $A, B \in \mathcal{I}$ and $|A| > |B| \Rightarrow \exists e \in A \setminus B$ such that $B \cup \{e\} \in \mathcal{I}$

Let $\mathcal{N} = \{1, 2, 3, 4\}$. The smallest matroid that contains $\{1, 2\}$ and $\{3, 4\}$ is ???

[Nemhauser and Wolsey, 1999]

Can we project onto non-convex sets?



PMAP-0:

- Matroid structured sparse models:

$$\mathcal{M} = (\mathcal{N}, \mathcal{I} \subseteq 2^{\mathcal{N}}), \mathcal{N} = \{1, \dots, N\}$$

$\mathcal{N}$: ground set

$\mathcal{I}$: base set

Definition:

non-emptiness	1. $\emptyset \in \mathcal{I}$
heredity	2. $A \in \mathcal{I}$ and $B \subseteq A \Rightarrow B \in \mathcal{I}$
exchange	3. $A, B \in \mathcal{I}$ and $\|A\| > \|B\| \Rightarrow \exists e \in A \setminus B$ such that $B \cup \{e\} \in \mathcal{I}$

Let $\mathcal{N} = \{1, 2, 3, 4\}$. The smallest matroid that contains $\{1, 2\}$ and $\{3, 4\}$

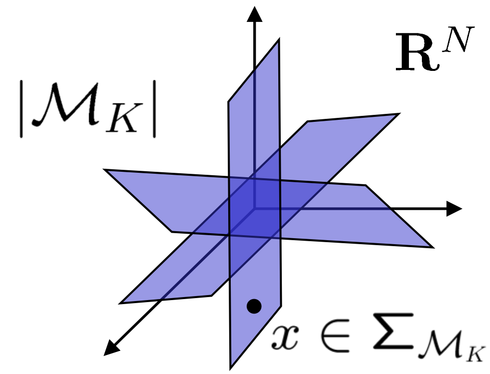
$\mathcal{I} = \{$ ∅, *by the non-emptiness property*

{1}, {2}, {3}, {4}, {1,2}, {3,4}, *by the heredity property*

{1,3}, {1,4}, {2,3}, {2,4} *by the exchange property*

}

Can we project onto non-convex sets?



PMAP-0:

- Matroid structured sparse models:

$$\mathcal{M} = (\mathcal{N}, \mathcal{I} \subseteq 2^{\mathcal{N}}), \mathcal{N} = \{1, \ldots, N\}$$

Definition:
1. $\emptyset \in \mathcal{I}$
2. $A \in \mathcal{I}$ and $B \subseteq A \Rightarrow B \in \mathcal{I}$
3. $A, B \in \mathcal{I}$ and $|A| > |B| \Rightarrow \exists e \in A \setminus B$ such that $B \cup \{e\} \in \mathcal{I}$

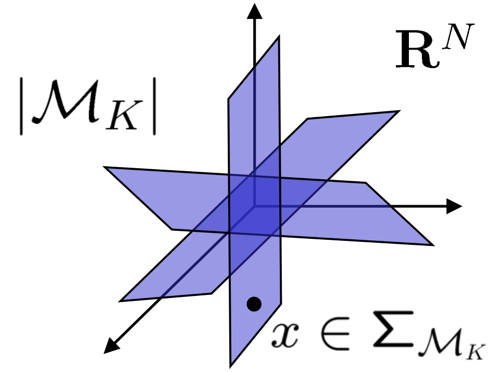
Greedy basis algorithm efficiently solves $\arg\max_{\mathcal{S}:\mathcal{S}\in\mathcal{M}} \sum_{i\in\mathcal{S}} w_i^2$

sort $\mathcal{N}$ in decreasing order by weight w_i^2

start with empty set: $\mathcal{S}_0 = \emptyset$

1. $\mathcal{R}_i = \{r_i \in \mathcal{N} \setminus \mathcal{S}_i\}$ while keeping the order
2. $r = \arg\max_j \{w_j^2 : (j \in \mathcal{R}_i) \wedge (\mathcal{S}_i \cup \{j\} \in \mathcal{I})\}$
3. $\mathcal{S}_{i+1} = \mathcal{S}_i \cup \{r\}$

Can we project onto non-convex sets?



PMAP-0:

- Matroid structured sparse models:

$$\mathcal{M} = (\mathcal{N}, \mathcal{I} \subseteq 2^{\mathcal{N}}), \mathcal{N} = \{1, \ldots, N\}$$

Definition:
1. $\emptyset \in \mathcal{I}$
2. $A \in \mathcal{I}$ and $B \subseteq A \Rightarrow B \in \mathcal{I}$
3. $A, B \in \mathcal{I}$ and $|A| > |B| \Rightarrow \exists e \in A \setminus B$ such that $B \cup \{e\} \in \mathcal{I}$

Greedy basis algorithm efficiently solves matroid constrained problems

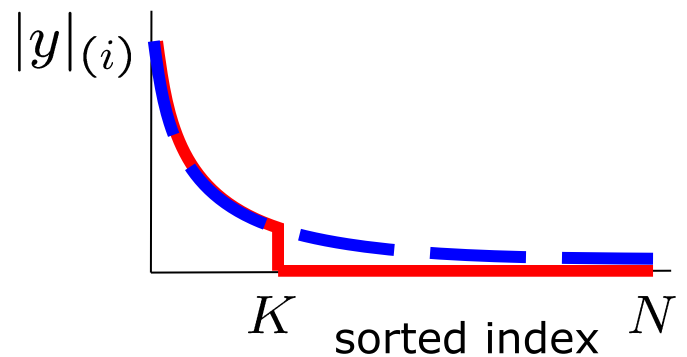
Examples:

1. uniform matroid: $\mathcal{I} = \{\mathcal{S} : \mathcal{S} \subseteq \mathcal{N}, |\mathcal{S}| \leq K\}$

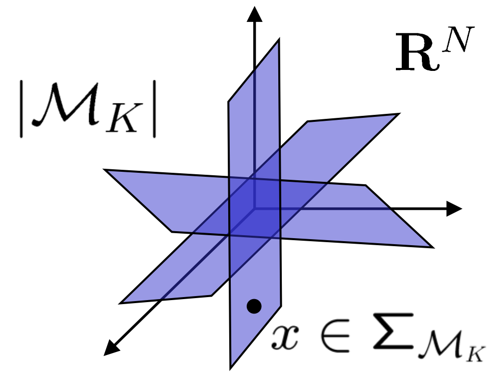
$$\underbrace{\text{prox}_{f_2}(y) = \arg \min_{x : x \in \Sigma_K} \|x - y\|}$$

hard thresholding!

$$H_K(y)$$



Can we project onto non-convex sets?



PMAP-0:

- Matroid structured sparse models:

$$\mathcal{M} = (\mathcal{N}, \mathcal{I} \subseteq 2^{\mathcal{N}}), \mathcal{N} = \{1, \ldots, N\}$$

Definition:
1. $\emptyset \in \mathcal{I}$
2. $A \in \mathcal{I}$ and $B \subseteq A \Rightarrow B \in \mathcal{I}$
3. $A, B \in \mathcal{I}$ and $|A| > |B| \Rightarrow \exists e \in A \setminus B$ such that $B \cup \{e\} \in \mathcal{I}$

Greedy basis algorithm efficiently solves matroid constrained problems

Examples:

[Kyrillidis and C, 2011]

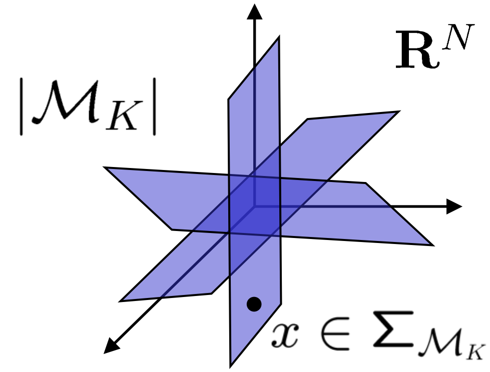
1. uniform matroid <> simple sparsity

intersection with the following matroids (result is still a matroid!*)

2. partition matroid <> distributed sparsity
3. graphic matroid <> spanning tree sparsity
4. matching matroid <> graph matching sparsity

*: in general, the intersection of two matroids is not a matroid.

Can we project onto non-convex sets?



PMAP-0:

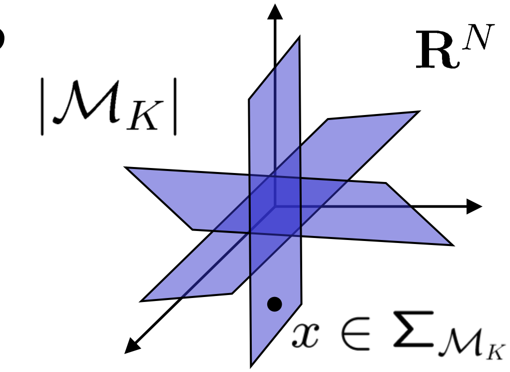
- Linear support constraints:

Definition: $\Sigma_{\mathcal{M}_K} = \bigcup_{\forall z \in \mathfrak{Z}} \text{supp}(z),$ where $\mathfrak{Z} := \left\{ z \in \{0,1\}^N : Az \leq b \right\}$

A and b	<>	integral
first row of A	<>	all 1's
first entry of b	<>	K

[Kyrillidis and C, 2011]

Can we project onto non-convex sets?

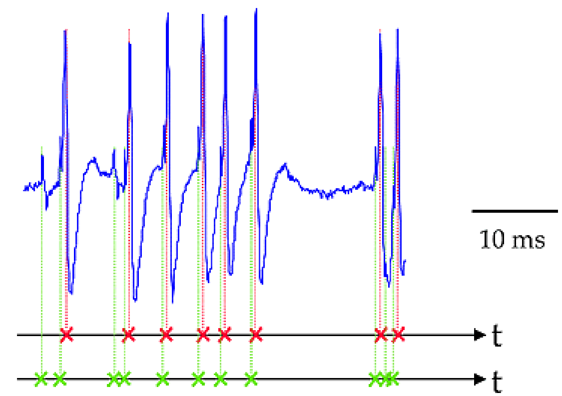


PMAP-0:

- Linear support constraints:

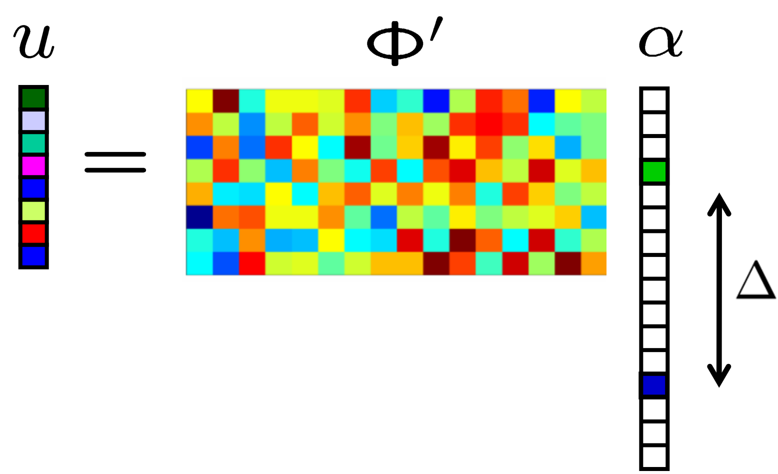
Definition: $\Sigma_{\mathcal{M}_K} = \bigcup_{\forall z \in \mathfrak{Z}} \text{supp}(z), \text{ where } \mathfrak{Z} := \{z \in \{0,1\}^N : Az \leq b\}$

Example: neuronal spike model

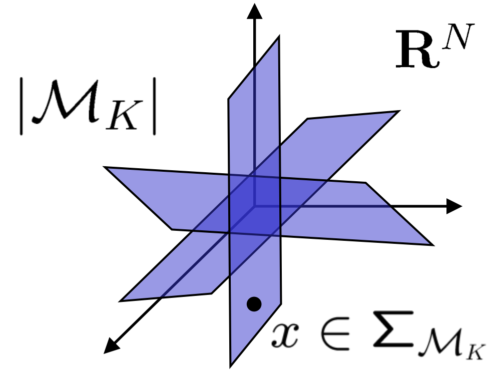


$z \in \{0,1\}^N$: binary support variables

$$
\begin{aligned}
z_1 + z_2 + \ldots + z_N &\leq K \\
z_1 + z_2 + \ldots + z_\Delta &\leq 1 \\
z_2 + z_3 + \ldots + z_{\Delta+1} &\leq 1 \\
&\vdots \\
z_{N-\Delta+1} + z_{N-\Delta+2} + \ldots + z_N &\leq 1
\end{aligned}
$$



Can we project onto non-convex sets?



PMAP-0:

- Linear support constraints:

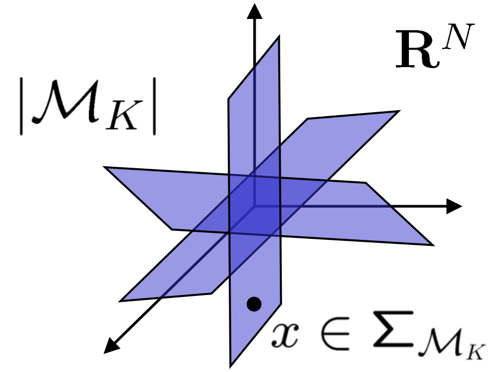
Definition: $\Sigma_{\mathcal{M}_K} = \bigcup_{\forall z \in \mathfrak{Z}} \text{supp}(z)$, where $\mathfrak{Z} := \left\{ z \in \{0,1\}^N : Az \leq b \right\}$

We can use LP can relax the LS constrained ILPs:

$$\arg\min_{z} \left\{ \rho^T z : z \in [0,1]^N, Az \leq b \right\} \qquad \rho_i = -|y_i|^2$$

...but, when is the result binary?

Can we project onto non-convex sets?



PMAP-0:

- Linear support constraints:

Definition: $\Sigma_{\mathcal{M}_K} = \bigcup_{\forall z \in \mathfrak{Z}} \text{supp}(z)$, where $\mathfrak{Z} := \left\{ z \in \{0,1\}^N : Az \leq b \right\}$

LP can *exactly* solve the LS constrained ILPs:

$$\arg\min_{z} \left\{ \rho^T z : z \in [0,1]^N, Az \leq b \right\} \qquad \rho_i = -|y_i|^2$$

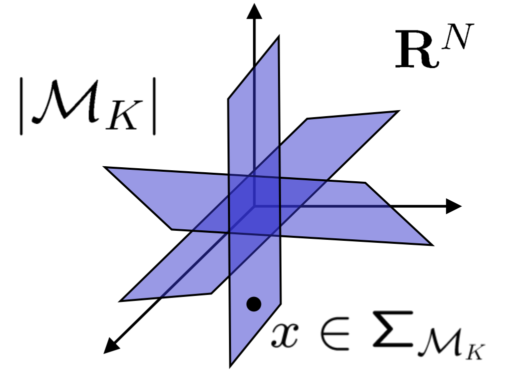
...when A is totally unimodular (TU)*! [Nemhauser and Wolsey, 1999]

– the determinant of each square submatrix is {-1,0,1}

Examples: interval matrices, perfect matrices, network matrices

*: if we want LP relaxation to work for all b, TU is a necessary condition.

Can we project onto non-convex sets?

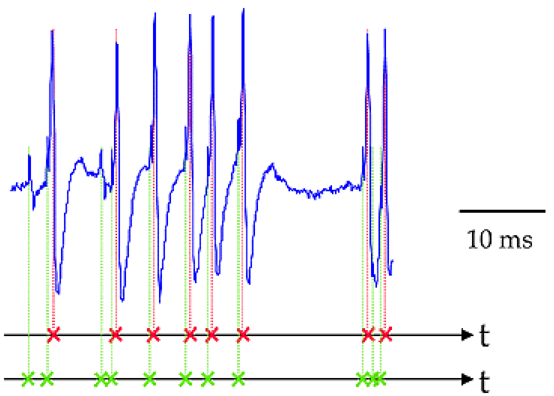


PMAP-0:

- Linear support constraints:

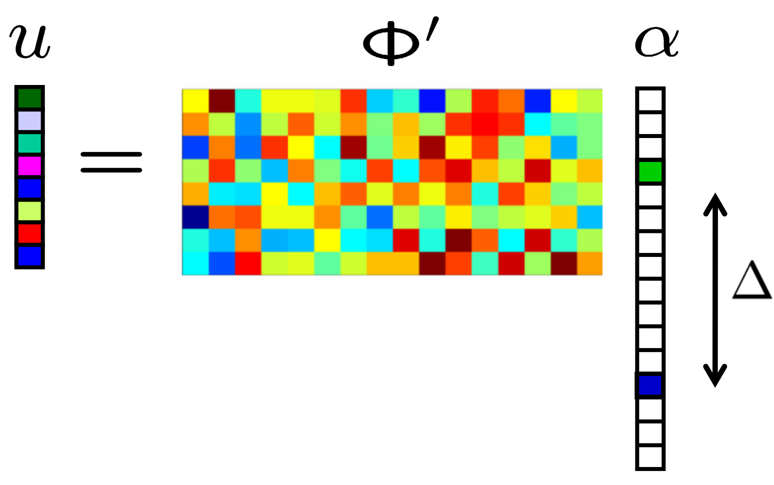
Definition: $\Sigma_{\mathcal{M}_K} = \bigcup_{\forall z \in \mathfrak{Z}} \text{supp}(z), \text{ where } \mathfrak{Z} := \left\{ z \in \{0,1\}^N : Az \leq b \right\}$

Example: neuronal spike model



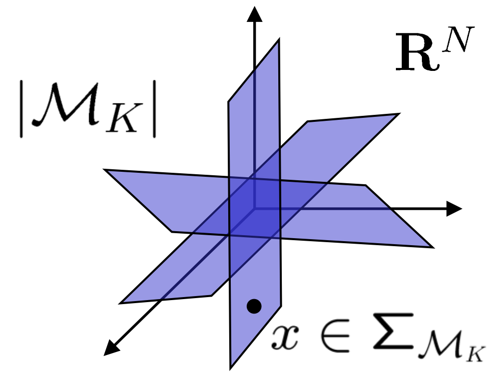
$z \in \{0,1\}^N$: binary support variables

$$\left.\begin{aligned} z_1 + z_2 + \ldots + z_N &\leq K \\ z_1 + z_2 + \ldots + z_\Delta &\leq 1 \\ z_2 + z_3 + \ldots + z_{\Delta+1} &\leq 1 \\ &\vdots \\ z_{N-\Delta+1} + z_{N-\Delta+2} + \ldots + z_N &\leq 1 \end{aligned}\right\} \text{TU}$$



[Hegde, Duarte, and C, 2009]

Can we project onto non-convex sets?



PMAP-0:

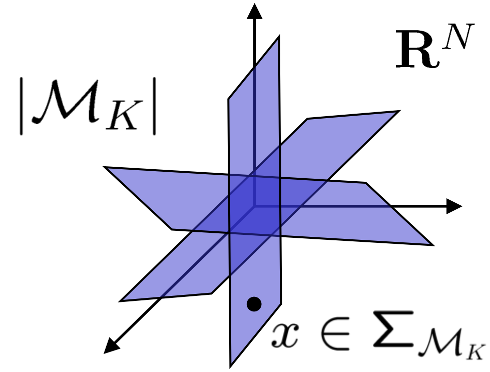
- prox-sparse models

Definition: define algorithmically!

$$\operatorname{prox}_{f_2}(y) = \arg \min_{x : x \in \Sigma_{\mathcal{M}_K}} \|x - y\| \qquad g(x) = 0$$

[Kyrillidis and C, 2011]

Can we project onto non-convex sets?



PMAP-0:

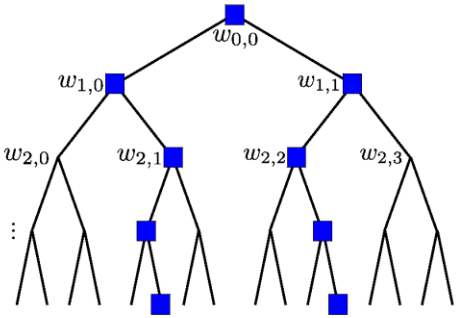
- prox-sparse models

Definition: define algorithmically!

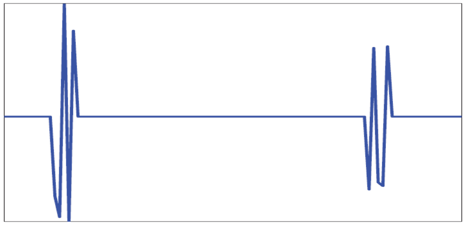
$$\text{prox}_{f_2}(y) = \arg \min_{x : x \in \Sigma_{\mathcal{M}_K}} \|x - y\| \qquad g(x) = 0$$

Example: clustered sparsity models

- tree-sparse <> dynamic program



- clustered sparse <> dynamic program



[Baraniuk, C, Wakin 2010; Baraniuk, C, Duarte, Hegde 2010]

Can we project onto non-convex sets?

Pop-quiz: A prox with convex and non-convex terms

Let us consider $f_2(x) = \|x\|_1 + \iota_{\{x : \|x\|_0 \leq K\}}(x)$ $g(x) = \|x\|_1$

$$\mathrm{prox}_{f_2}(y) = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|y - x\|_2^2 + f_2(x)$$

Is it PMAP-0?

Can we project onto non-convex sets?

Pop-answer: A prox with convex and non-convex terms

Let us consider $f_2(x) = \|x\|_1 + \iota_{\{x : \|x\|_0 \le K\}}(x)$ $\qquad g(x) = \|x\|_1$

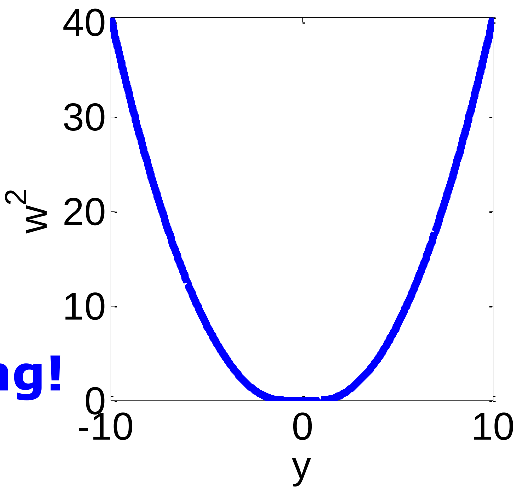
$$\mathrm{prox}_{f_2}(y) = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|y - x\|_2^2 + f_2(x)$$

$$\mathrm{supp}\left(\mathrm{prox}_{f_2}(y)\right) = \arg\max_{\mathcal{S} : |\mathcal{S}| \le K} F(\mathcal{S}; y)$$

$$F(\mathcal{S}; y) = \tfrac{1}{2}\|y\|^2 - \min_{x : \mathrm{supp}(x) = \mathcal{S}} \tfrac{1}{2}\|y - x\|_2^2 + \|x\|_1$$

$$\Rightarrow F(\mathcal{S}; y) = \textstyle\sum_{i \in \mathcal{S}} w_i^2$$

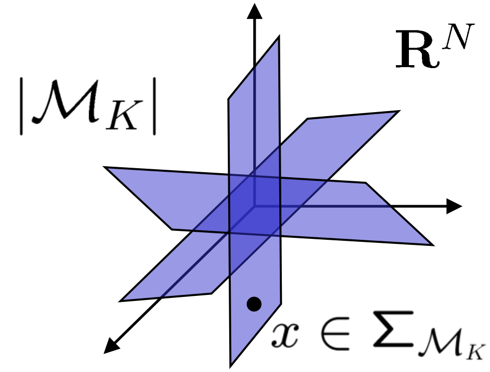
$$w_i^2 = y_i \times \mathrm{soft}(y_i, 1) - \tfrac{1}{2}|\mathrm{soft}(y_i, 1)|^2 - |\mathrm{soft}(y_i, 1)|$$



Hard thresholding followed by soft thresholding!

YES: certified PMAP-0

Can we project onto non-convex sets?



PMAP-epsilon: $F(\widehat{\mathcal{S}}_\epsilon; y) \geq (1-\epsilon)\max_{\mathcal{S}\in\bar{\mathcal{M}}_K} F(\mathcal{S}; y)$

- **Knapsack**

  multi-knapsack constraints

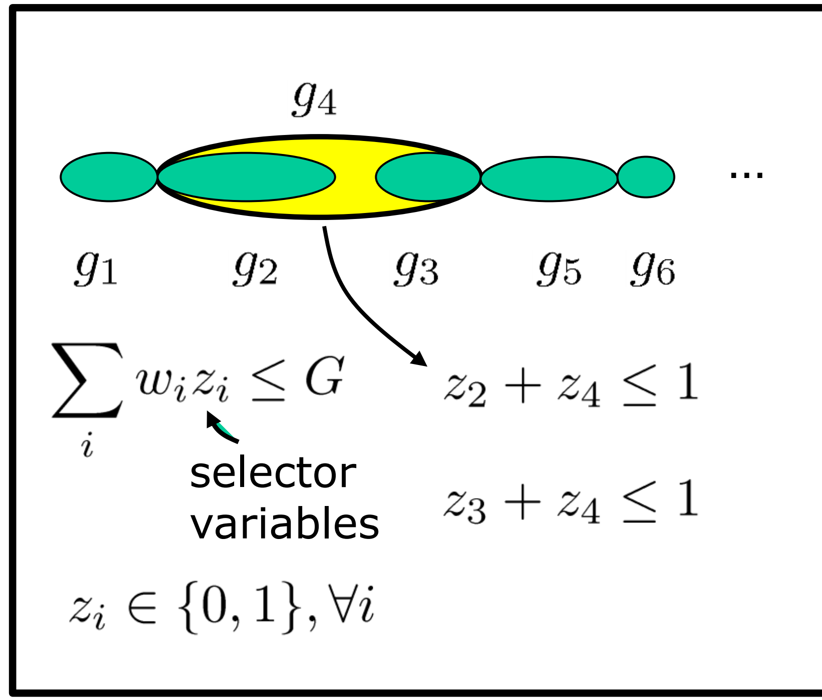
  weighted multi-knapsack

  Ex: Nested group sparse problems

  quadratically-constrained

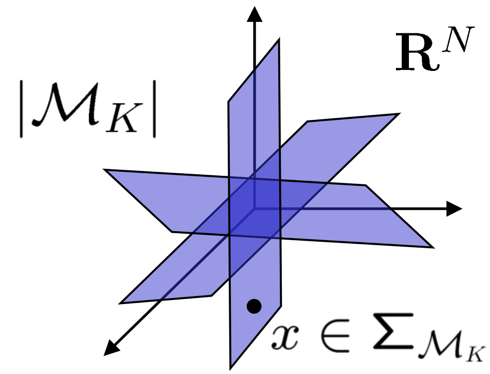
- **Define algorithmically!**

  approximate solutions for computational reasons



[Kyrillidis and C, 2011]

Can we project onto non-convex sets?



PMAP-epsilon: $F(\widehat{\mathcal{S}}_\epsilon; y) \geq (1-\epsilon)\max_{\mathcal{S}\in\bar{\mathcal{M}}_K} F(\mathcal{S}; y)$

- **Knapsack**

  multi-knapsack constraints

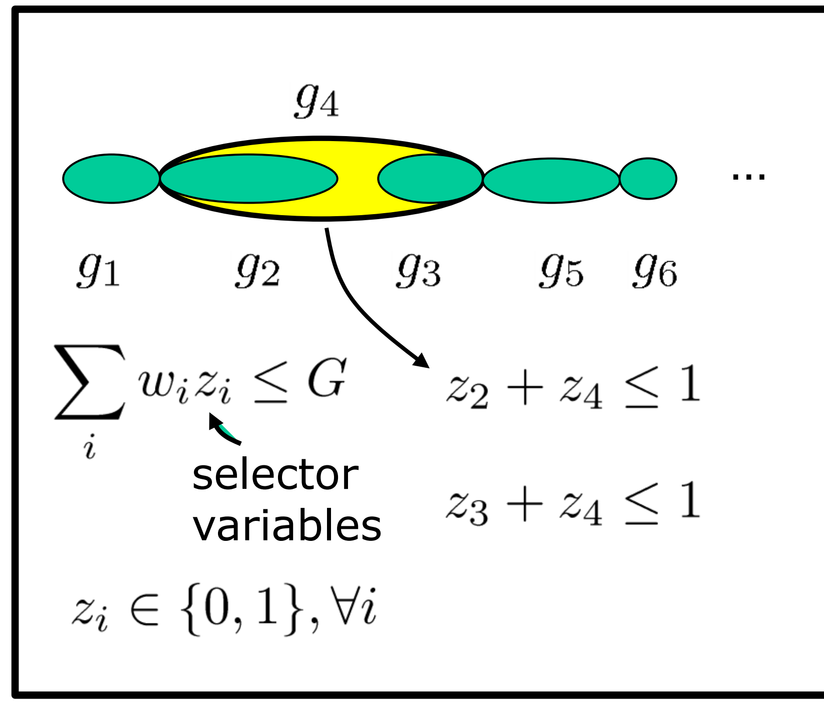
  weighted multi-knapsack

  Ex: Nested group sparse problems

  quadratically-constrained

- **Define algorithmically!**

  approximate solutions for computational reasons

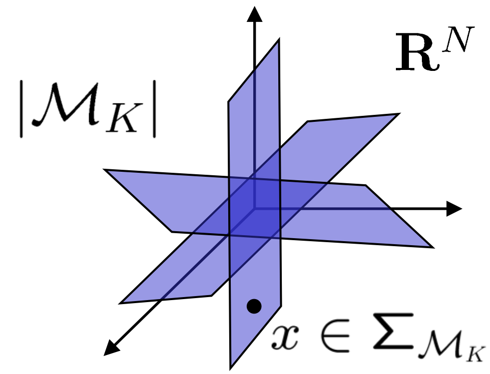


- **Pairwise overlapping groups** <> quadratic binary w/ cardinality cons.

$$\max_{\mathcal{S}:\mathcal{S}\in\bar{\mathcal{M}}_K} F(S;y) = -\min\left\{\sum_{i>j}\|(y)_{g_i\cap g_j}\|_2^2 z_i z_j - \sum_i \|(y)_{g_i}\|_2^2 z_i : \sum_i z_i \leq G\right\}.$$

we can only approximate... and epsilon is large!

Can we project onto non-convex sets?



PMAP-epsilon: $F(\widehat{\mathcal{S}}_\epsilon; y) \geq (1-\epsilon)\max_{\mathcal{S}\in\bar{\mathcal{M}}_K} F(\mathcal{S}; y)$

- **Knapsack**

  multi-knapsack constraints

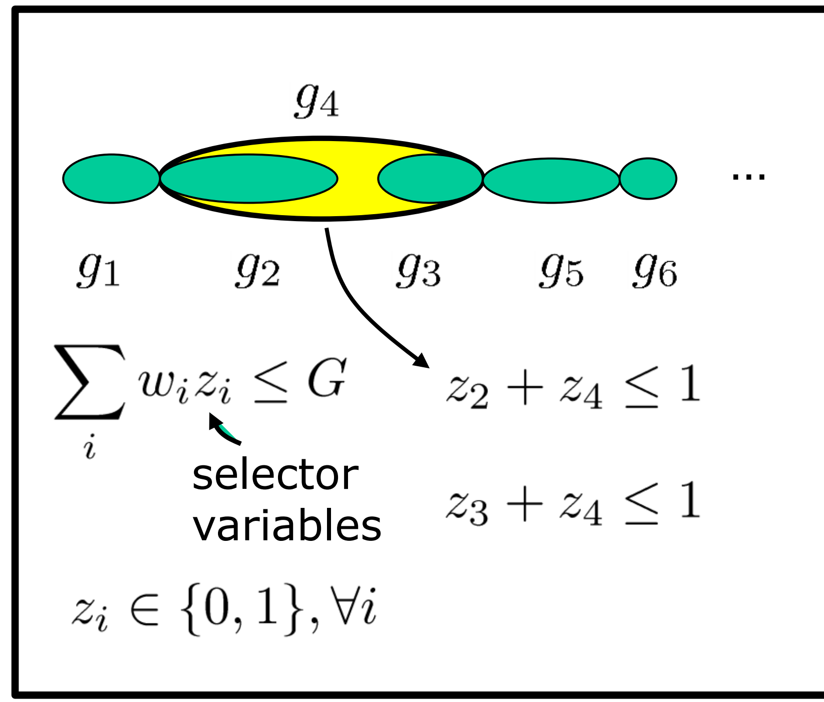
  weighted multi-knapsack

  Ex: Nested group sparse problems

  quadratically-constrained

- **Define algorithmically!**

  approximate solutions for computational reasons

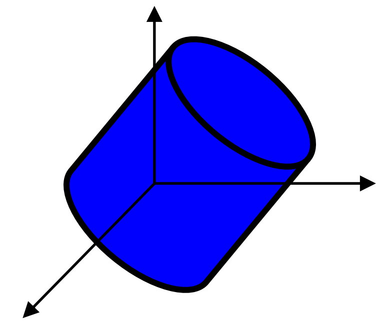


- **Pairwise overlapping groups** <> quadratic binary w/ cardinality cons.

- **Multi-knapsack + multi-matroids** [Lee et al., 2009]

we can only approximate... and epsilon is large!

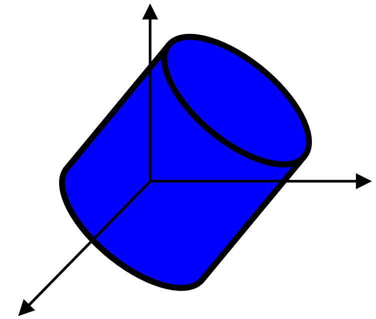
Can we project onto non-convex sets?



Matrix examples!

- Rank constrained projections $\text{prox}_{f_2}(Y) = \arg \min_{X:\text{rank}(X) \leq R} \|X - Y\|_F$

Can we project onto non-convex sets?



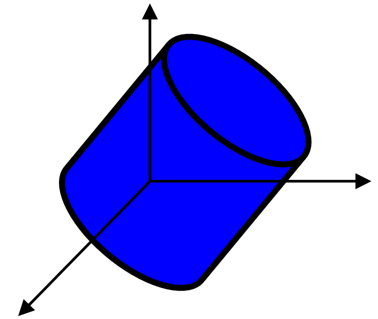
Matrix examples!

- Rank constrained projections $\text{prox}_{f_2}(Y) = \arg \min_{X:\text{rank}(X) \leq R} \|X - Y\|_F$

$$\arg \min_{X:\text{rank}(X) \leq R} \|X - Y\|_F = \arg \min_{X:\text{rank}(X) \leq R} \|X - U\Lambda_Y V^T\|_F$$

singular value decomposition

Can we project onto non-convex sets?



Matrix examples!

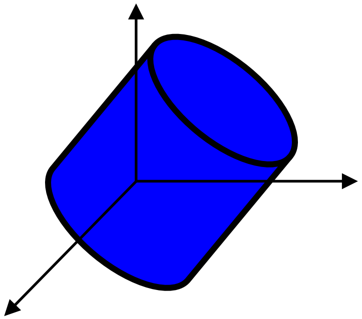
- Rank constrained projections $\mathrm{prox}_{f_2}(Y) = \arg \min_{X:\mathrm{rank}(X)\leq R} \|X - Y\|_F$

$$\arg \min_{X:\mathrm{rank}(X)\leq R} \|X - Y\|_F = \arg \min_{X:\mathrm{rank}(X)\leq R} \|X - U\Lambda_Y V^T\|_F$$

$$= \arg \min_{X:\mathrm{rank}(X)\leq R} \|U^T X V - \Lambda_Y\|_F$$

invariance to unitary transform

Can we project onto non-convex sets?



Matrix examples!

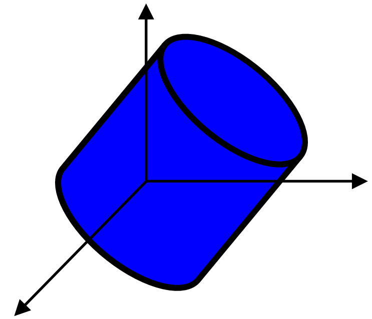
- Rank constrained projections $\operatorname{prox}_{f_2}(Y) = \arg \min_{X : \operatorname{rank}(X) \leq R} \|X - Y\|_F$

$$\arg \min_{X : \operatorname{rank}(X) \leq R} \|X - Y\|_F = \arg \min_{X : \operatorname{rank}(X) \leq R} \|X - U\Lambda_Y V^T\|_F$$

$$= U \left(\arg \min_{\tilde{X} : \operatorname{rank}(\tilde{X}) \leq R} \|\tilde{X} - \Lambda_Y\|_F \right) V^T$$

sparse approximation problem!

Can we project onto non-convex sets?



Matrix examples!

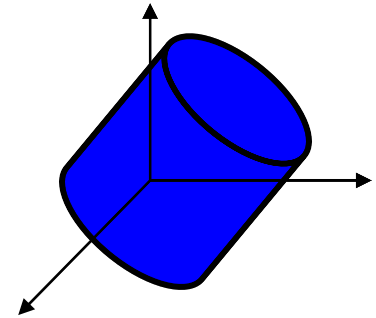
- Rank constrained projections $\mathrm{prox}_{f_2}(Y) = \arg \min_{X:\mathrm{rank}(X)\leq R} \|X - Y\|_F$

$$\arg \min_{X:\mathrm{rank}(X)\leq R} \|X - Y\|_F = \arg \min_{X:\mathrm{rank}(X)\leq R} \|X - U\Lambda_Y V^T\|_F$$

$$= U H_R(\Lambda_Y) V^T$$

singular value (hard) thresholding

Can we project onto non-convex sets?



Matrix examples!

- Rank constrained projections $\text{prox}_{f_2}(Y) = \arg \min_{X:\text{rank}(X) \leq R} \|X - Y\|_F$

$$\arg \min_{X:\text{rank}(X) \leq R} \|X - Y\|_F = \arg \min_{X:\text{rank}(X) \leq R} \|X - U\Lambda_Y V^T\|_F$$

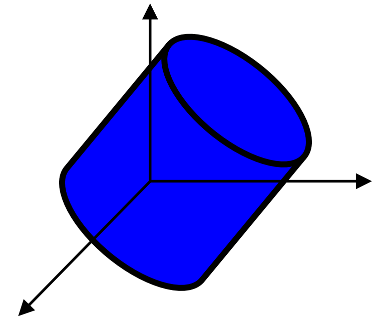
$$= U H_R(\Lambda_Y) V^T$$

singular value (hard) thresholding

- Non-convex spectral projections <> sets described by their eigenvalue properties
  - exact projections >> basic operations on eigenvalues

[Lewis and Malick 2008]

Can we project onto non-convex sets?



Matrix examples!

- Rank constrained projections $\text{prox}_{f_2}(Y) = \arg \min_{X:\text{rank}(X)\leq R} \|X - Y\|_F$

- Non-convex spectral projections <> sets described by their eigenvalue properties

- epsilon-approximate projections (note the difference with PMAP)

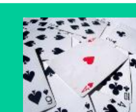
$$\|\text{prox}^{\epsilon}_{f_2}(Y) - Y\|_F \leq (1+\epsilon) \min_{X:\text{rank}(X)\leq R} \|X - Y\|_F$$

Two highlights:

- structure from randomness/power methods [Halko, Martinsson, Tropp, 2010]
- column subset selection approaches [Boutsidis, Mahoney, Drineas, 2010]

Recovery algorithms for low-dimensional models

Now that we have projections...

	Non-convex $\binom{N}{K}$	**Convex** 	**Probabilistic** 
Encoding	combinatorial / manifolds	atomic norm / convex relaxation	compressible / sparse priors

A common criteria covering a broad set of applications:

$$\min_X \|u - \Phi(X)\|^2 \text{ s.t. } X = S + L, \|S\|_0 \leq K, \text{rank}(L) \leq R$$

- *affine rank minimization, matrix completion, robust PCA…*

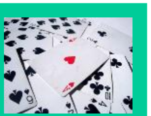
[Candes and Recht 2009; Waters, Sankaranayanan, Baraniuk, 2011]

A common algorithm: ***projected gradient***

$\|S\|_0 = \#\{S_i \neq 0\}$

Recovery algorithms for low-dimensional models

To highlight the salient differences, we will consider

	Non-convex $\binom{N}{K}$	**Convex** 	**Probabilistic** 
Encoding	combinatorial / manifolds	atomic norm / convex relaxation	compressible / sparse priors

compressive sensing recovery

$$\min_{x:\|x\|_0 \leq K} \|u - \Phi x\|^2$$

A common algorithm: ***projected gradient***

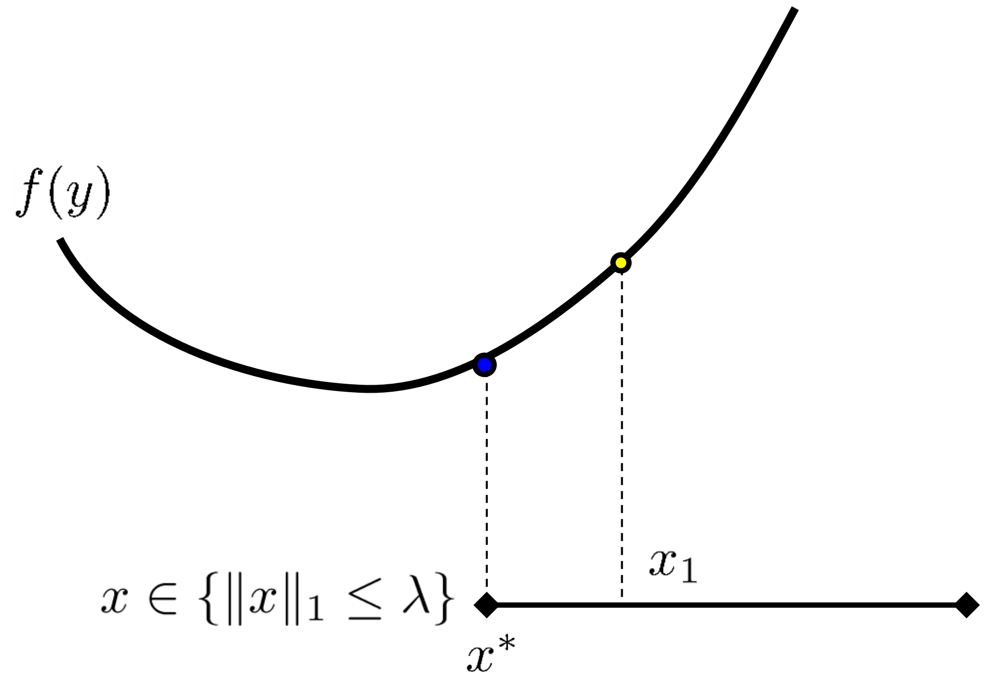
$$\|x\|_0 = \#\{x_i \neq 0\}$$

A tale of two algorithms

- Soft thresholding

$$f(x) = \|u - \Phi x\|^2$$

$\min_{x:\|x\|_1 \leq \lambda} f(x)$

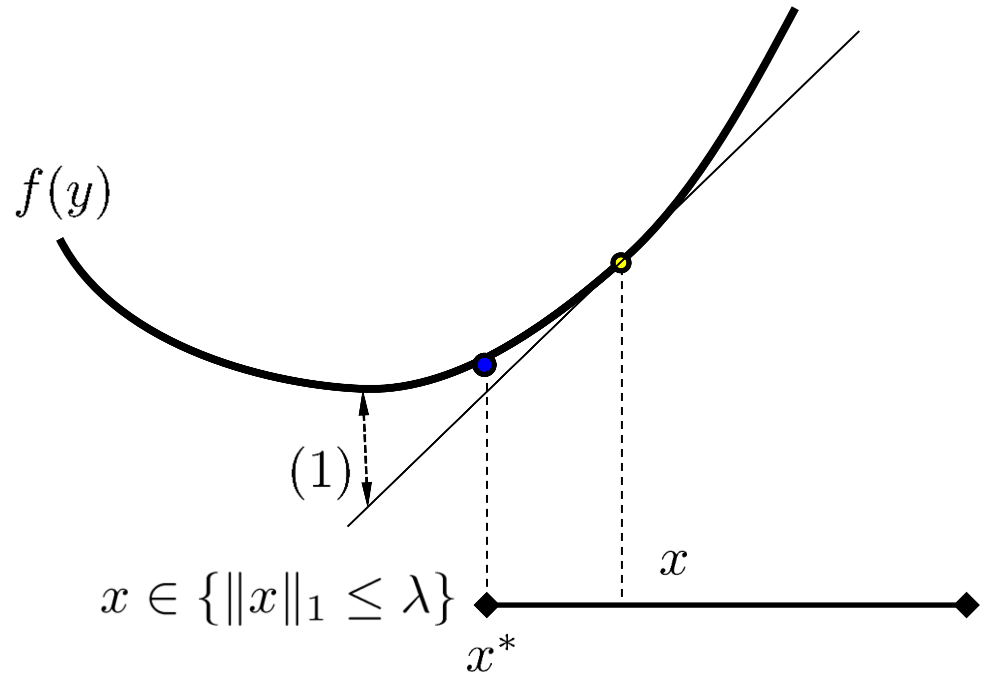


A tale of two algorithms

- Soft thresholding

$$f(x) = \|u - \Phi x\|^2$$

$\min_{x:\|x\|_1 \leq \lambda} f(x)$



Structure in optimization:

$$(1) \qquad f(y) - f(x) - \langle \nabla f(x), y - x \rangle \quad = \quad \|\Phi(y - x)\|^2 \qquad \forall x, y \in \mathcal{R}^N,$$

A tale of two algorithms

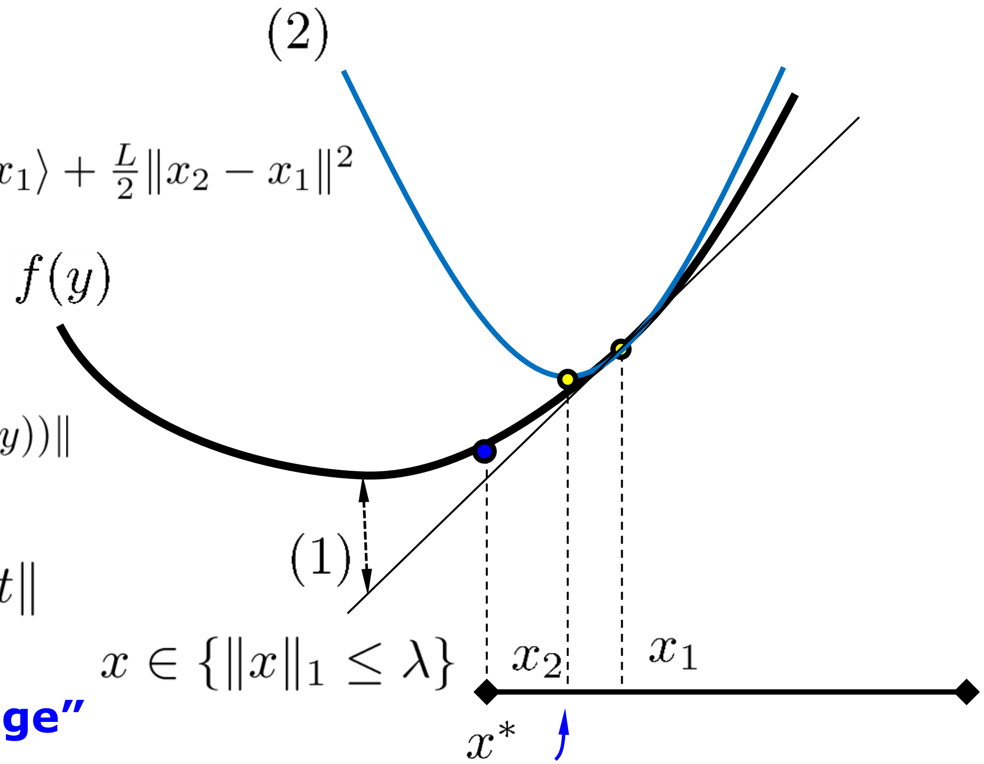
- Soft thresholding

$$f(x) = \|u - \Phi x\|^2$$

$\min_{x:\|x\|_1 \le \lambda} f(x)$

$$U(x_2, x_1) = f(x_1) + \langle \nabla f(x_1), x_2 - x_1 \rangle + \frac{L}{2}\|x_2 - x_1\|^2$$

majorization-minimization



$$\arg\min_{\|x\|_1 \le \lambda} U(x, y) = \arg\min_{\|x\|_1 \le \lambda} \|x - (y - \frac{1}{L}\nabla f(y))\|$$

$$\mathrm{St}_{\{\|x\|_1 \le \lambda\}}(t) = \arg\min_{\|x\|_1 \le \lambda} \|x - t\|$$

Key actor: "least absolute shrinkage"

$$x_{i+1} = \mathrm{St}_{\{\|x\|_1 \le \lambda\}}\left(x_i - \tfrac{1}{L}\nabla f(x_i)\right)$$

$$\begin{array}{llcll} (1) & f(y) - f(x) - \langle \nabla f(x), y - x \rangle & = & \|\Phi(y - x)\|^2 & \forall x, y \in \mathcal{R}^N, \\ (2) & f(y) - f(x) - \langle \nabla f(x), y - x \rangle & \le & \frac{L}{2}\|y - x\|^2 & L = 2\|\Phi\|^2, \forall x, y \in \mathcal{R}^N, \end{array}$$

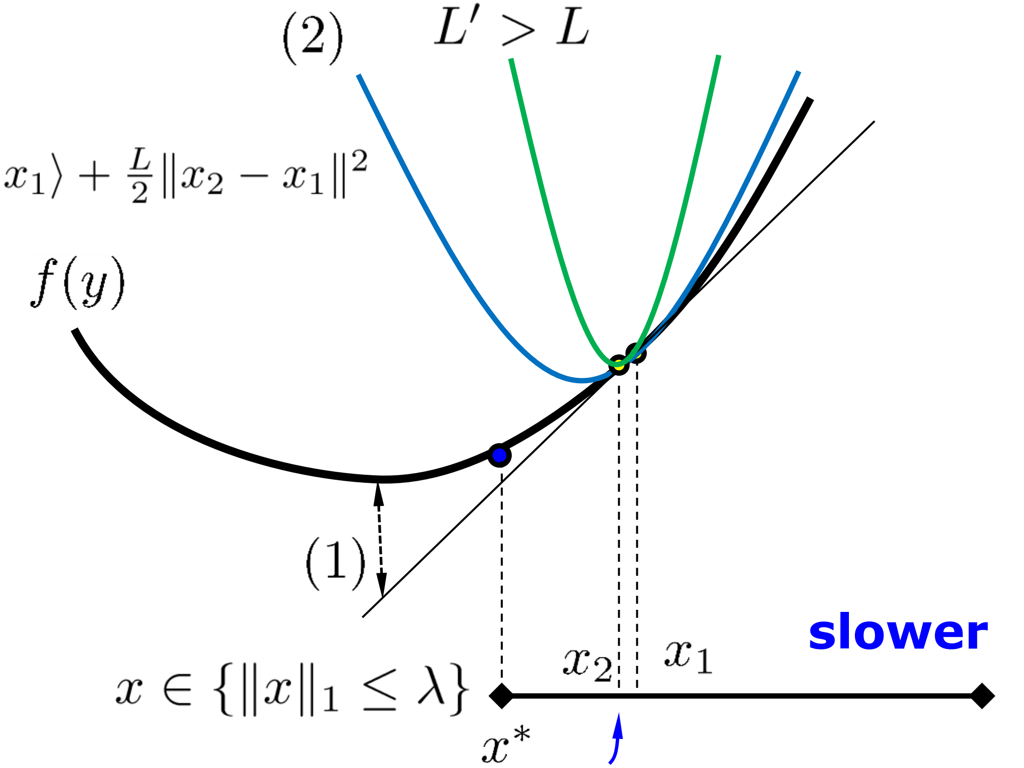
A tale of two algorithms

- Soft thresholding

$$f(x) = \|u - \Phi x\|^2$$

$$\min_{x:\|x\|_1 \leq \lambda} f(x)$$

$$U(x_2, x_1) = f(x_1) + \langle \nabla f(x_1), x_2 - x_1 \rangle + \frac{L}{2}\|x_2 - x_1\|^2$$



$$x_{i+1} = \mathrm{St}_{\{\|x\|_1 \leq \lambda\}}\left(x_i - \frac{1}{L'}\nabla f(x_i)\right)$$

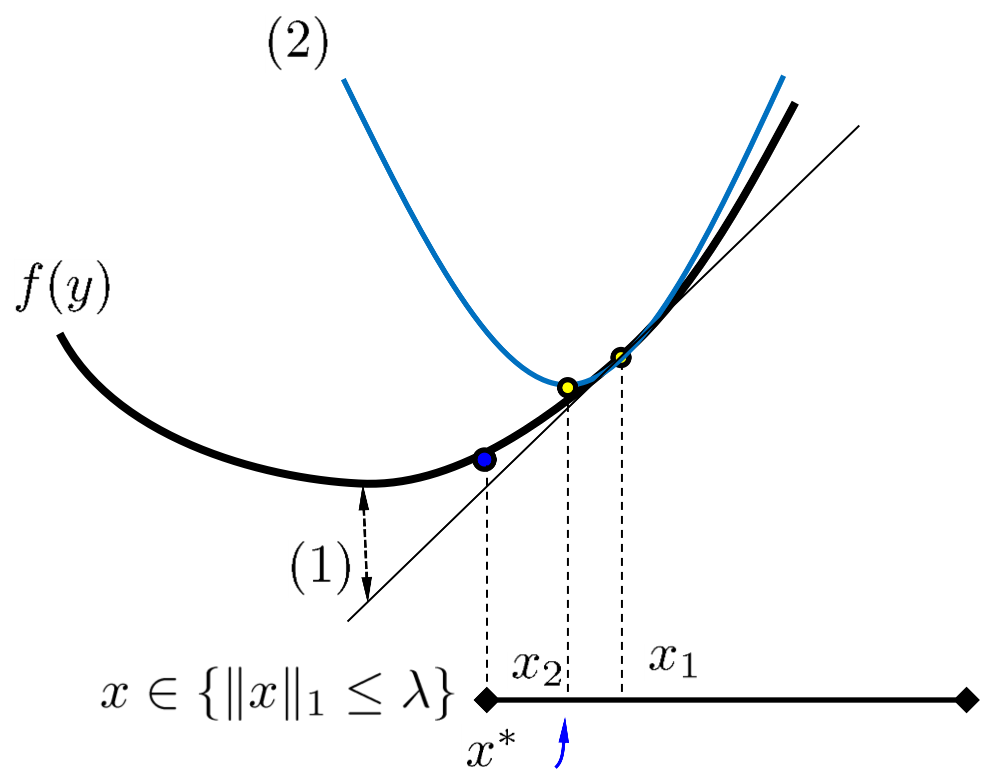
$$\begin{array}{llcll} (1) & f(y) - f(x) - \langle \nabla f(x), y - x \rangle & = & \|\Phi(y-x)\|^2 & \forall x, y \in \mathcal{R}^N, \\ (2) & f(y) - f(x) - \langle \nabla f(x), y - x \rangle & \leq & \frac{L}{2}\|y - x\|^2 & L = 2\|\Phi\|, \forall x, y \in \mathcal{R}^N, \end{array}$$

A tale of two algorithms

- Soft thresholding

$$f(x) = \|u - \Phi x\|^2$$

$\min_{x:\|x\|_1 \leq \lambda} f(x)$



$$x_{i+1} = \mathrm{St}_{\{\|x\|_1 \leq \lambda\}} \left(x_i - \tfrac{1}{L}\nabla f(x_i)\right)$$

- Is x* what we are looking for?

  local "unverifiable" assumptions:

  - ERC/URC/RSC condition
  - coherence based conditions …

  (local → global / dual certification)

[Buhlmann and van de Geer 2011]

A tale of two algorithms

$\binom{N}{K}$

- Hard thresholding

$$f(x) = \|u - \Phi x\|^2$$

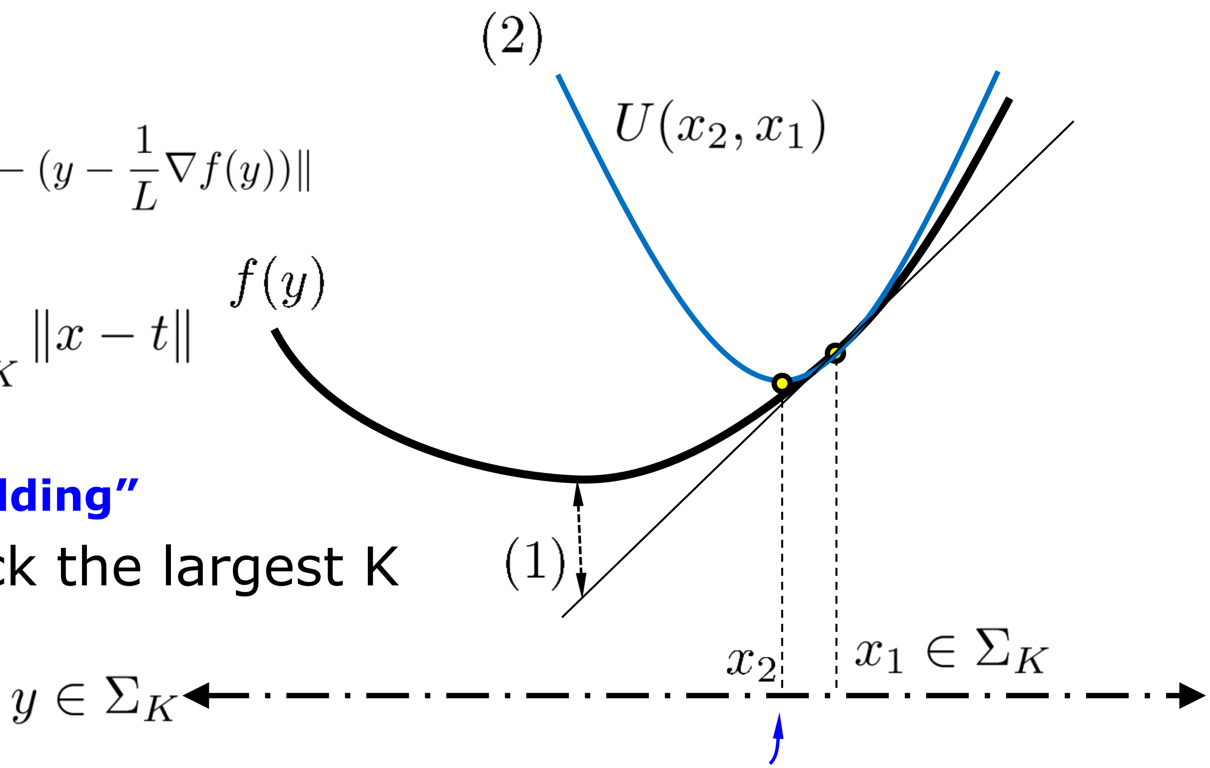
$\min_{x:\|x\|_0 \leq K} f(x)$

$$\arg\min_{\|x\|_0 \leq K} U(x, y) = \arg\min_{\|x\|_0 \leq K} \|x - (y - \frac{1}{L}\nabla f(y))\|$$

$$\mathrm{H}_{\{\|x\|_0 \leq K\}}(t) = \arg\min_{\|x\|_0 \leq K} \|x - t\|$$

Key actor: "hard thresholding"

ALGO: sort and pick the largest K



$$x_{i+1} = \mathrm{H}_{\{\|x\|_0 \leq K\}}\left(x_i - \frac{1}{L}\nabla f(x_i)\right)$$

A tale of two algorithms

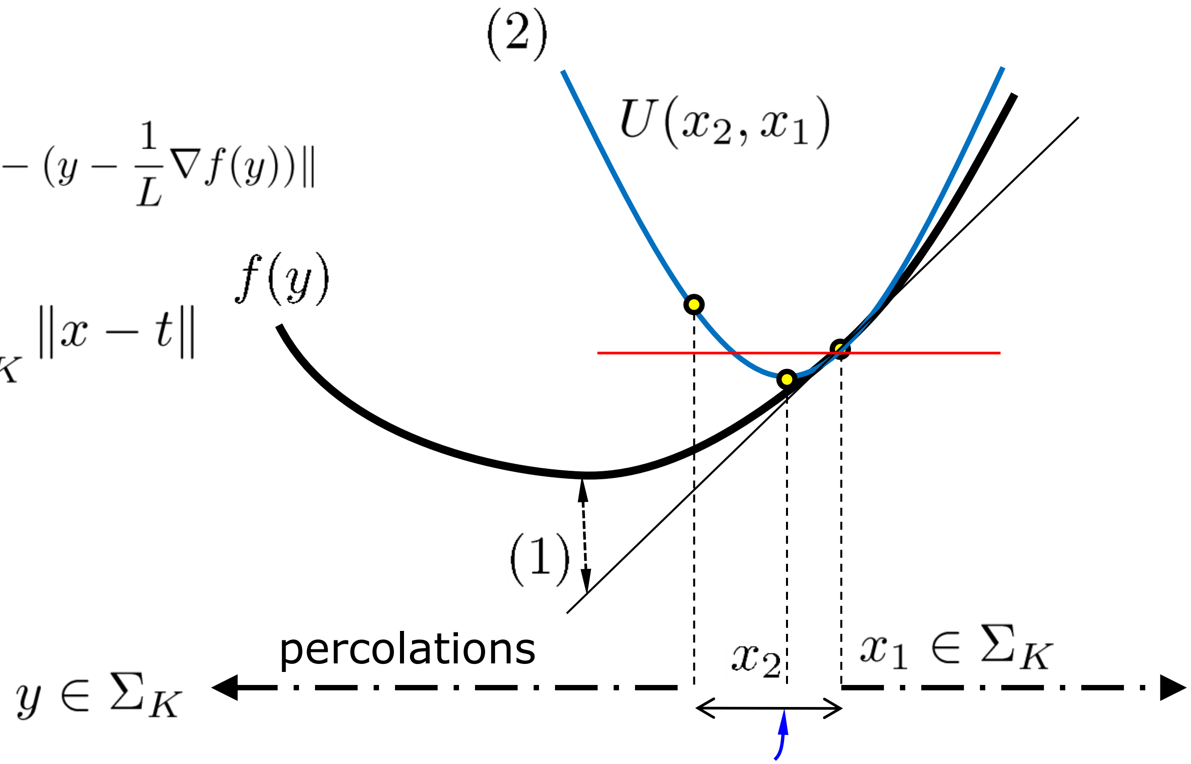
- Hard thresholding

$$f(x) = \|u - \Phi x\|^2$$

$\min_{x:\|x\|_0 \leq K} f(x)$

$$\arg \min_{\|x\|_0 \leq K} U(x,y) = \arg \min_{\|x\|_0 \leq K} \|x - (y - \frac{1}{L}\nabla f(y))\|$$

$$\mathrm{H}_{\{\|x\|_0 \leq K\}}(t) = \arg \min_{\|x\|_0 \leq K} \|x - t\|$$



$$x_{i+1} = \mathrm{H}_{\{\|x\|_0 \leq K\}}\left(x_i - \tfrac{1}{L}\nabla f(x_i)\right)$$

What could possibly go wrong with this naïve approach?

[C, 2011]

A tale of two algorithms

$\binom{N}{K}$

- Hard thresholding

$$f(x) = \|u - \Phi x\|^2$$

$\min_{x:\|x\|_0 \leq K} f(x)$

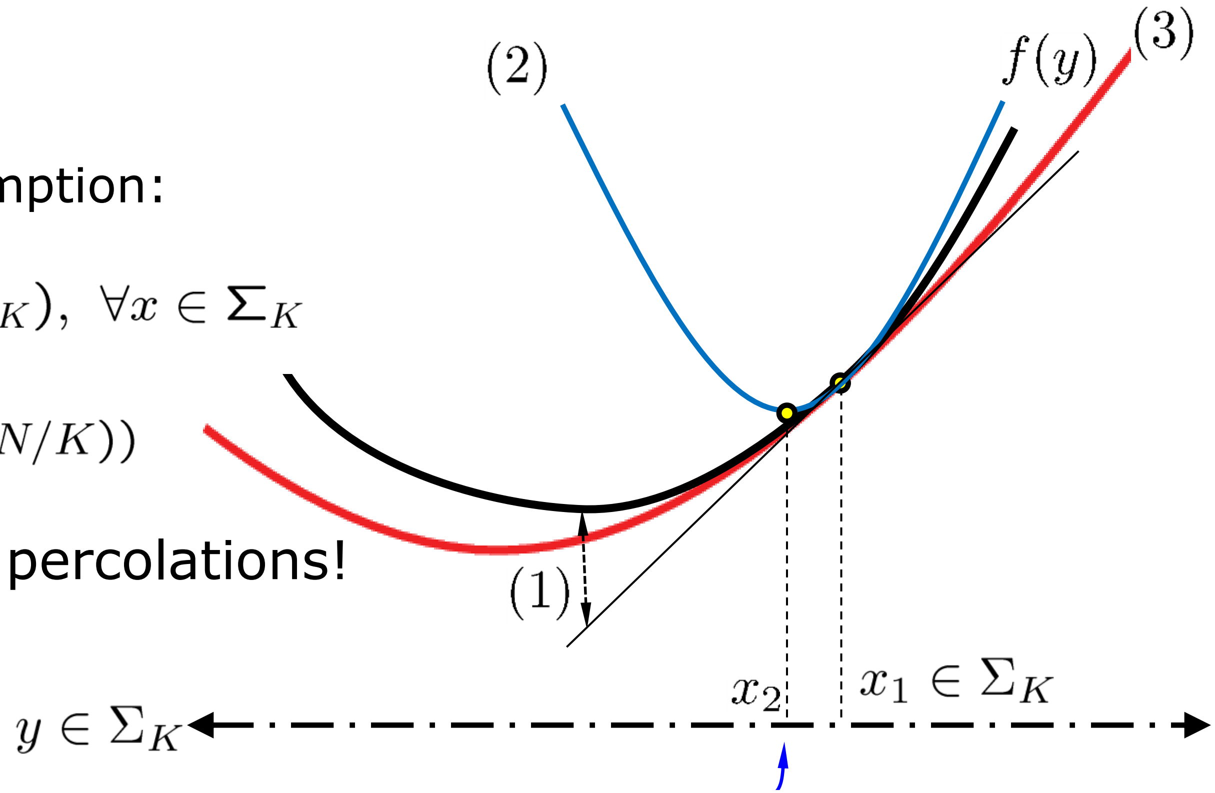
Global "unverifiable" assumption:

$$(1-\delta_K) \leq \frac{\|\Phi x\|_2^2}{\|x\|_2^2} \leq (1+\delta_K), \quad \forall x \in \Sigma_K$$

RIP condition $M = O(K \log(N/K))$

$\Longrightarrow$ we can tiptoe among percolations!

$\delta_{2K} < 1/3$



another variant has $\delta_{3K} < 1/2$

$$x_{i+1} = \mathrm{H}_{\{\|x\|_0 \leq K\}}\left(x_i - \frac{1}{L_{2K}} \nabla f(x_i)\right)$$

$$\begin{array}{llcll}
(1) & f(y) - f(x) - \langle \nabla f(x), y - x \rangle & = & \|\Phi(y-x)\|^2 & \forall x, y \in \mathcal{R}^N, \\
(2) & f(y) - f(x) - \langle \nabla f(x), y - x \rangle & \leq & \frac{L_{2K}}{2}\|y-x\|^2 & L_{2K} = 2(1+\delta_{2K}), \forall x, y \in \Sigma_K, \\
(3) & f(y) - f(x) - \langle \nabla f(x), y - x \rangle & \geq & \frac{\mu_{2K}}{2}\|y-x\|^2 & \mu_{2K} = 2(1-\delta_{2K}), \forall x, y \in \Sigma_K,
\end{array}$$

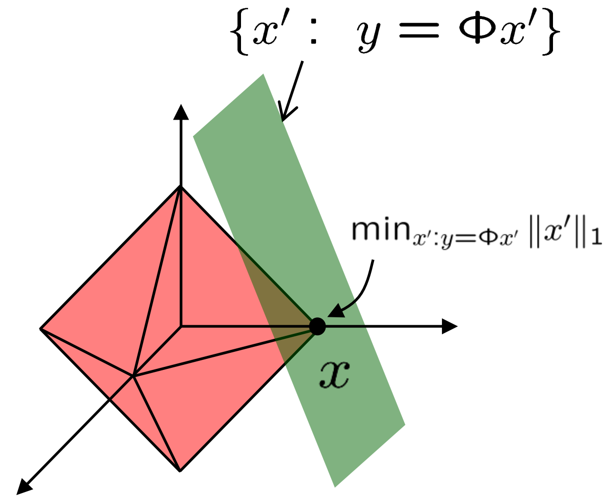
Restricted Isometry Property

- **Model:** *K*-sparse coefficients

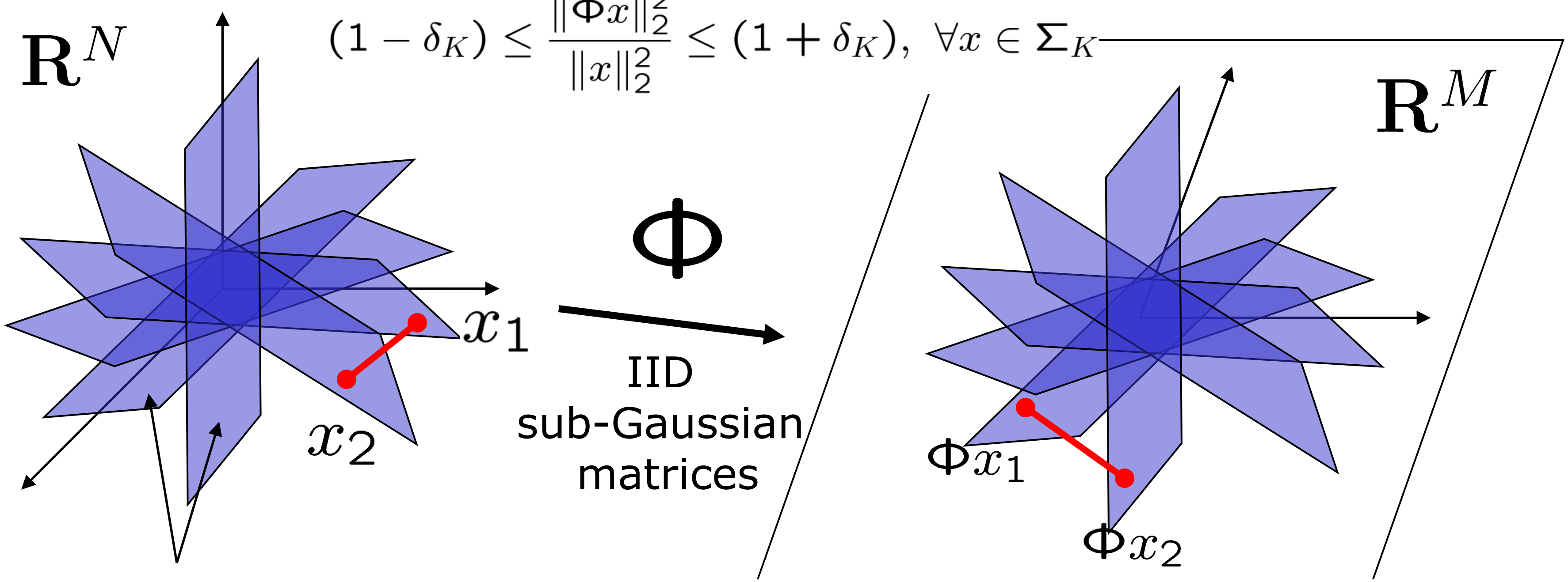
  Remark: implies convergence of convex relaxations also

  e.g., $\delta_{2K} < .465$ is sufficient for BP

- **RIP:** stable embedding



$$(1 - \delta_K) \leq \frac{\|\Phi x\|_2^2}{\|x\|_2^2} \leq (1 + \delta_K), \;\; \forall x \in \Sigma_K$$



$$M = O(K \log(N/K))$$

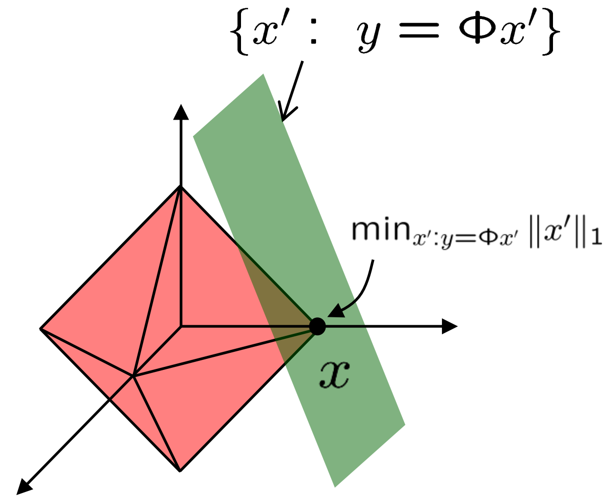
[Candes 2008; Baraniuk et al. 2008]

Restricted Isometry Property

- **Model:** *K*-sparse coefficients

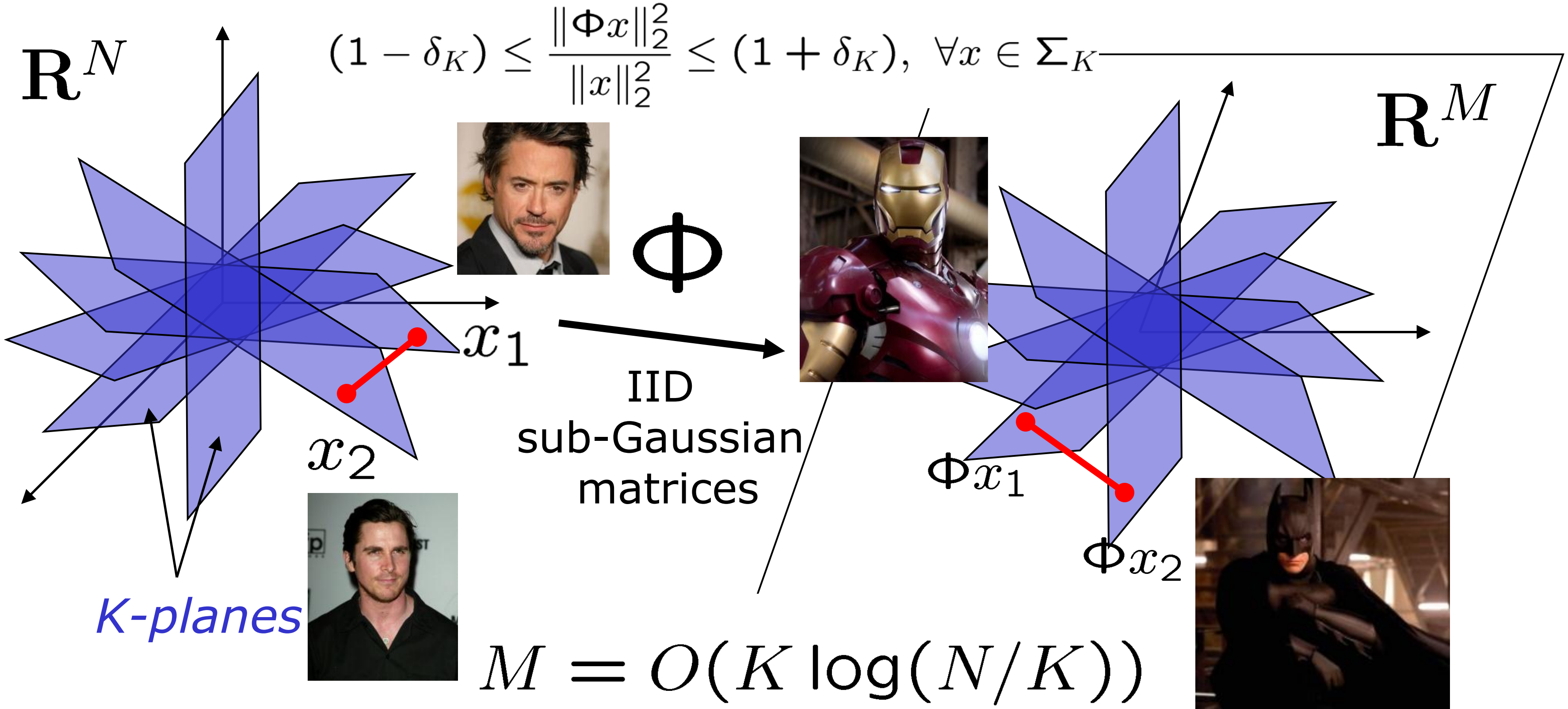
  Remark: implies convergence of convex relaxations also

  e.g., $\delta_{2K} < .465$ is sufficient for BP



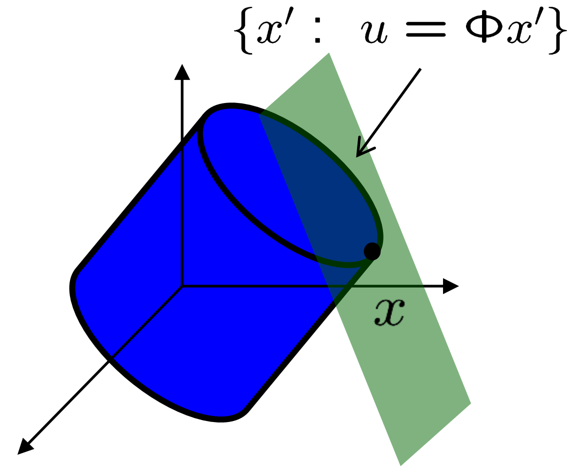
- **RIP:** stable embedding

$$(1-\delta_K) \leq \frac{\|\Phi x\|_2^2}{\|x\|_2^2} \leq (1+\delta_K), \;\; \forall x \in \Sigma_K$$



$$M = O(K \log(N/K))$$

Restricted Isometry Property for Matrices!



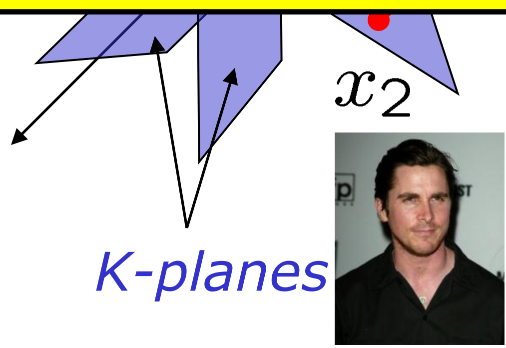
- **Model:** *rank-R matrices*

  Remark: bi-Lipschitz embedding of low-rank matrices

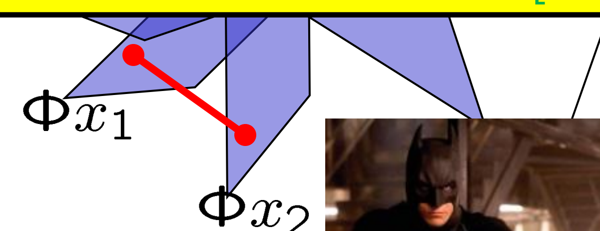
- **RIP:** stable embedding

$$(1-\delta_R) \leq \frac{\|\Phi X\|_F^2}{\|X\|_F^2} \leq (1+\delta_R), \; \forall X : \text{rank}(X) \leq R$$

[Plan 2011]



IID sub-Gaussian matrices



$$M = O(R(2N-R))$$

Projected gradient method for non-convex sets

- Model-based hard thresholding $f(x) = \|u - \Phi x\|^2$

$\min_{x : x \in \Sigma_{\mathcal{M}_K}} f(x)$

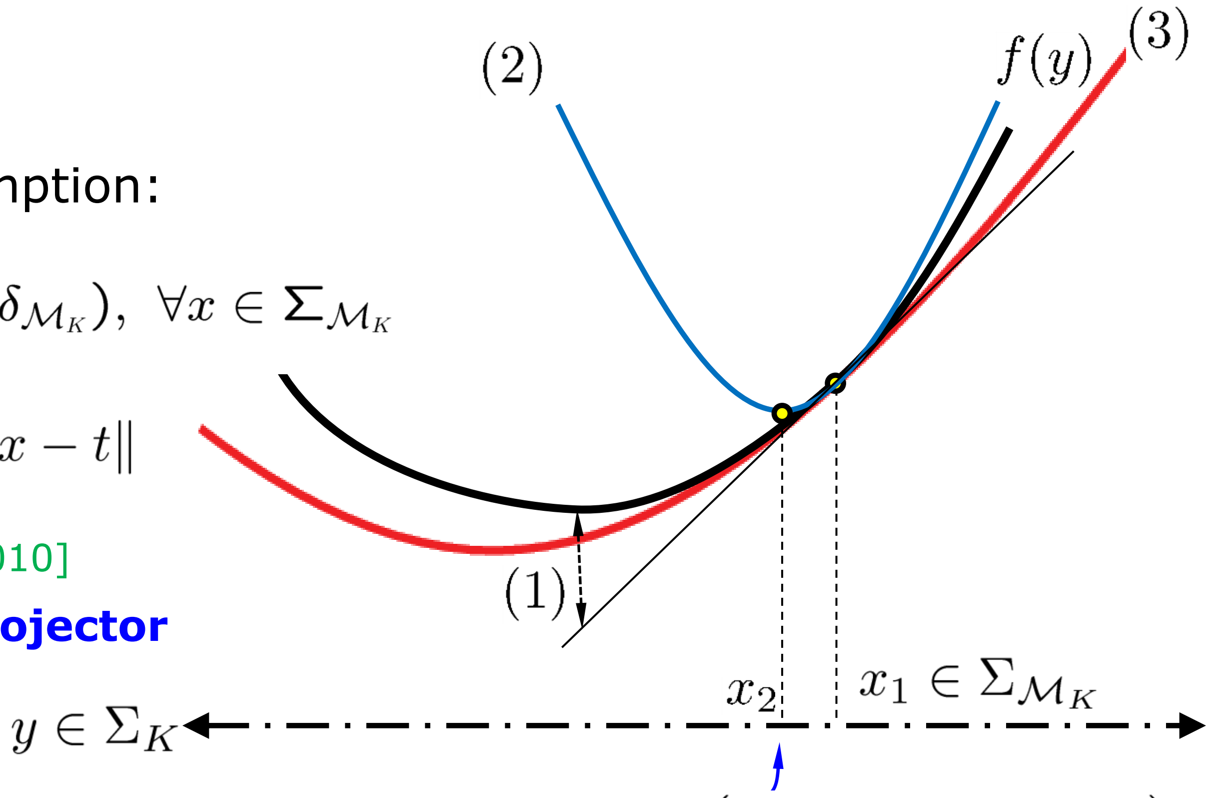
Global "unverifiable" assumption:

$$(1 - \delta_{\mathcal{M}_K}) \leq \frac{\|\Phi x\|_2^2}{\|x\|_2^2} \leq (1 + \delta_{\mathcal{M}_K}), \; \forall x \in \Sigma_{\mathcal{M}_K}$$

$$\mathrm{H}_{\Sigma_{\mathcal{M}_K}}(t) = \arg \min_{x : x \in \Sigma_{\mathcal{M}_K}} \|x - t\|$$

[Baraniuk, C, Duarte, Hegde 2010]

Key actor: non-convex projector

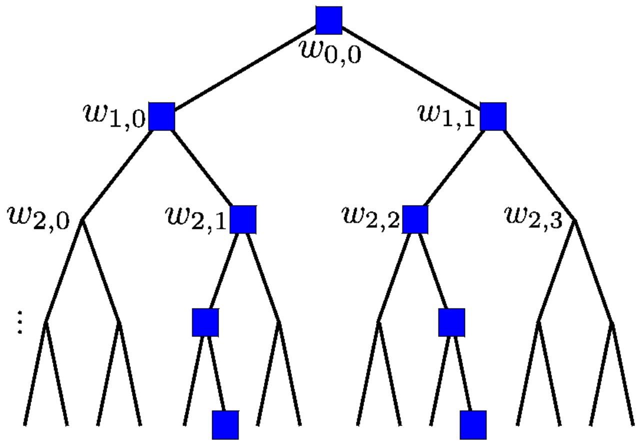


$\delta_{\mathcal{M}_{2K}} < 1/3$

$$x_{i+1} = \mathrm{H}_{\Sigma_{\mathcal{M}_K}}\left(x_i - \frac{1}{L_{\mathcal{M}_{2K}}} \nabla f(x_i)\right)$$

$$\begin{array}{llcll}
(1) & f(y) - f(x) - \langle \nabla f(x), y - x \rangle & = & \|\Phi(y - x)\|^2 & \forall x, y \in \mathcal{R}^N, \\
(2) & f(y) - f(x) - \langle \nabla f(x), y - x \rangle & \leq & \frac{L_{\mathcal{M}_{2K}}}{2}\|y - x\|^2 & L_{\mathcal{M}_{2K}} = 2(1 + \delta_{\mathcal{M}_{2K}}), \forall x, y \in \Sigma_{\mathcal{M}_{2K}}, \\
(3) & f(y) - f(x) - \langle \nabla f(x), y - x \rangle & \geq & \frac{\mu_{\mathcal{M}_{2K}}}{2}\|y - x\|^2 & \mu_{\mathcal{M}_{2K}} = 2(1 - \delta_{\mathcal{M}_{2K}}), \forall x, y \in \Sigma_{\mathcal{M}_{2K}},
\end{array}$$

Example: tree-sparse recovery



- **Model:** *K*-sparse coefficients
  + significant coefficients lie on a rooted subtree

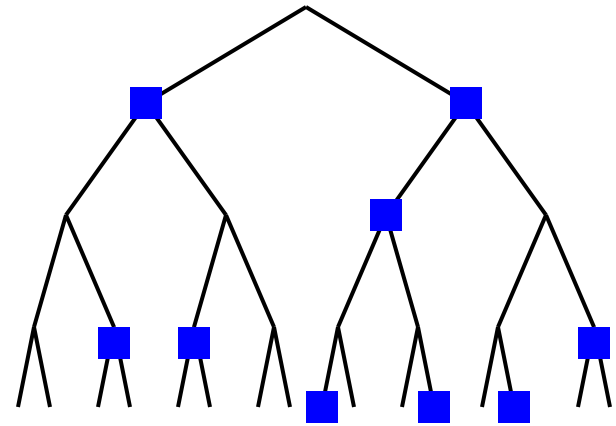
- **Sparse approx:** find best set of coefficients
  - sorting
  - hard thresholding

- **Tree-sparse approx:** find best rooted subtree of coefficients
  - condensing sort and select [Baraniuk]
  - dynamic programming [Donoho]

[Baraniuk, C, Duarte, Hegde 2010]

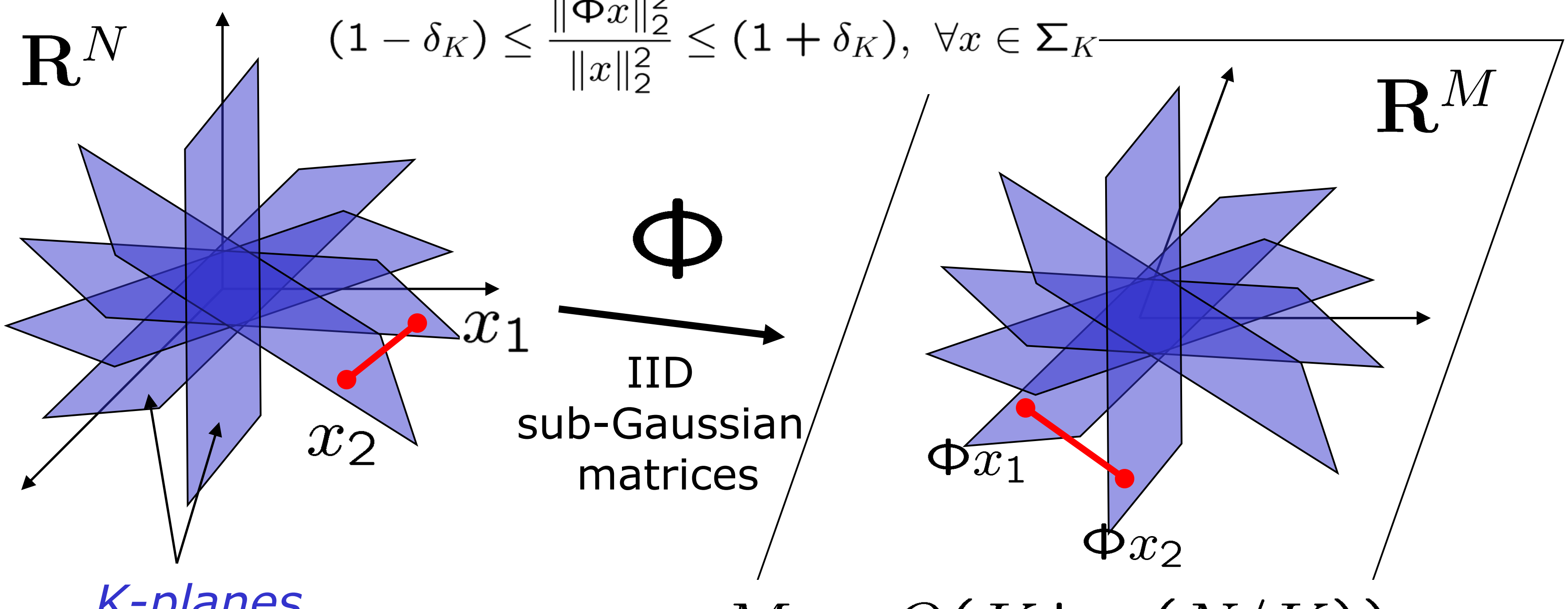
Example: tree-sparse recovery

- **Model:** *K*-sparse coefficients



- **RIP:** stable embedding

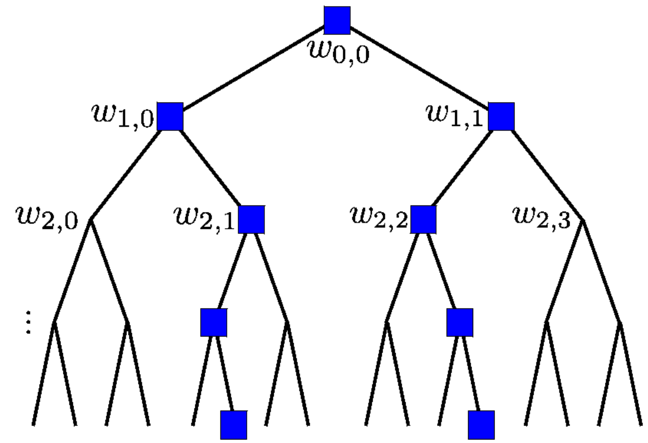
$$(1-\delta_K) \le \frac{\|\Phi x\|_2^2}{\|x\|_2^2} \le (1+\delta_K), \;\; \forall x \in \Sigma_K$$



$$M = O(K \log(N/K))$$

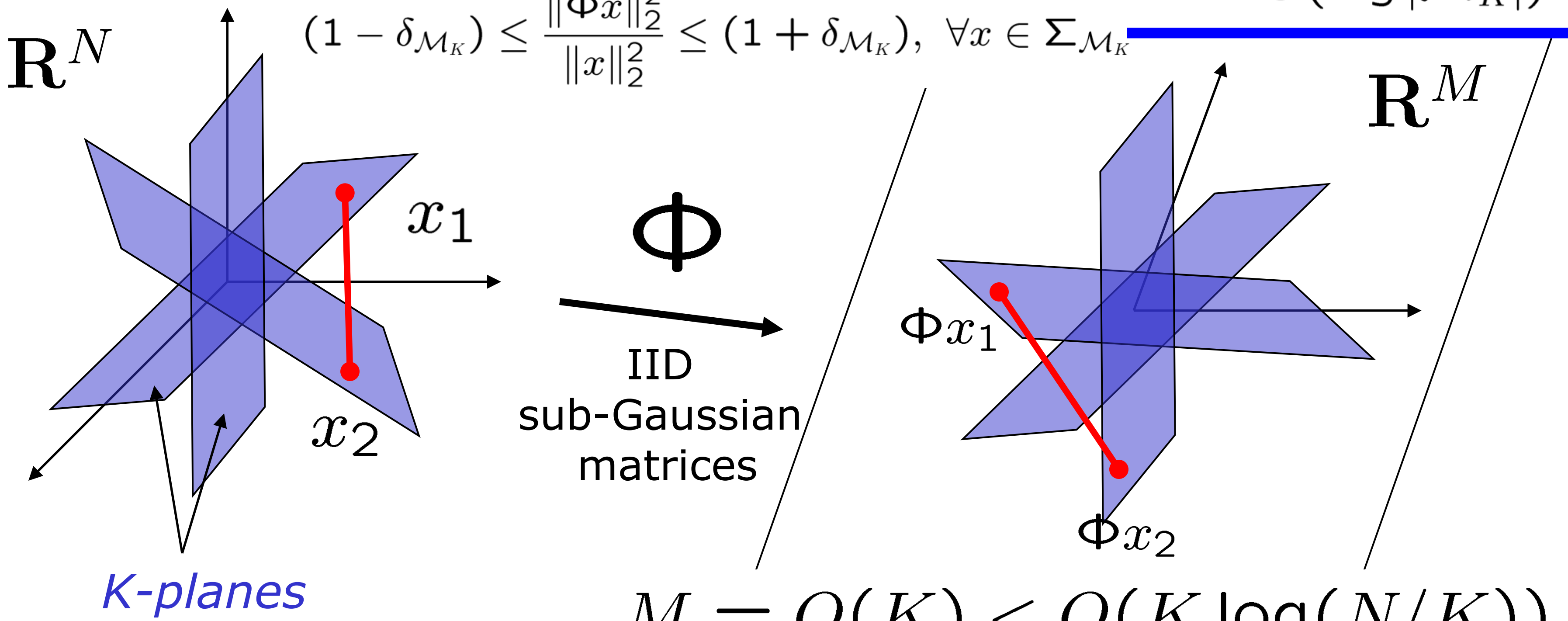
Example: tree-sparse recovery

- **Model:** K-sparse coefficients
  + significant coefficients lie on a rooted subtree
- **Tree-RIP:** stable embedding



$$M = O(\log|\mathcal{M}_K|)$$

$$(1-\delta_{\mathcal{M}_K}) \leq \frac{\|\Phi x\|_2^2}{\|x\|_2^2} \leq (1+\delta_{\mathcal{M}_K}), \ \forall x \in \Sigma_{\mathcal{M}_K}$$



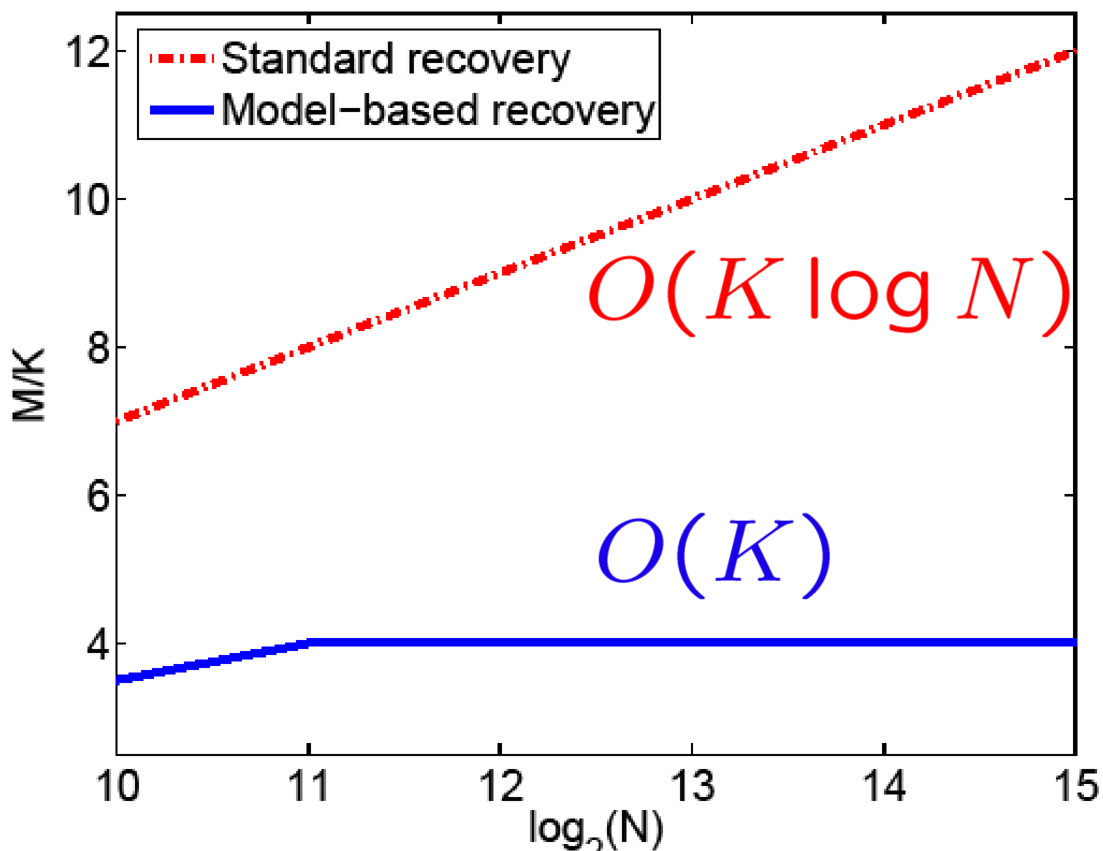
$$M = O(K) < O(K \log(N/K))$$

Example: tree-sparse recovery

- Number samples for correct recovery

- Piecewise cubic signals + wavelets

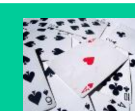
- Models/algorithms:
  - compressible (CoSaMP)
  - tree-compressible (tree-CoSaMP)



[Baraniuk, C, Duarte, Hegde 2010]

Recovery algorithms for low-dimensional models

The Clash Operator

	Non-convex $\binom{N}{K}$	Convex 	Probabilistic 
Encoding	combinatorial / manifolds	atomic norm / convex relaxation	compressible / sparse priors
Example	$\min_{x:\|x\|_0 \le K} \|u - \Phi x\|^2$	$\min_{x:\|x\|_1 \le \lambda} \|u - \Phi x\|^2$	$E\{x \mid u\}$
Algorithm	IHT, CoSaMP, SP, ALPS, OMP...	Basis pursuit, Lasso, basis pursuit denoising...	Variational Bayes, EP, Approximate message passing (AMP)...

$$\widehat{x}_{\text{Clash}} = \arg\min_{x:\|x\|_0 \le K, \|x\|_1 \le \lambda} \|u - \Phi x\|^2$$

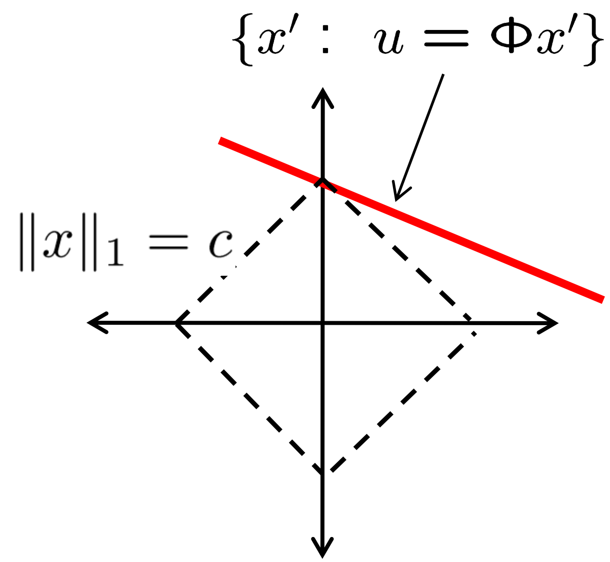
$$\|x\|_0 = \#\{x_i \neq 0\}$$

[Kyrillidis and C, 2011]

Recovery algorithms for low-dimensional models

$$\widehat{x} = \arg\min \|x\|_0 \text{ s.t. } u = \Phi x$$

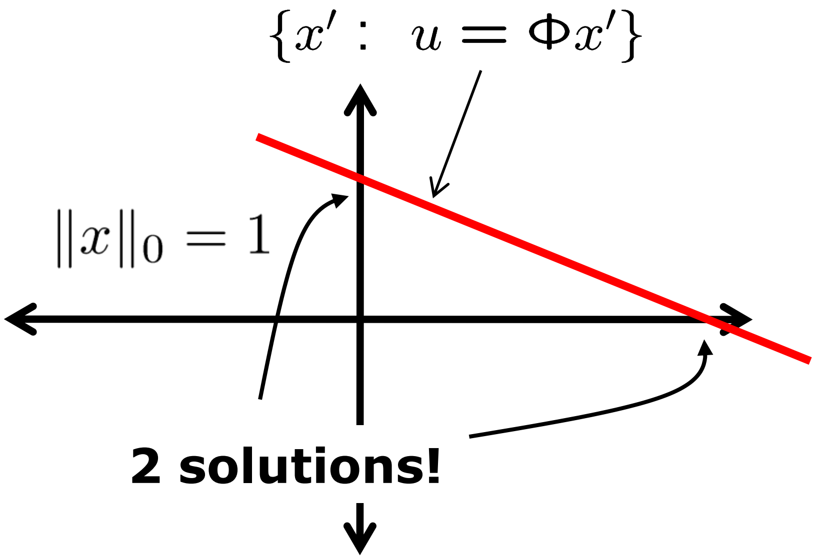
$$\widehat{x} = \arg\min \|x\|_1 \text{ s.t. } u = \Phi x$$



Recovery algorithms for low-dimensional models

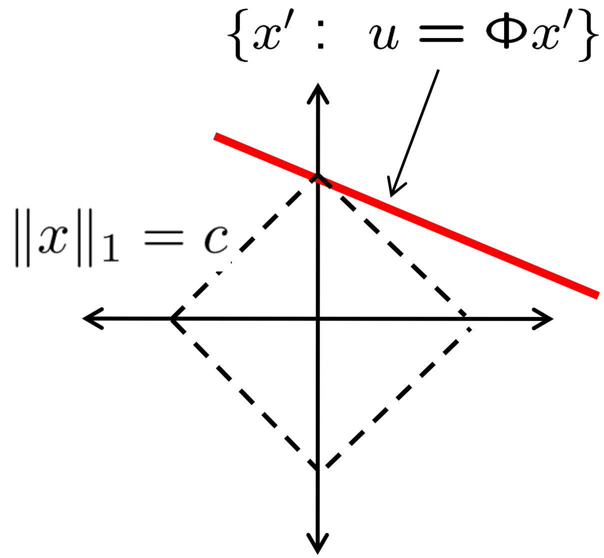
- A subtle issue

$$\widehat{x} = \arg\min \|x\|_0 \text{ s.t. } u = \Phi x$$



Which one is correct?

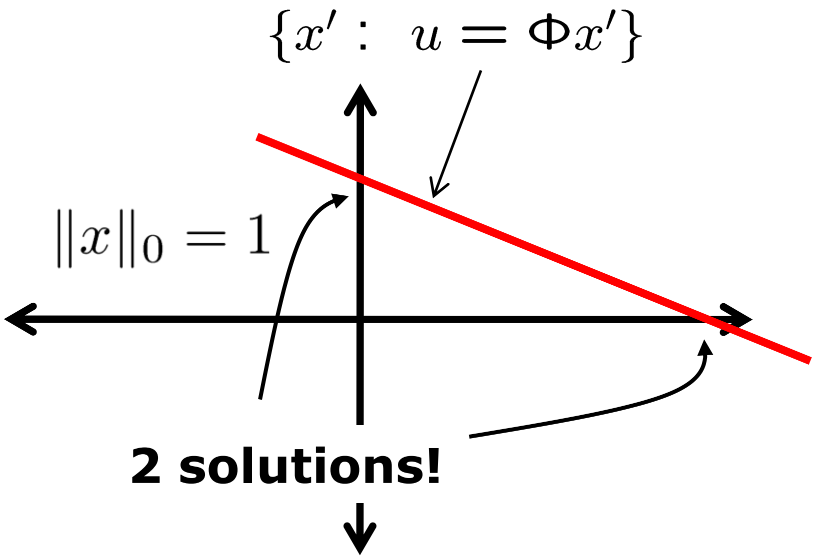
$$\widehat{x} = \arg\min \|x\|_1 \text{ s.t. } u = \Phi x$$



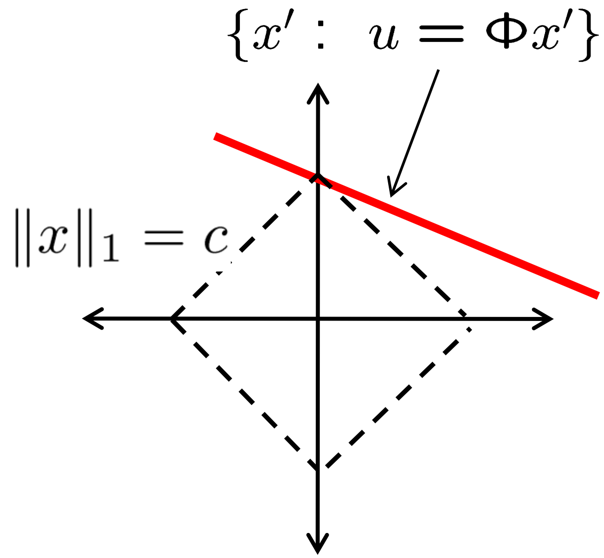
Recovery algorithms for low-dimensional models

- A subtle issue

$$\widehat{x} = \arg\min \|x\|_0 \text{ s.t. } u = \Phi x$$



$$\widehat{x} = \arg\min \|x\|_1 \text{ s.t. } u = \Phi x$$

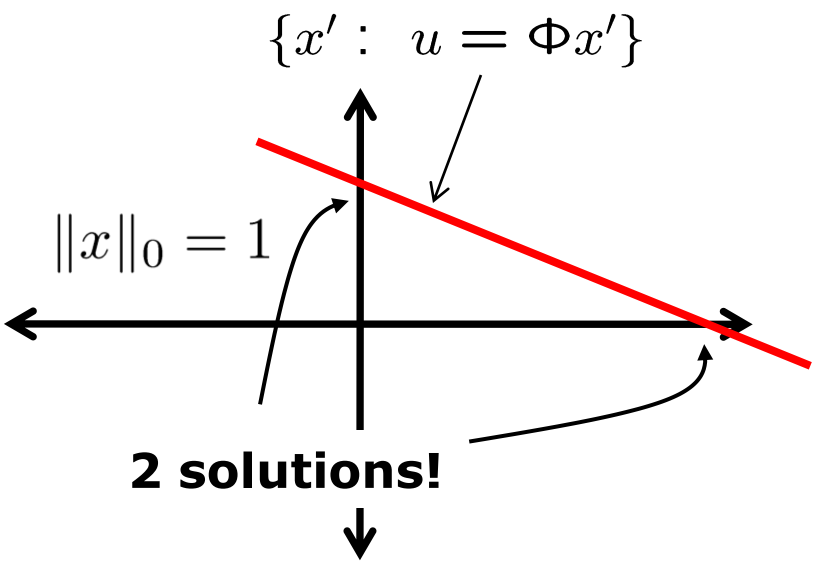


"Greed is good." – Joel Tropp 2004

Recovery algorithms for low-dimensional models

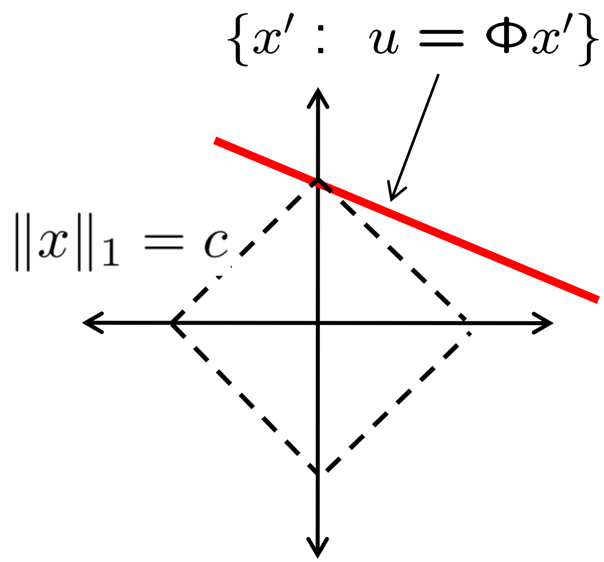
- A subtle issue

$$\widehat{x} = \arg\min \|x\|_0 \text{ s.t. } u = \Phi x$$



Which one is correct?

$$\widehat{x} = \arg\min \|x\|_1 \text{ s.t. } u = \Phi x$$

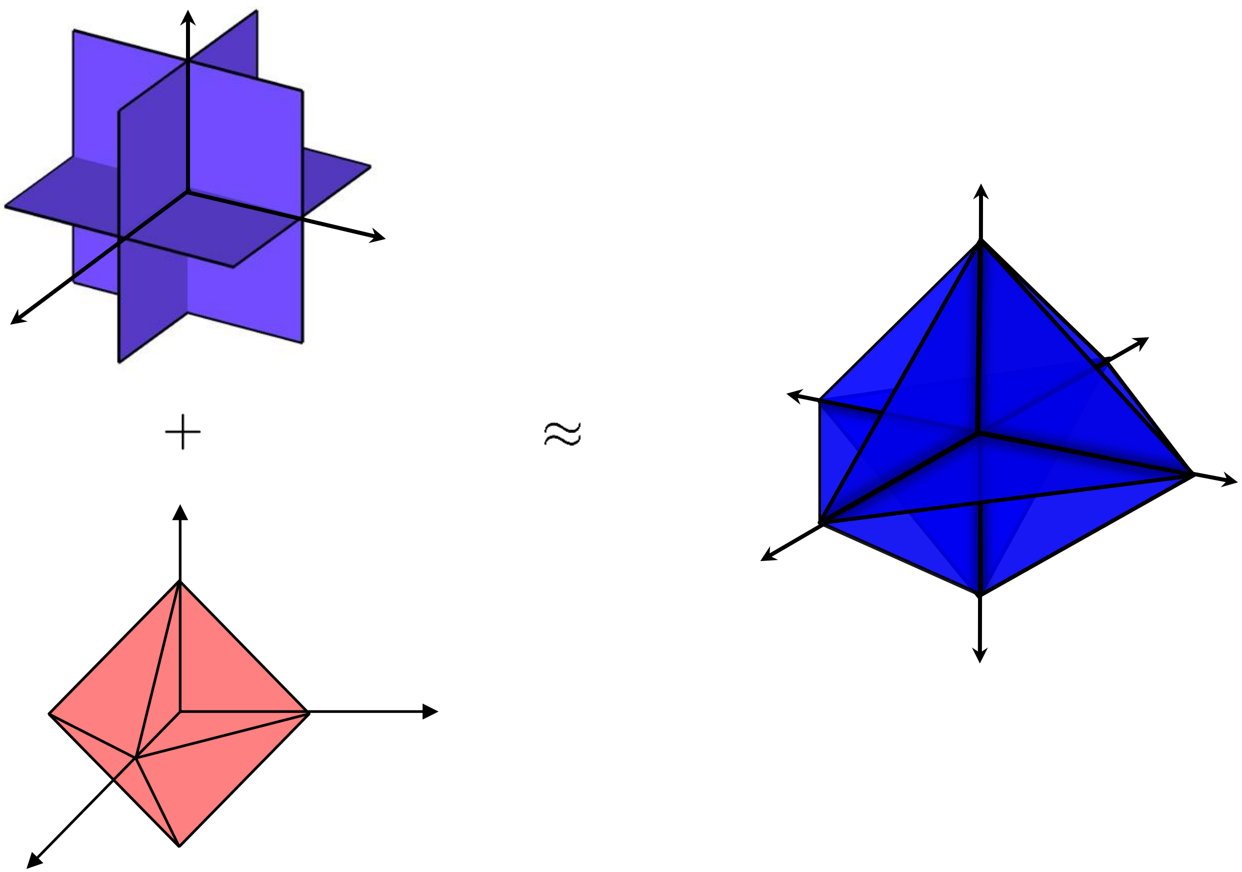


The CLASH algorithm

combinatorial selection
+
least absolute shrinkage

$$\mathrm{H}_{\{\|x\|_0 \le K\}}(t) = \arg \min_{\|x\|_0 \le K} \|x - t\|$$

$$\mathrm{St}_{\{\|x\|_1 \le \lambda\}}(t) = \arg \min_{\|x\|_1 \le \lambda} \|x - t\|$$



The CLASH algorithm

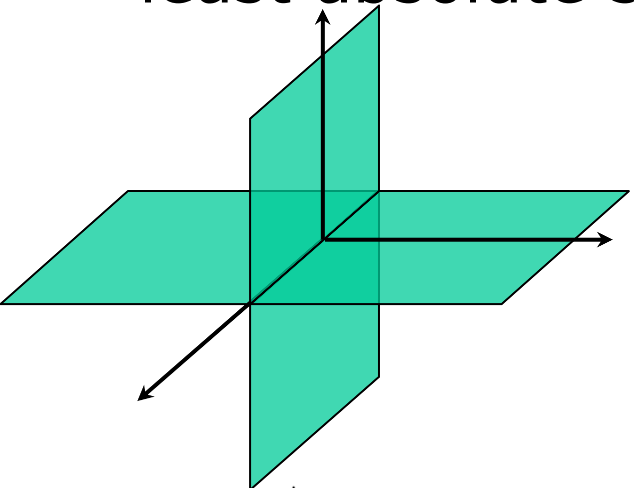
combinatorial selection

+

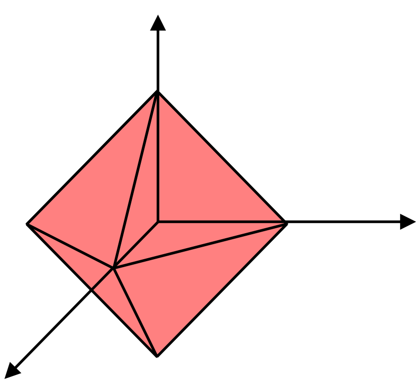
least absolute shrinkage

$$\mathrm{H}_{\Sigma_{\mathcal{M}_K}}(y) = \arg \min_{x : x \in \Sigma_{\mathcal{M}_K}} \|x - y\|$$

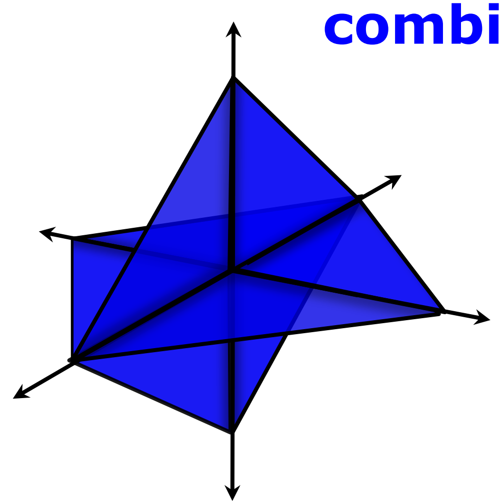
$$\mathrm{St}_{\{\|x\|_1 \leq \lambda\}}(t) = \arg \min_{\|x\|_1 \leq \lambda} \|x - t\|$$



$+$



$\approx$

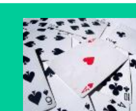


combinatorial origami

http://lions.epfl.ch/CLASH

Recovery algorithms for low-dimensional models

The Clash Operator

	Non-convex $\binom{N}{K}$	Convex 	Probabilistic 
Encoding	combinatorial / manifolds	atomic norm / convex relaxation	compressible / sparse priors
Example	$\min_{x:\|x\|_0\leq K} \|u-\Phi x\|^2$	$\min_{x:\|x\|_1\leq \lambda} \|u-\Phi x\|^2$	$E\{x\|u\}$
Algorithm	IHT, CoSaMP, SP, ALPS, OMP…	Basis pursuit, Lasso, basis pursuit denoising…	Variational Bayes, EP, Approximate message passing (AMP)…

$$\widehat{x}_{\text{Clash}} = \arg\min_{x:\|x\|_0\leq K, \|x\|_1\leq\lambda} \|u-\Phi x\|^2$$

The idea is much more general

$$\widehat{x}_{\text{Normed Pursuit}} = \arg\min_{x:\|x\|_0\leq K, \|x\|_*\leq\lambda} \|u-\Phi x\|^2$$

$$\|x\|_0 = \#\{x_i \neq 0\}$$

[Kyrillidis, Puy, and C, 2012]

Recovery algorithms for low-dimensional models

- Using projected gradient with exact non-convex projections
  with RIP/ERC/URC/RSC…
- **Exact low-dimensional model**
  - noise-free measurements: exact recovery
  - noisy measurements: stable recovery
- **Approximately low-dimensional model**
  - recovery as good as *K*-model-sparse approximation

$$\underbrace{\|x - \widehat{x}\|_{\ell_2}}_{\text{recovery error}} \leq C_1 \log\left(\frac{N}{K}\right) \underbrace{\frac{\|x - x_{\mathcal{M}_K}\|_{\ell_1}}{K^{1/2}}}_{\text{signal } K\text{-term model approx error}} + \underbrace{C_2 \epsilon}_{\text{noise}}$$

[Baraniuk, C, Duarte, Hegde 2010]

Recovery algorithms for low-dimensional models

- Using projected gradient with exact non-convex projections
  with RIP/ERC/URC/RSC…
- **Exact low-dimensional model**
  - noise-free measurements: exact recovery
  - noisy measurements: stable recovery
- **Approximately low-dimensional model**
  - recovery as good as *K*-model-sparse approximation

$$\|x - \widehat{x}\|_{\ell_2} \leq C_1 \log\left(\frac{N}{K}\right) \frac{\|x - x_{\mathcal{M}_K}\|_{\ell_1}}{K^{1/2}} + C_2 \epsilon$$

recovery error — signal *K*-term model approx error — noise

  - the bound remains qualitatively the same for other models!!!

Recovery algorithms for low-dimensional models

- Projected gradient with (non)exact non-convex projections
  without RIP/ERC/URC/RSC…

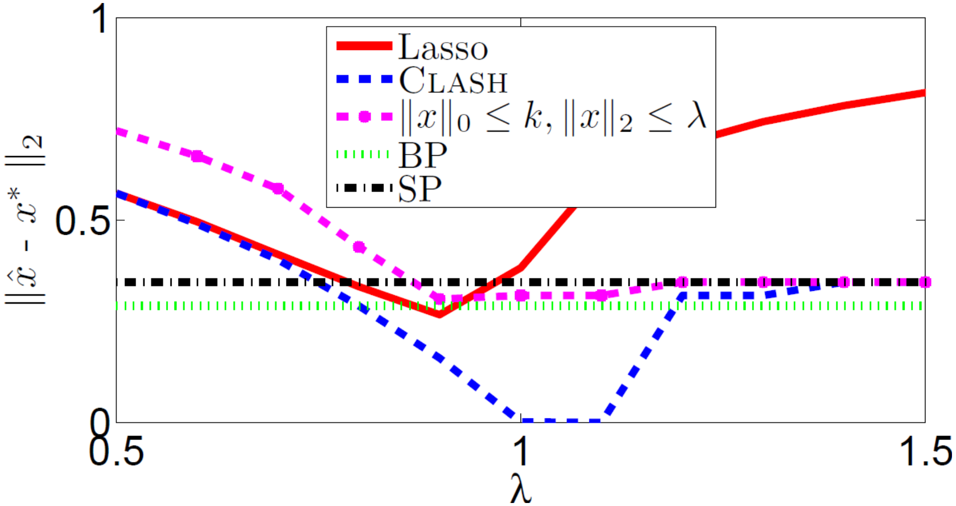
- **Not much!**
  - convergence to stationary point with ***Kurdyka-Lojasiewicz***

    [Attouch et al., 2010]

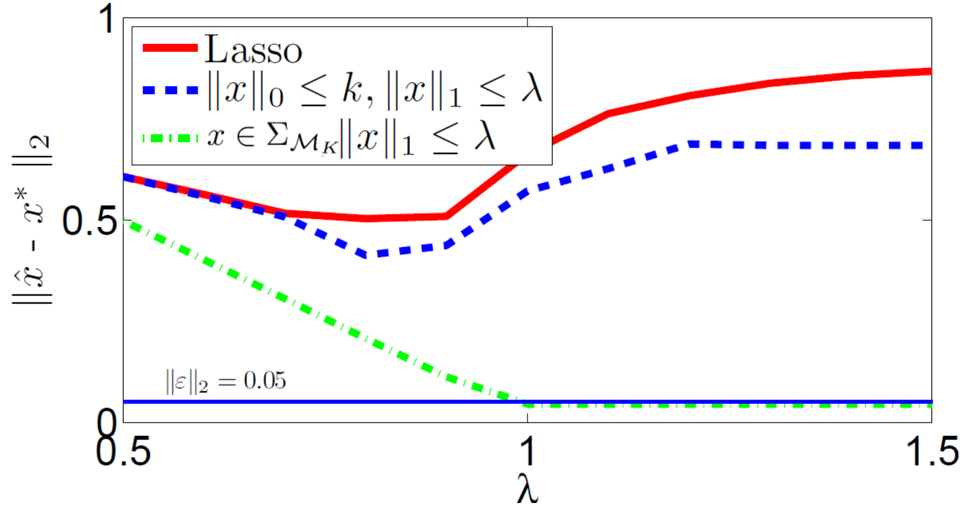
Examples

$$\widehat{x} = \arg \min_{x:\mathrm{supp}(x)\in\Sigma_{\mathcal{M}_K}, \|x\|_* \leq \lambda} \|u - \Phi x\|^2$$

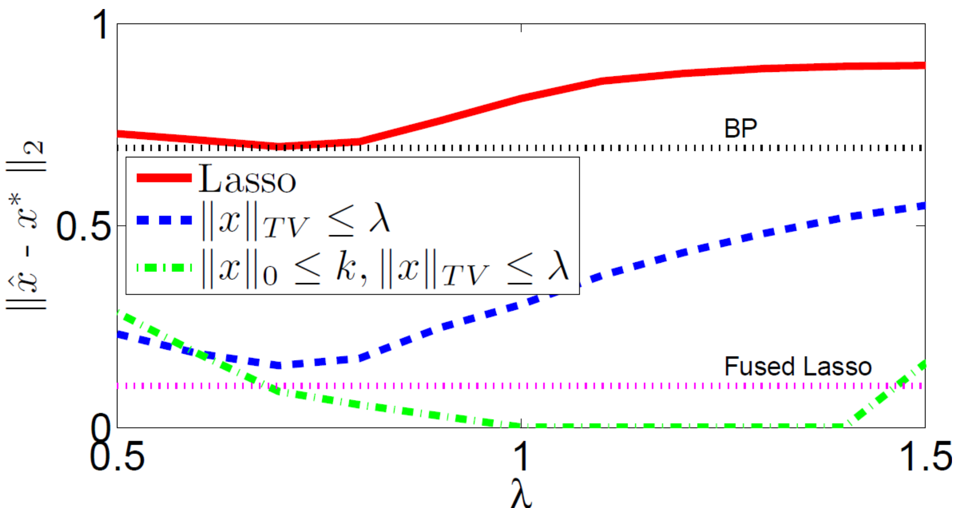
CLASH



Structured Sparsity



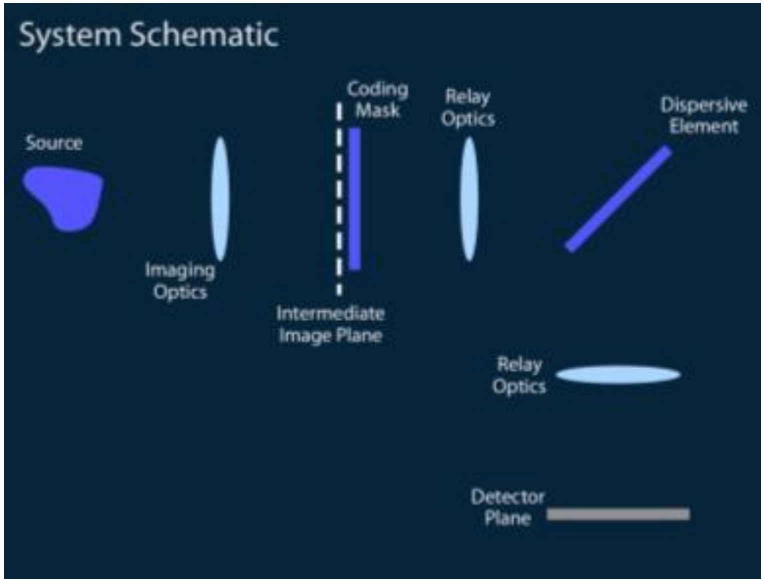
Norm Constraints

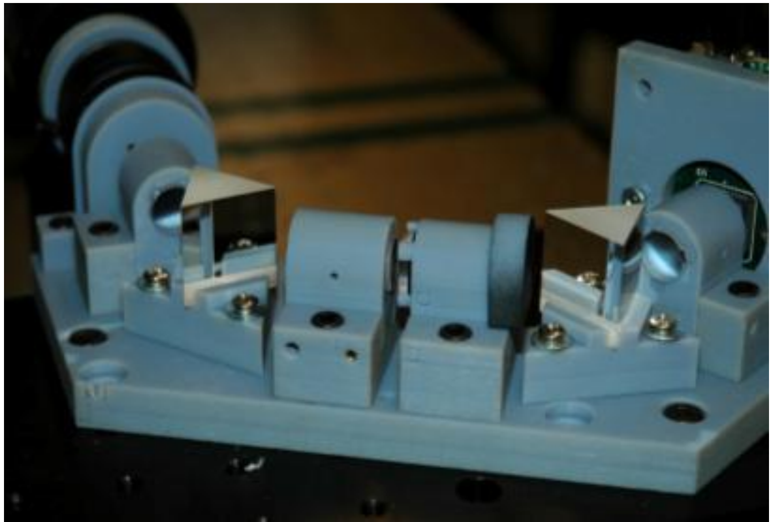


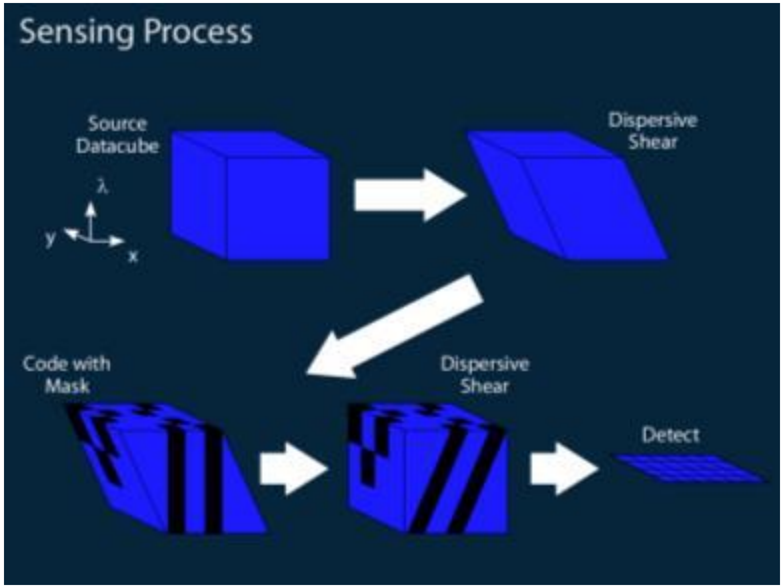
[Kyrillidis, Puy, and C, 2012]

Examples

Coded Aperture Snapshot Spectral Imager



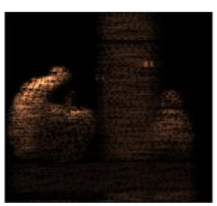


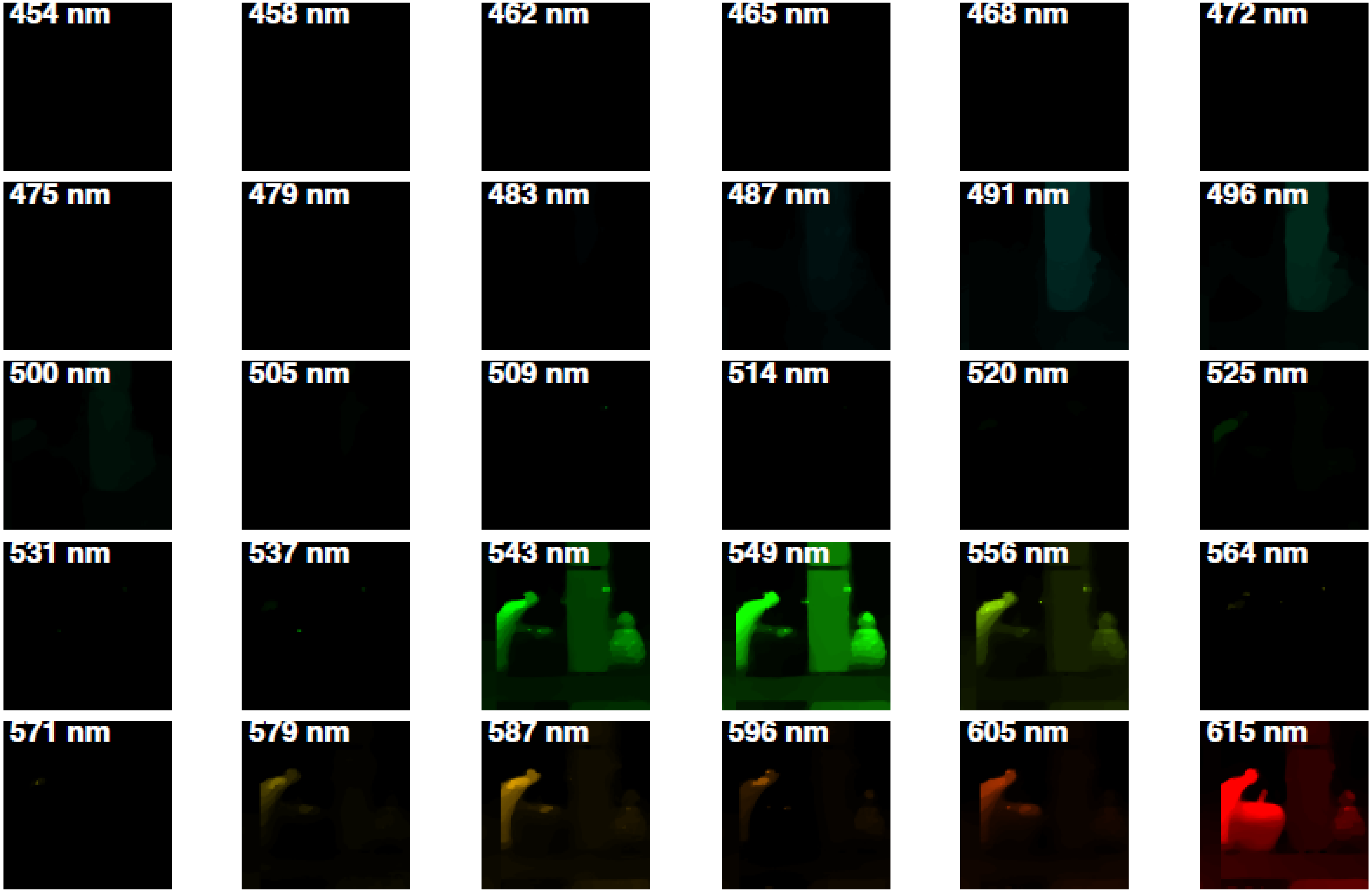


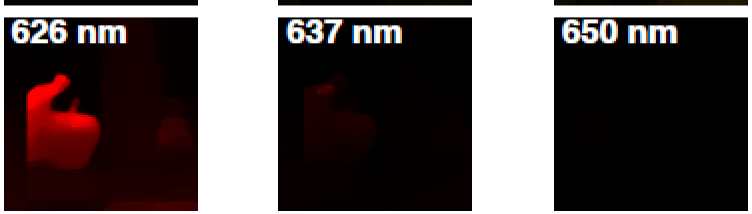
http://www.disp.duke.edu/projects/CASSI/

Examples

Total variation



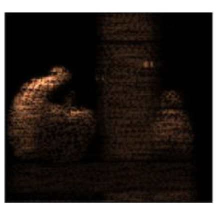




TWiST results are @
http://www.disp.duke.edu/projects/CASSI/

Examples

TV-CLASH



TV + sparsity in wavelets

[Kyrillidis, Puy, and C, 2012]

Acceleration of non-convex algorithms

- Several approaches

  step-size selection

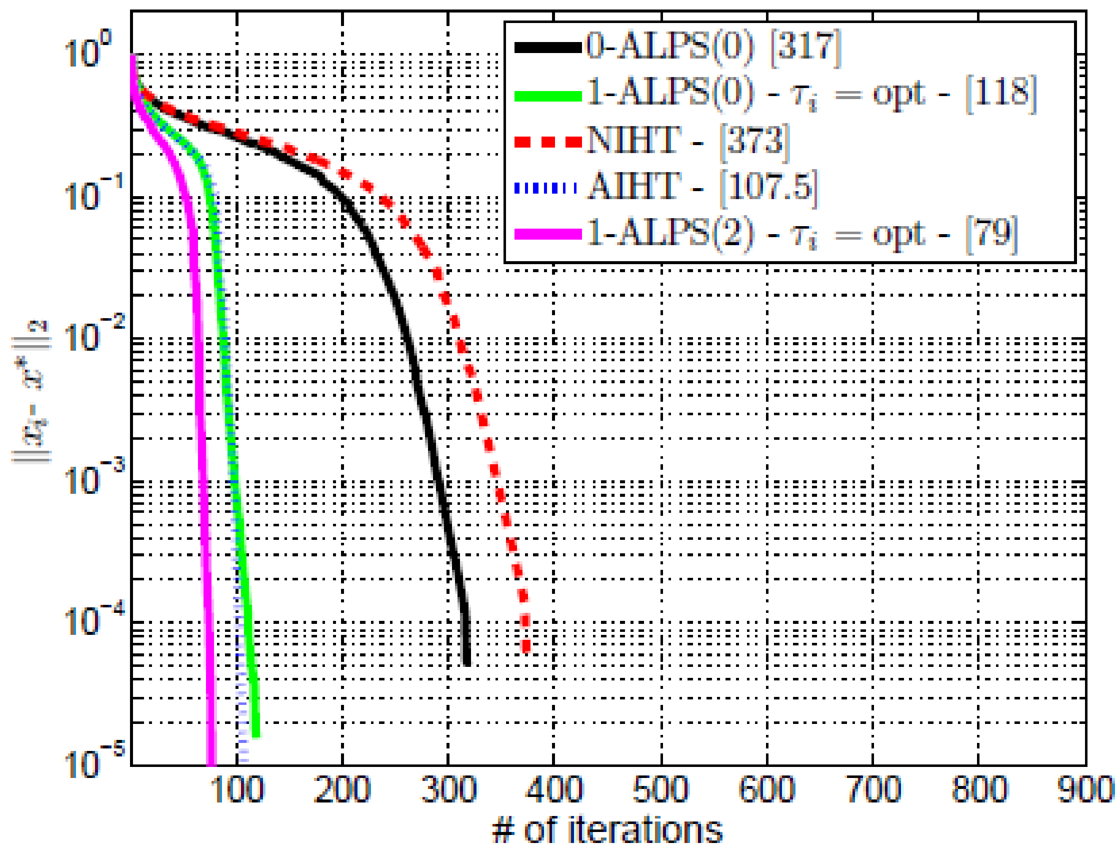
  memory based methods similar to Nesterov acceleration / double overrelaxation

  non-convex splitting

  (adaptive) block coordinate descent

  epsilon-approximate projections

$$x_{i+1} = \mathrm{H}_{\Sigma_{\mathcal{M}_K}}\left(y_i - \mu_i \nabla f(y_i)\right)$$
$$y_{i+1} = x_{i+1} + \tau_i(x_{i+1} - x_i)$$



[Kyrillidis and C, 2011]

Acceleration of non-convex algorithms

- Several approaches

  step-size selection

  memory based methods similar to Nesterov acceleration / double overrelaxation

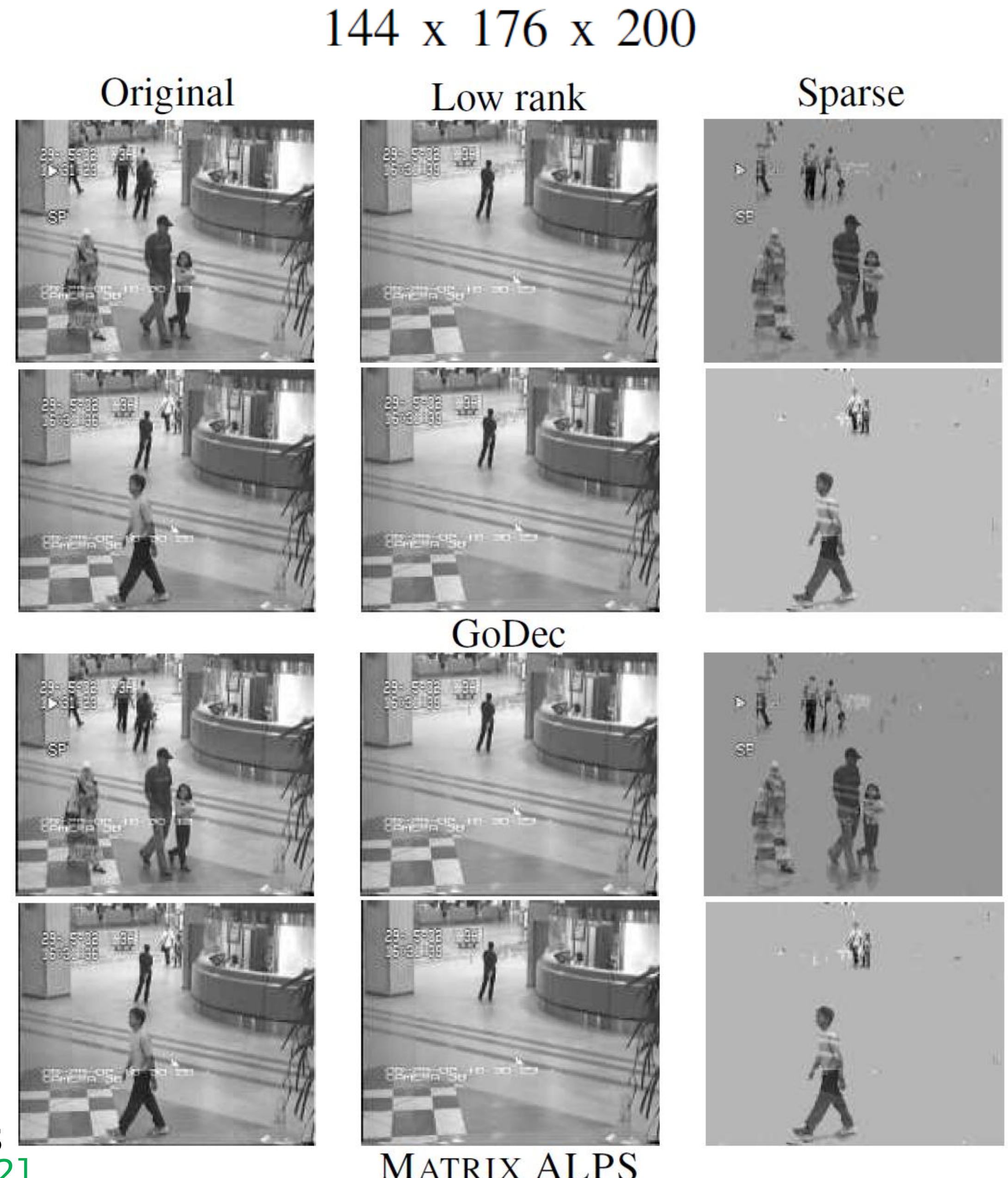
34.8s

  non-convex splitting

  (adaptive) block coordinate descent

  epsilon-approximate projections

15.8s

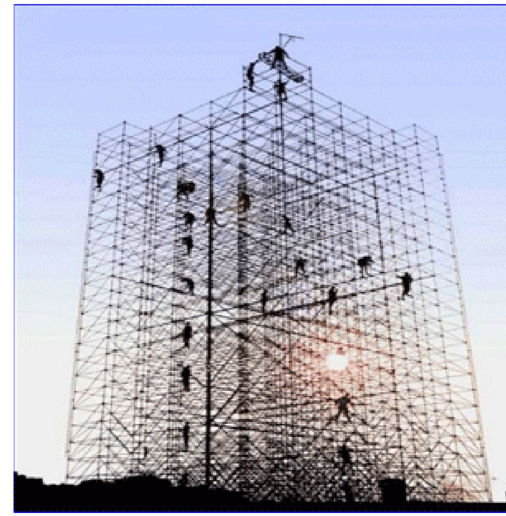


[Zhou and Tao 2011; Kyrillidis and C, 2012]

Final remarks

- non-convex algorithms <> low-dimensional scaffold
  - possible performance gains
  - non-convexifiable priors
  - matching prox operator with optimal space/time bounds

    complexity of structured approximation



- non-convex algorithms vs. convex algorithms
  - no clear winner / scenario dependent
  - decades of research in both

References

M. Afonso, J. Bioucas-Dias, M. Figueiredo, "Fast image recovery using variable splitting and constrained optimization", *IEEE Transactions on Image Processing, vol. 19, 2010.*

H. Attouch, J. Bolte, P. Redont, A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Lojasiewicz inequality", *Math. Oper. Research*, 2010

O. Axelsson*, Iterative Solution Methods,* Cambridge University Press, 1996.

R. Baraniuk, V. Cevher, M. Duarte, C. Hegde, "Model-based compressive sensing", *IEEE Transactions on Information Theory*, vol. 56, 2010.

R. Baraniuk, M. Davenport, R. de Vore, M. Wakin, "A Simple Proof of the Restricted Isometry Property for Random Matrices", *Constructive Approximation*, 2008.

R. Baraniuk, V. Cevher, M. Wakin, "Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective", *Proceedings of the IEEE*, vol. 98, 2010.

J. Barzilai and J. Borwein, "Two point step size gradient methods," *IMA Journal of Numer. Anal., vol. 8, 1988.*
R. Basri, D. Jacobs, "Lambertian Reflectance and Linear Subspaces", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, 2003.

A. Beck, M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Science, vol. 2,* 2009.

J. Bioucas-Dias, M. Figueiredo, "A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Transactions on Image Processing, vol. 16, 2007.*

P. Boufounos, R. Baraniuk, "1-bit compressed sensing", *Proceedings of the Conference on Information Science and Systems*, Princeton, 2008.

C. Boutsidis, M. Mahoney, P. Drineas, "An improved approximation algorithm for the column subset selection problem", *Proc. 20th Annual ACM/SIAM Symposium on Discrete Algorithms,* New York*, NY,* 2008.

References

P. Bühlmann, S. van der Geer, *Statistics for High-Dimensional Data,* Springer, 2011.

E. Candès, "The restricted isometry property and its implications for compressed sensing", *Comptes Rendus Mathematique,* vol. 346, 2008.

E. Candès and B. Recht, "Exact matrix completion via convex optimization", *Foundations of Computational Mathematics*, vol. 9, 2009.

L. Carin, R. Baraniuk, V. Cevher, D. Dunson, M. Jordan, G. Sapiro, M. Wakin, "Learning low-dimensional signal models", *IEEE Signal Processing Magazine*, vol. 28, 2010.

V. Cevher, "An ALPS view of sparse recovery", *Proc. ICASSP,* 2011.

V. Chandrasekaran, B. Recht, P. Parrilo, A. Willsky, "The convex geometry of linear inverse problems", submitted, 2010.

R. Chartrand, W. Yin, "Iteratively reweighted algorithms for compressive sensing", *Proc. ICASSP*, 2008

S. Chen, D. Donoho, M. Saunders, "Atomic decomposition by Basis Pursuit", *SIAM Review*, vol. 43, 2001.

P. Combettes, V. Wajs, "Signal recovery by proximal forward-backward splitting", *SIAM Journal Multiscale Modeling and Simulation, vol. 4,* 2005.

J. Eckstein, D. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators", *Mathematical Programming, vol. 5, 1992.*

M. Figueiredo and J. Bioucas-Dias, "Restoration of Poissonian images using alternating direction optimization", *IEEE Transactions on Image Processing,* vol. 19, 2010.

M. Figueiredo, R. Nowak, S. Wright, "Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems", *IEEE Journal of Selected Topics in Signal Processing,* vol. 1, 2007.

References

D. Gabay, B. Mercier, “A dual algorithm for the solution of nonlinear variational problems via finite-element approximations”, *Computers and Mathematics with Application, vol. 2, 1976.*

R. Glowinski, A. Marroco, “Sur l’approximation, par elements finis d’ordre un, et la resolution, par penalisation-dualité d’une classe de problemes de Dirichlet non lineares,” *Rev. Française d’Automatique, 1975.*

Y. Gordon, “On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$”, in *Geometric Aspects of Functional Analysis,* Springer, 1988.

E. Hale, W. Yin, Y. Zhang, “Fixed-point continuation for l1-minimization: Methodology and convergence”, *SIAM Journal on Optimization*, vol. 19, 2008.

N. Halko, P.-G. Martinsson, J. Tropp, “Finding structure with randomness: stochastic algorithms for constructing approximate matrix decompositions”, *SIAM Review*, vol. 53, 2011.

C. Hegde, M. Duarte, V. Cevher, “Compressive sensing recovery of spike trains using a structured sparsity model”, *Proceedings of SPARS'09*, Saint-Malo, France, 2009.

M. Hestenes, “Multiplier and gradient methods”, *Journal of Optimazion Theory andApplications, vol. 4, 1969.*

A. Kyrillidis, V. Cevher, “Recipes for hard thresholding methods”, Tech. Rep., EPFL, 2011.

J. Lee, V. Mirrokni, V. Nagarajan, M. Sviridenko, “Non-monotone submodular maximization under matroid and knapsack constraints”, *Proc. 41st Annual ACM Symposium on Theory of Computing, Bethesda, MD,* 2009.

A. Lewis, J. Malick, “Alternating projections on manifolds”, *Math. of Operations Research*, vol. 33, 2008.

D. Lorenz, “Constructing test instances for basis pursuit denoising”, submitted, 2011.

N. Meinshausen, P. Bühlmann, “High-dimensional graphs and variable selection with the lasso”, *The Annals of Statistics*, vol. 34, pp. 1436-1462, 2006.

J.-J. Moreau, “Proximité et dualité dans un espace hilbertien,” *Bull. Soc. Mathematiques de France, vol. 93, 1965.*

References

G. Nemhauser, L. Wolsey, *Integer and combinatorial optimization*, Wiley, 1988.

S. Osher, M. Burger, D. Goldfarb, J. Xu, W. Yin, “An iterative regularization method for total variation-based image restoration”, *SIAM Journal on Multiscale Modeling and Simulation*, vol. 4, 2005.

Y. Plan, “Compressed sensing, sparse approximation, low-rank matrix estimation”, PhD Thesis, Caltech, 2011

M. Powell, “A method for nonlinear constraints in minimization problems”, in *Optimization,* Academic Press*, 1969.*

H. Raguet, J. Fadili, G. Peyré, “Generalized Forward-Backward splitting”, Tech. report, Hal-00613637, 2011.

S. Setzer, G. Steidl, T. Teuber, “Deblurring Poissonian images by split Bregman techniques,” *Journal of Visual Communication and Image Representation, 2010.*

K.-C. Toh , S. Yun, “An accelerated proximal gradient algorithm for nuclear norm regularized least squares Problems”, *Pacific Journal of Optimization*, vol. 6, 2010.

A. Waters, A. Sankaranarayanan, R. Baraniuk, ”SpaRCS: recovering low-rank and sparse matrices from compressive measurements”, *Neural Information Processing Systems,* 2011.

S. Wright, R. Nowak, M. Figueiredo, “Sparse reconstruction by separable approximation,” *IEEE Transactions on Signal Processing, vol. 57,* 2009.

W. Yin, S. Osher, D. Goldfarb, J. Darbon, “Bregman iterative algorithms for l1-minimization with applications to compressed sensing”, *SIAM Journal on Imaging Science, vol. 1, 2008.*

T. Zhou, D. Tao, “Godec: randomized low-rank & sparse matrix decomposition in noisy case,” *Proc. International Conference on Machine Learning*, Bellevue, WA, 2011.