The Abdus Salam
International Centre
for Theoretical Physics
50th Anniversary 1964-2014

2585-27

Joint ICTP-TWAS School on Coherent State Transforms, Time-Frequency and Time-Scale Analysis, Applications

2 - 20 June 2014

Numerical algorithms for sparse recovery

I. Loris
ULB, Brussels
Belgium

Numerical algorithms for sparse recovery (part 1)

Ignace Loris

Université Libre de Bruxelles

Coherent state transforms,
time-frequency and time-scale analysis, applications

Trieste, Italy, June 2–21, 2014

Theme: Sparse recovery

- What is it?

  "finding an answer without asking too many questions, knowing the answer is simple"

- In mathematical language:
  Solving an under-determined linear system

$$Ku = y \qquad (y, K \text{ known})$$

  for u,

  - when number of y_i is much smaller than number of u_i,
  - but knowing that many of the u_i are zero ("sparsity").

- Example: recovering an object (image) from incomplete measurements knowing that its wavelet transform is sparse

Mathematical framework

- Solve linear relations between unknown u and measurement data y:

$$Ku = y$$

- Here:
  - $y =$ data vector (known)
  - $K =$ linear operator(known)
  - $u =$ model vector (unknown)

- Problems: insufficient data, inconsistent (noisy) data, ill-conditioning of K:

$$\rightarrow \textit{No solution or no unique solution}$$

- Minimize a penalized least-squares functional:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \text{penalty}$$

NB: minimizer of $f(u)$ $\overset{\text{def.}}{\leftrightarrow}$ $\arg\min_u f(u)$

Penalization strategy

- How to enforce sparsity on solutions of $Ku = y$?
- Minimize a penalized least-squares functional:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \text{penalty}$$

  with a judiciously chosen 'sparsity promoting' penalty.
- A trade-off between sparsity promotion and tractability (convexity)
- See [6, 3]

Penalization strategy

- How to enforce sparsity on solutions of $Ku = y$?
- Minimize a penalized least-squares functional:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \text{penalty}$$

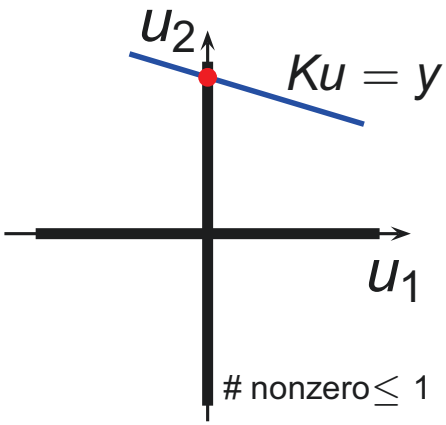
  with a judiciously chosen 'sparsity promoting' penalty.
- A trade-off between sparsity promotion and tractability (convexity)
- See [6, 3]

Penalization strategy

- How to enforce sparsity on solutions of $Ku = y$?
- Minimize a penalized least-squares functional:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \text{penalty}$$

  with a judiciously chosen 'sparsity promoting' penalty.
- A trade-off between sparsity promotion and tractability (convexity)



'combinatorial'

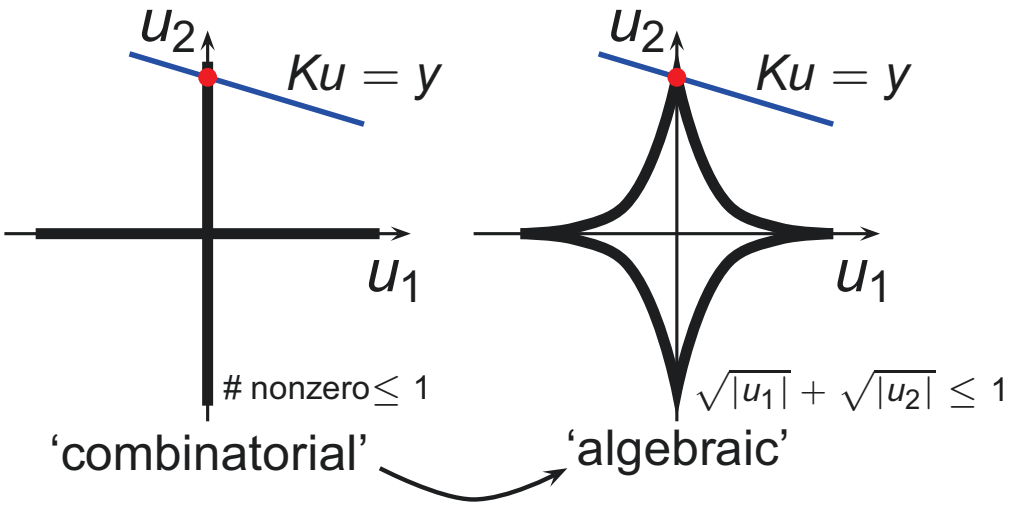
- See [6, 3]

Penalization strategy

- How to enforce sparsity on solutions of $Ku = y$?
- Minimize a penalized least-squares functional:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \text{penalty}$$

  with a judiciously chosen 'sparsity promoting' penalty.
- A trade-off between sparsity promotion and tractability (convexity)



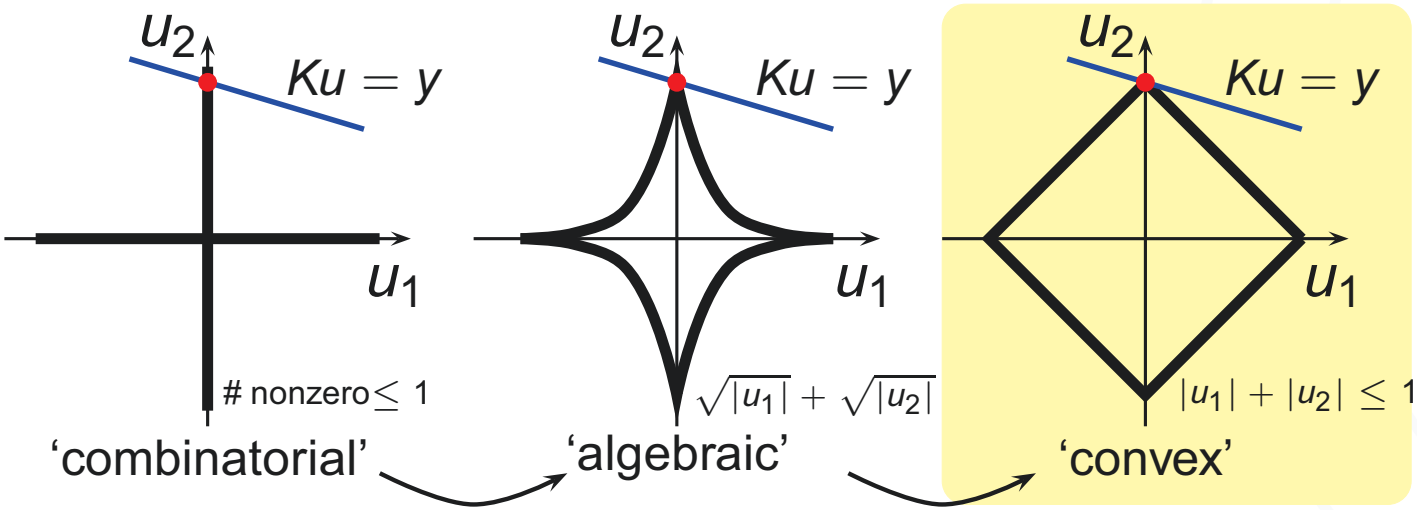
- See [6, 3]

Penalization strategy

- How to enforce sparsity on solutions of $Ku = y$?
- Minimize a penalized least-squares functional:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \text{penalty}$$

  with a judiciously chosen 'sparsity promoting' penalty.
- A trade-off between sparsity promotion and tractability (convexity)



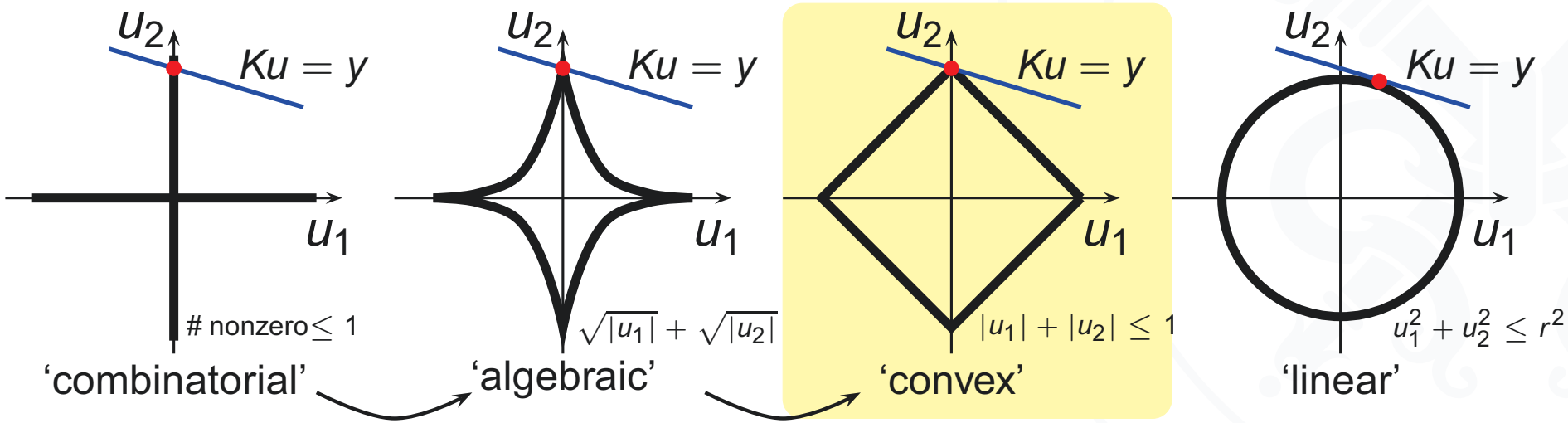
- See [6, 3]

Penalization strategy

- How to enforce sparsity on solutions of $Ku = y$?
- Minimize a penalized least-squares functional:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \text{penalty}$$

  with a judiciously chosen 'sparsity promoting' penalty.
- A trade-off between sparsity promotion and tractability (convexity)



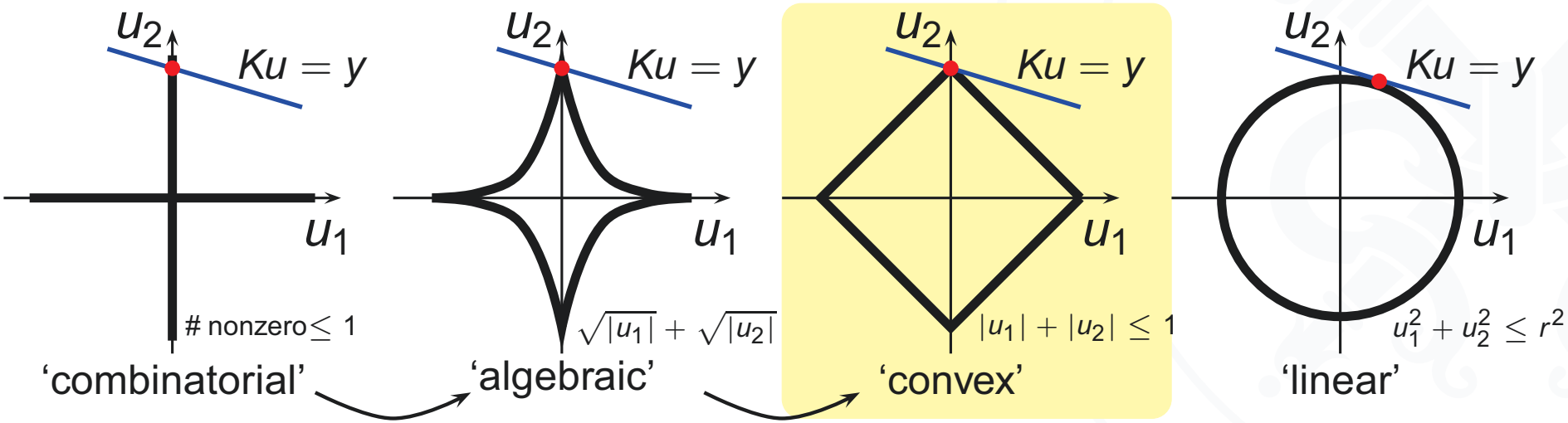
- See [6, 3]

Penalization strategy

- How to enforce sparsity on solutions of $Ku = y$?
- Minimize a penalized least-squares functional:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \text{penalty}$$

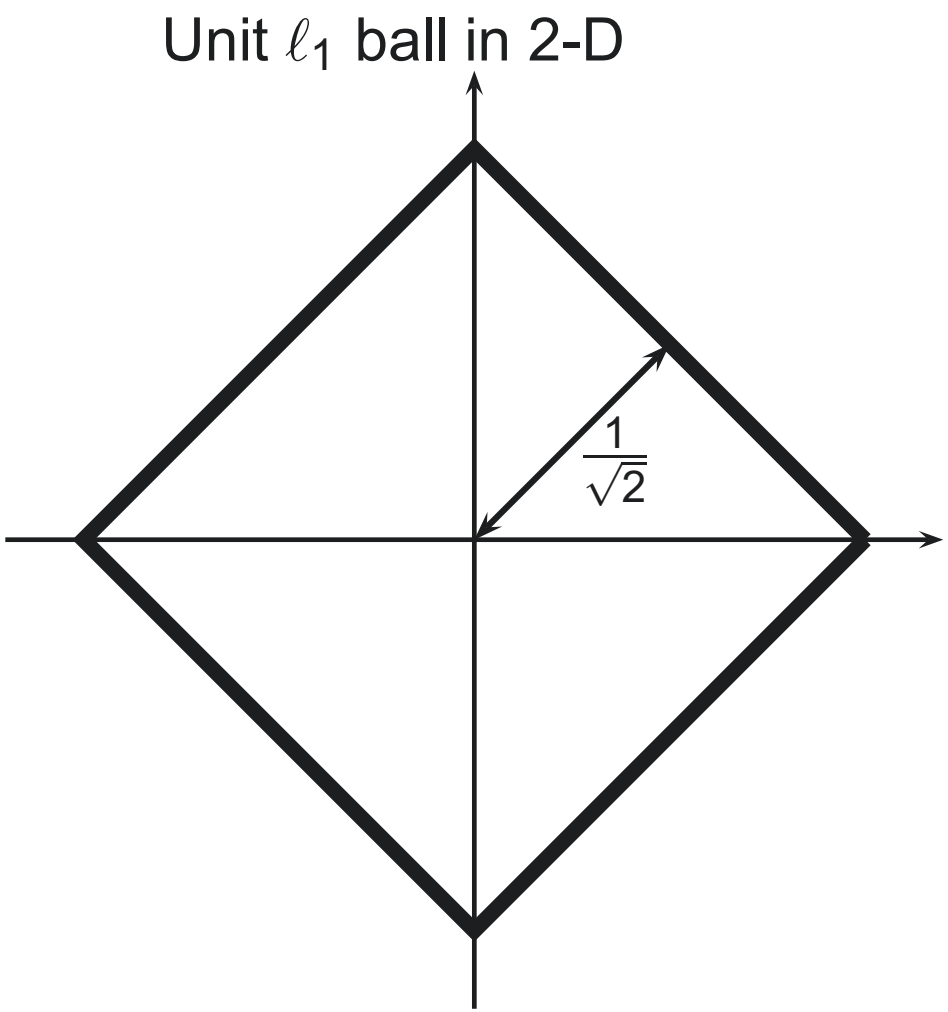
  with a judiciously chosen 'sparsity promoting' penalty.
- A trade-off between sparsity promotion and tractability (convexity)



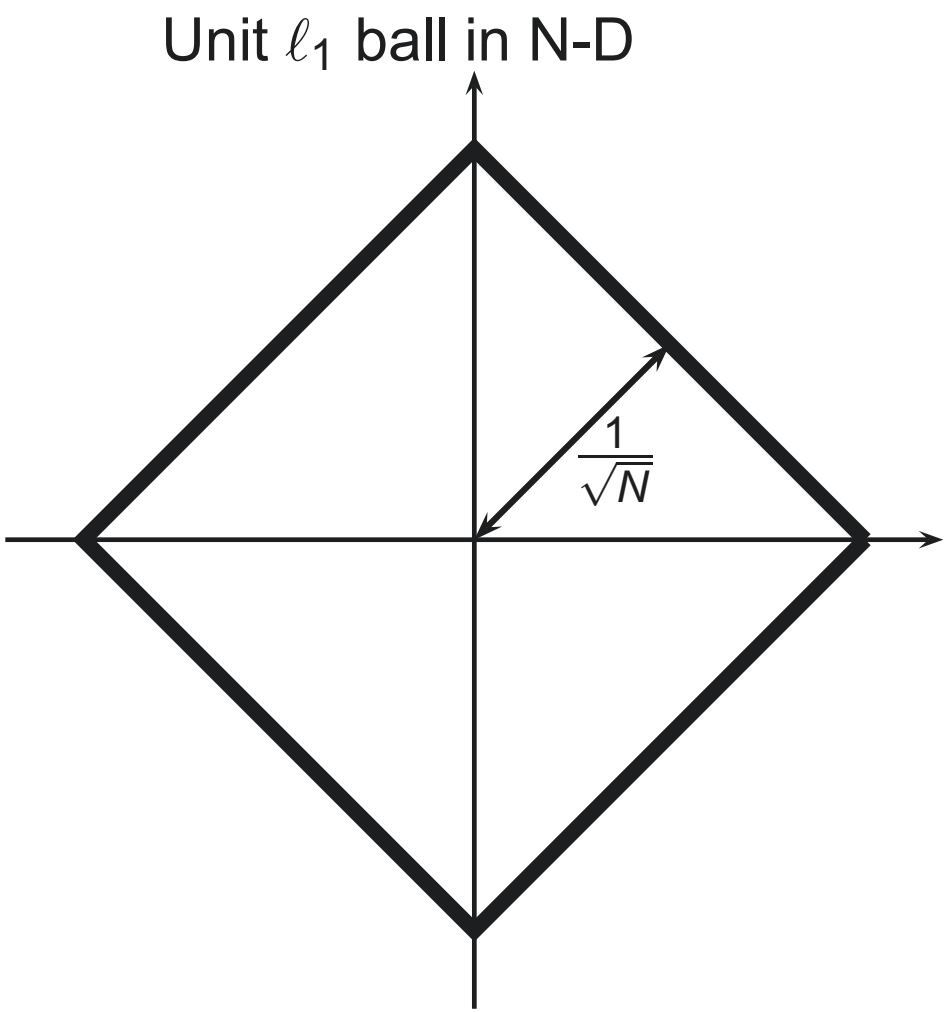
ℓ_1-norm penalty $\|u\|_1 \stackrel{\text{def}}{=} \sum_i |u_i|$ promotes sparsity and is tractable

- See [6, 3]

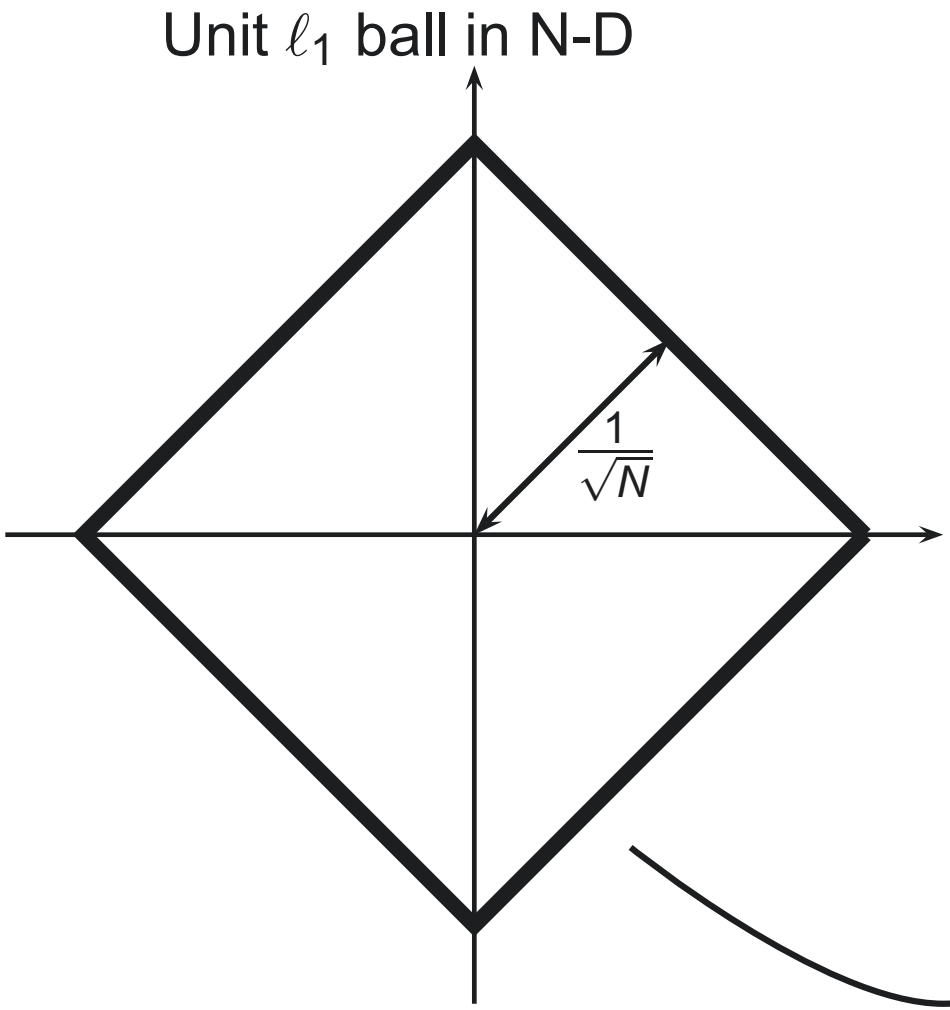
Sparsity and ℓ_1 norm

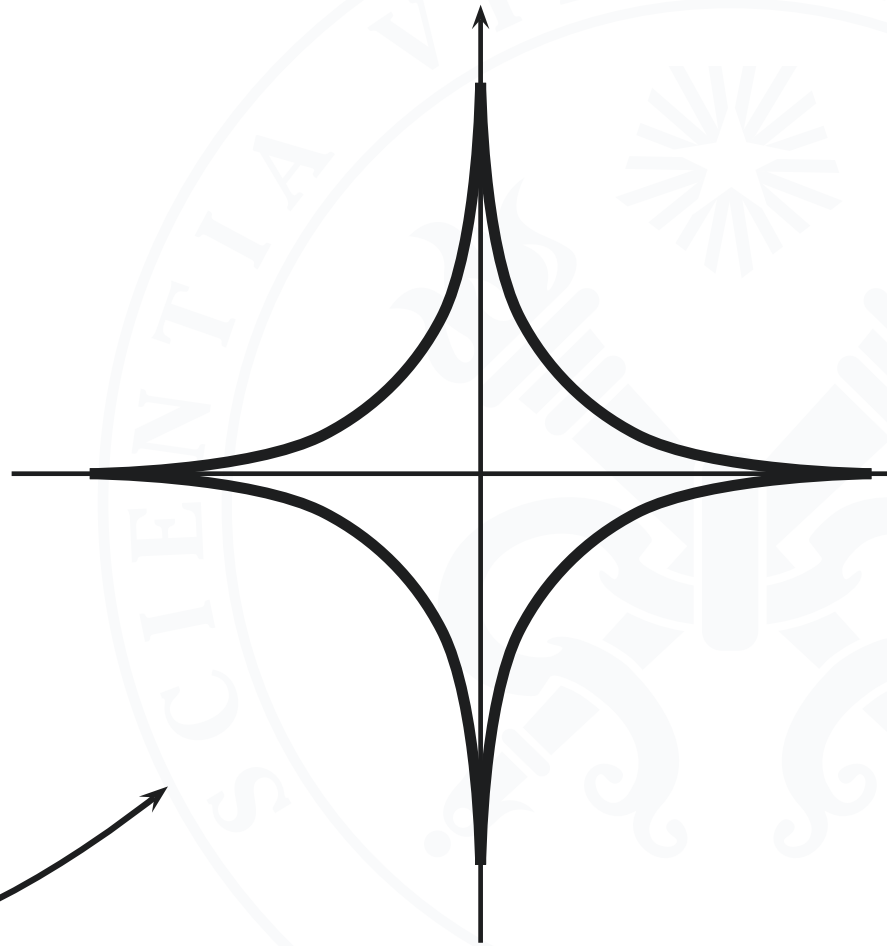


Sparsity and ℓ_1 norm



Sparsity and ℓ_1 norm





"looks like"
(when N is large)

Analysis sparsity vs. synthesis sparsity (1)

- Analysis-style sparsity:
  - Find an (approximate) solution to $Ku = y$ and
  - *Require* that certain linear combinations Au of unknown u are sparse:

$$u \quad \text{with many} \quad (Au)_i = 0$$

  Here A (analysis operator) is explicitly known.

- Corresponding optimization problem:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \lambda\|Au\|_1 \qquad (1)$$

- Primary example of (1) is total variation (TV) penalty in imaging: $A =$ local gradient of an image:

$$\|Au\|_1 = \sum_{\text{pixels}} \sqrt{(\Delta_x u)^2 + (\Delta_y u)^2}$$

Analysis sparsity vs. synthesis sparsity (1)

- Analysis-style sparsity:
  - Find an (approximate) solution to $Ku = y$ and
  - *Require* that certain linear combinations Au of unknown u are sparse:

$$u \quad \text{with many} \quad (Au)_i = 0$$

  Here A (analysis operator) is explicitly known.

- Corresponding optimization problem:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \lambda\|Au\|_1 \tag{1}$$

- Primary example of (1) is total variation (TV) penalty in imaging: $A =$ local gradient of an image:

$$\|Au\|_1 = \sum_{\text{pixels}} \sqrt{(\Delta_x u)^2 + (\Delta_y u)^2}$$

Analysis sparsity vs. synthesis sparsity (1)

- Analysis-style sparsity:
  - Find an (approximate) solution to $Ku = y$ and
  - *Require* that certain linear combinations Au of unknown u are sparse:

$$u \quad \text{with many} \quad (Au)_i = 0$$

  Here A (analysis operator) is explicitly known.

- Corresponding optimization problem:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \lambda\|Au\|_1 \tag{1}$$

- Primary example of (1) is total variation (TV) penalty in imaging: $A =$ local gradient of an image:

$$\|Au\|_1 = \sum_{\text{pixels}} \sqrt{(\Delta_x u)^2 + (\Delta_y u)^2}$$

Analysis sparsity vs. synthesis sparsity (2)

- Synthesis-style sparsity:
  Express unknown u as a sparse linear combination of a set of known basis/frame/dictionary vectors:

$$u = Sv \quad \text{with many} \quad v_i = 0$$

- Synthesis-style sparsity (express $u = Sv$ with v sparse):

$$v^{\text{rec}} = \arg\min_v \tfrac{1}{2}\|KSv - y\|_2^2 + \lambda\|v\|_1 \quad \text{and} \quad u^{\text{rec}} = Sv^{\text{rec}} \qquad (2)$$

- Example of (2): sparse combination of wavelets

- If $AS = SA = 1$ then:

  synthesis sparsity u^{rec} = *analysis sparsity* u^{rec}

Analysis sparsity vs. synthesis sparsity (2)

- Synthesis-style sparsity:
  Express unknown u as a sparse linear combination of a set of known basis/frame/dictionary vectors:

$$u = Sv \quad \text{with many} \quad v_i = 0$$

- Synthesis-style sparsity (express $u = Sv$ with v sparse):

$$v^{\text{rec}} = \arg\min_v \tfrac{1}{2}\|KSv - y\|_2^2 + \lambda\|v\|_1 \quad \text{and} \quad u^{\text{rec}} = Sv^{\text{rec}} \qquad (2)$$

- Example of (2): sparse combination of wavelets
- If $AS = SA = 1$ then:
  synthesis sparsity u^{rec} = *analysis sparsity* u^{rec}

Analysis sparsity vs. synthesis sparsity (2)

- Synthesis-style sparsity:
  Express unknown u as a sparse linear combination of a set of known basis/frame/dictionary vectors:

$$u = Sv \quad \text{with many} \quad v_i = 0$$

- Synthesis-style sparsity (express $u = Sv$ with v sparse):

$$v^{\text{rec}} = \arg\min_v \tfrac{1}{2}\|KSv - y\|_2^2 + \lambda\|v\|_1 \quad \text{and} \quad u^{\text{rec}} = Sv^{\text{rec}} \qquad (2)$$

- Example of (2): sparse combination of wavelets

- If $AS = SA = 1$ then:

  synthesis sparsity u^{rec} = *analysis sparsity* u^{rec}

Analysis sparsity vs. synthesis sparsity (2)

- Synthesis-style sparsity:
  Express unknown u as a sparse linear combination of a set of known basis/frame/dictionary vectors:

$$u = Sv \quad \text{with many} \quad v_i = 0$$

- Synthesis-style sparsity (express $u = Sv$ with v sparse):

$$v^{\text{rec}} = \arg\min_v \tfrac{1}{2}\|KSv - y\|_2^2 + \lambda\|v\|_1 \quad \text{and} \quad u^{\text{rec}} = Sv^{\text{rec}} \qquad (2)$$

- Example of (2): sparse combination of wavelets

- If $AS = SA = 1$ then:

  synthesis sparsity u^{rec} = *analysis sparsity* u^{rec}

Cost functions for 'sparse recovery'

- Many $u_i = 0$, then use penalty of type $\lambda\|u\|_1$:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \lambda\|u\|_1 \tag{3}$$

- Many $(Au)_i = 0$, then use penalty of type $\lambda\|Au\|_1$:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \lambda\|Au\|_1 \tag{4}$$

(A not necessarily invertible)

- (3) is a special case of (4): $A = 1$ or change of variables if $\exists A^{-1}$

Cost functions for 'sparse recovery'

- Many $u_i = 0$, then use penalty of type $\lambda\|u\|_1$:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \lambda\|u\|_1 \tag{3}$$

- Many $(Au)_i = 0$, then use penalty of type $\lambda\|Au\|_1$:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \lambda\|Au\|_1 \tag{4}$$

(A not necessarily invertible)

- (3) is a special case of (4): $A = 1$ or change of variables if $\exists A^{-1}$

Cost functions for 'sparse recovery'

- Many $u_i = 0$, then use penalty of type $\lambda\|u\|_1$:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \lambda\|u\|_1 \tag{3}$$

- Many $(Au)_i = 0$, then use penalty of type $\lambda\|Au\|_1$:

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \lambda\|Au\|_1 \tag{4}$$

(A not necessarily invertible)

- (3) is a special case of (4): $A = 1$ or change of variables if $\exists A^{-1}$

Goal

- Write *iterative algorithms* for finding the *numerical solutions* to the following optimization problems

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \lambda\|u\|_1 \tag{5}$$

and

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \lambda\|Au\|_1 \tag{6}$$

where K, y and A are given.

- Problem (5) in part 1, problem (6) in part 2.
- These objective functions are convex $\Rightarrow$ study the problem in the framework of *convex optimization*
- NB: “course”=definitions, properties, proofs, exercises!

Goal

- Write *iterative algorithms* for finding the *numerical solutions* to the following optimization problems

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \lambda\|u\|_1 \tag{5}$$

and

$$u^{\text{rec}} = \arg\min_u \tfrac{1}{2}\|Ku - y\|_2^2 + \lambda\|Au\|_1 \tag{6}$$

where K, y and A are given.

- Problem (5) in part 1, problem (6) in part 2.
- These objective functions are convex $\Rightarrow$ study the problem in the framework of *convex optimization*

- NB: “course”=definitions, properties, proofs, exercises!

Preliminaries

- Real d-dimensional space $\mathbb{R}^d$
- Inner product of $u, v \in \mathbb{R}^d$: $\langle u, v \rangle = \sum_{i=1}^{d} u_i v_i$
- Euclidean norm: $\|u\|_2 = \sqrt{\langle u, u \rangle} = \sqrt{\sum_{i=1}^{d} u_i^2}$
- Some special products:
  - $\|u \pm v\|_2^2 = \|u\|_2^2 + \|v\|_2^2 \pm 2\langle u, v \rangle$
  - $2\langle u, v \rangle = \pm\|u \pm v\|_2^2 \mp \|u\|_2^2 \mp \|v\|_2^2$
  - $\langle u, v \rangle = \dfrac{\|u + v\|_2^2 - \|u - v\|_2^2}{4}$
- $\langle u, Av \rangle = \langle A^T u, v \rangle$, where A is a linear operator (matrix)

Convex sets and convex functions

Definition

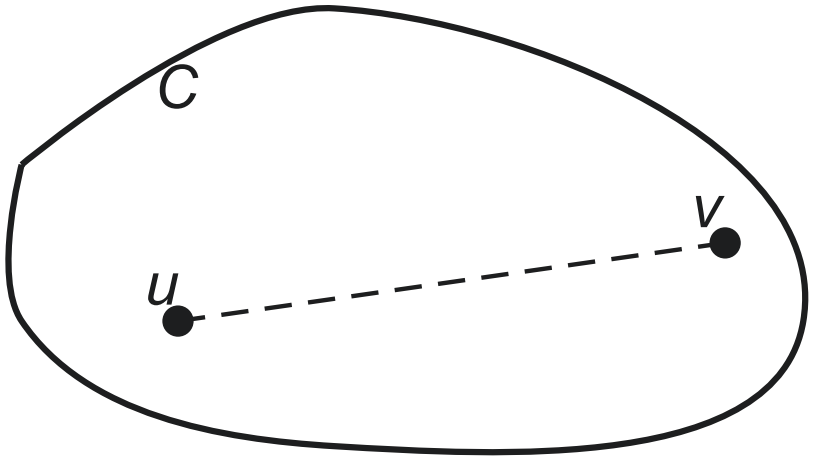
A set $C \subset \mathbb{R}^d$ is said to be convex if

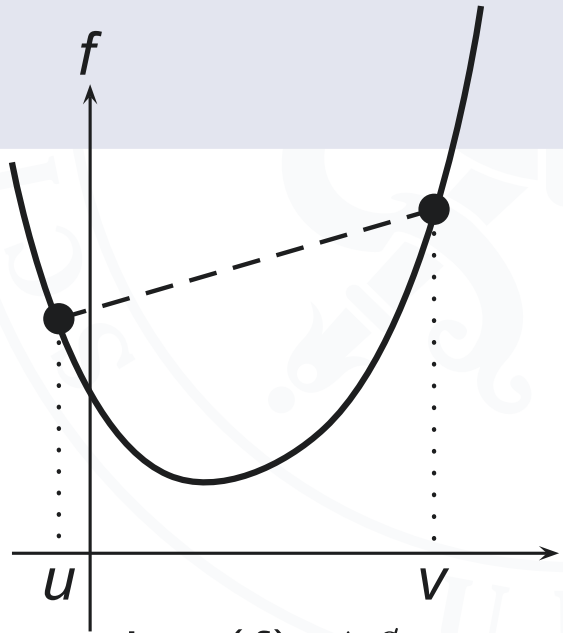
$$u, v \in C \qquad \Rightarrow \qquad \lambda u + (1-\lambda)v \in C \tag{7}$$

for all $\lambda \in [0,1]$. A function $f : \mathbb{R}^d \to \bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ is said to be convex if

$$f(\lambda u + (1-\lambda)v) \leq \lambda f(u) + (1-\lambda) f(v) \tag{8}$$

for all points $u, v \in \mathbb{R}^d$ and for all $\lambda \in [0,1]$.





NB: $\mathrm{dom}(f) = \{u | f(u) < +\infty\}$, f is 'proper' means $\mathrm{dom}(f) \neq \emptyset$

Why convex optimization?

Property

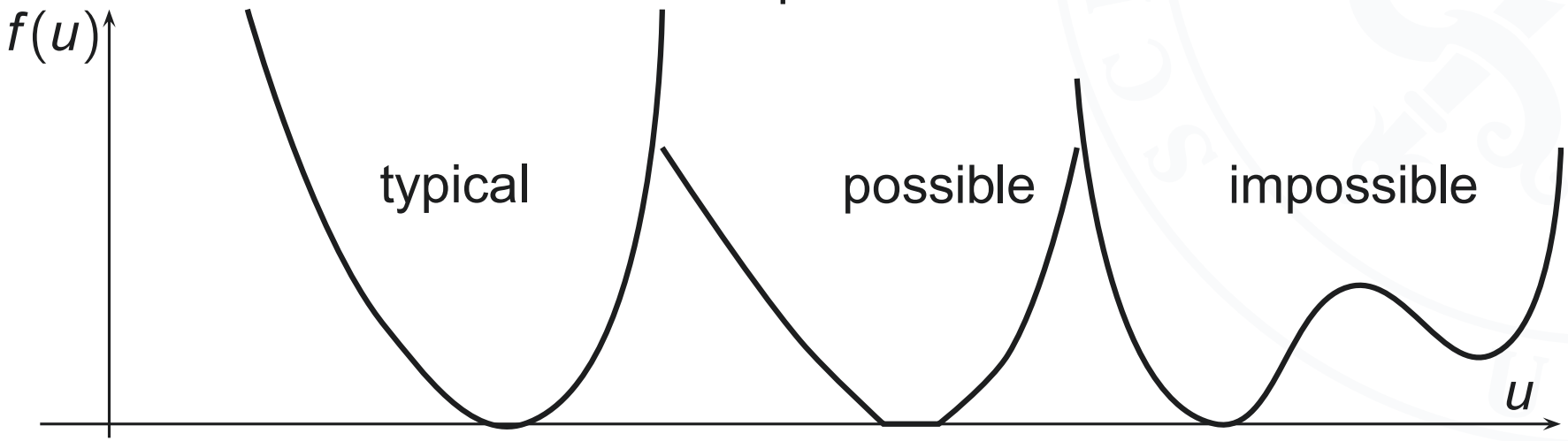
A local minimum of a convex function is necessarily a global minimum

Proof: Suppose u is a local minimum of the convex function f. If there is a point v where $f(v) < f(u)$ then

$$f(\lambda u + (1-\lambda)v) \leq \lambda f(u) + (1-\lambda)f(v) < \lambda f(u) + (1-\lambda)f(u) = f(u)$$

for all $\lambda \in [0, 1[$. But $\lambda u + (1-\lambda)v \xrightarrow{\lambda \to 1} u$ which contradicts the assumption. □

- $\rightarrow$ Easier than non-convex optimization



Convex functions: Examples

- $f(u) = \|u\|_2^2$ is a convex function
- Any norm on $\mathbb{R}^d$ is a convex function, e.g.:

$$\|u\|_1 = \sum_i |u_i|, \qquad \|u\|_2 = \left(\sum_i |u_i|^2 \right)^{1/2} \quad \text{and} \quad \|u\|_\infty = \max_i |u_i| \tag{9}$$

- ℓ_p-ball of radius R:

$$B_R^{(p)} = \{u \mid \|u\|_p \leq R\}, \qquad \text{for} \quad p = 1, 2, \infty. \tag{10}$$

  are convex sets
- The indicator functions of these convex sets:

$$i_{B_R^{(p)}}(u) = \begin{cases} 0 & \|u\|_p \leq R \\ +\infty & \|u\|_p > R \end{cases} \tag{11}$$

  are also convex functions

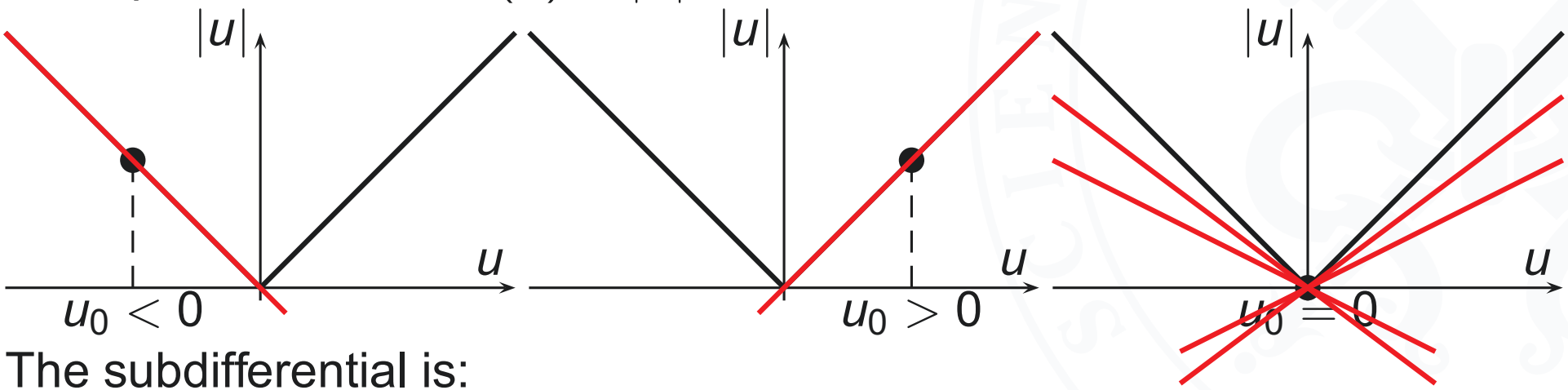
Subdifferential of a convex function

Definition

The subdifferential $\partial f(u_0)$ of f at the point $u_0 \in \mathbb{R}^d$ is the *set* of vectors $w \in \mathbb{R}^d$ such that

$$f(u) \geq f(u_0) + \langle w, u - u_0 \rangle \qquad \forall u \in \mathbb{R}^d. \tag{12}$$

Example: $f : \mathbb{R} \to \mathbb{R} : f(u) = |u|$ is not differentiable in $u_0 = 0$.



The subdifferential is:

$$\partial f(u) = \begin{cases} -1 & u < 0 \\ [-1, 1] & u = 0 \\ 1 & u > 0 \end{cases} \tag{13}$$

Subdifferential of a convex differentiable function

If f is differentiable at u, the subdifferential reduces to the usual gradient
Example: $f(u) = \frac{1}{2}\|Ku - y\|_2^2$

$$\partial f(u) = \nabla f(u) = K^T(Ku - y) \tag{14}$$

Indeed: need to show that $f(u) \geq f(u_0) + \langle \nabla f(u_0), u - u_0 \rangle \qquad \forall u$

$$\begin{aligned}
-f(u) + f(u_0) + \langle \nabla f(u_0), u - u_0 \rangle &= -\tfrac{1}{2}\|Ku - y\|_2^2 + \tfrac{1}{2}\|Ku_0 - y\|_2^2 \\
&\quad + \langle K^T(Ku_0 - y), u - u_0 \rangle \\
&= -\tfrac{1}{2}\|Ku\|_2^2 + \tfrac{1}{2}\|Ku_0\|_2^2 \\
&\quad \cancel{- \langle K(u_0 - u), y \rangle} \\
&\quad + \langle K^T(Ku_0 - y), u - u_0 \rangle \\
&= -\tfrac{1}{2}\|Ku\|_2^2 - \tfrac{1}{2}\|Ku_0\|_2^2 - \langle Ku_0, Ku \rangle \\
&= -\tfrac{1}{2}\|K(u_0 - u)\|_2^2 \\
&\leq 0 \qquad \forall v
\end{aligned}$$

NB: We will assume functions are subdifferentiable.

Characterization of the minimizer(s) of a convex function

Notation: minimizer of $f(u)$ $\overset{\text{def.}}{\Leftrightarrow}$ $\arg\min_u f(u)$

Property

$$\hat{u} \in \arg\min_u f(u) \quad \Leftrightarrow \quad 0 \in \partial f(\hat{u}). \tag{15}$$

Proof:

$$\begin{aligned} \hat{u} \in \arg\min_u f(u) &\Leftrightarrow f(u) \geq f(\hat{u}) \qquad \forall u \\ &\Leftrightarrow f(u) \geq f(\hat{u}) + \langle 0, u - \hat{u} \rangle \qquad \forall u \\ &\Leftrightarrow 0 \in \partial f(\hat{u}) \qquad \square \end{aligned}$$

Example: Suppose one wants to find a minimizer of $f(u) + g(u)$, where f is differentiable but g is not.

$$\begin{aligned} \hat{u} = \arg\min_u f(u) + g(u) &\Leftrightarrow 0 \in \partial \left(f(\hat{u}) + g(\hat{u}) \right) \\ &\Leftrightarrow 0 \in \nabla f(\hat{u}) + \partial g(\hat{u}) \\ &\Leftrightarrow \exists \hat{w} \in \partial g(\hat{u}) \quad \text{s.t.} \quad 0 = \nabla f(\hat{u}) + \hat{w} \end{aligned}$$

Proximal operator: definition

The projection onto the non-empty closed convex set C can be written as

$$P_C(u) = \arg\min_v \frac{1}{2}\|u - v\|_2^2 + i_C(v), \tag{16}$$

where i_C is the indicator function of C: $i_C(v) = \begin{cases} 0 & v \in C \\ +\infty & v \notin C. \end{cases}$

Definition (Proximal operator)

The proximal operator of a convex function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is defined as:

$$\text{prox}_f(u) = \arg\min_v \frac{1}{2}\|u - v\|_2^2 + f(v). \tag{17}$$

Remarks:

- f is assumed to be lower semi-continuous and proper ($\not\equiv +\infty$)
- prox_f is uniquely defined because $\|u - v\|_2^2$ is strictly convex
- 'standard tool for non-smooth, constrained, large-scale minimization problems' [9]

Proximal operator: elementary examples

It is easy to check that (exercise):

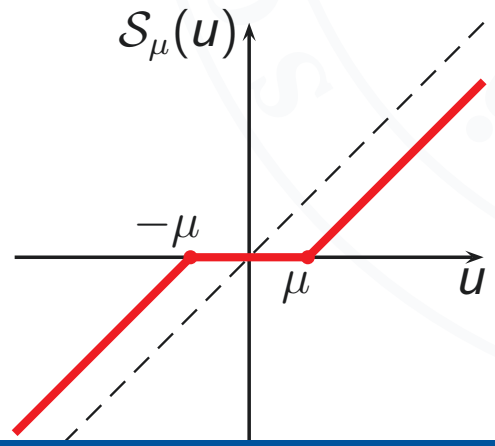
$$f(u) = \langle a, u \rangle + b \quad \Rightarrow \quad \text{prox}_f(u) = u - a \tag{18}$$

$$f(u) = \frac{\mu}{2}\|u\|_2^2 \quad \Rightarrow \quad \text{prox}_f(u) = \frac{u}{1+\mu} \tag{19}$$

$$f(u) = \mu|u| \quad \text{(1 var.)} \quad \Rightarrow \quad \text{prox}_f(u) = \begin{cases} 0 & |u| \leq \mu \\ u - \mu\,\text{sgn}(u) & |u| \geq \mu \end{cases} \tag{20}$$

$$f(u) = \mu\|u\|_2 \quad \Rightarrow \quad \text{prox}_f(u) = \begin{cases} 0 & \text{if} \quad \|u\|_2 \leq \mu \\ u - \mu\frac{u}{\|u\|_2} & \text{if} \quad \|u\|_2 \geq \mu \end{cases} \tag{21}$$

Proximity operator of $\mu|u|$ is "soft-thresholding": $\text{prox}_f(u) = \mathcal{S}_\mu(u)$



Proximal operator: Properties

Property

Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be a convex function, then:

$$g(u) = f(u - a) \quad \Rightarrow \quad \text{prox}_g(u) = a + \text{prox}_f(u - a). \tag{22}$$

Proof: One finds that:

$$\begin{aligned}
\text{prox}_g(u) &= \arg\min_v \tfrac{1}{2}\|u - v\|_2^2 + g(v) \\
&= \arg\min_v \tfrac{1}{2}\|u - v\|_2^2 + f(v - a) \\
&\overset{v=a+\tilde{v}}{=} a + \arg\min_{\tilde{v}} \tfrac{1}{2}\|u - a - \tilde{v}\|_2^2 + f(\tilde{v}) \\
&= a + \text{prox}_f(u - a)
\end{aligned}$$

□

NB: Similar formula for $\text{prox}_g(u)$ where $g(u) = f(\alpha u)$:

$$\text{prox}_g(u) = \frac{1}{\alpha}\text{prox}_{\alpha^2 f}(\alpha u)$$

Proximal operator: Properties

Property

Let $f_1 : \mathbb{R}^{d_1} \to \bar{\mathbb{R}}$ and $f_2 : \mathbb{R}^{d_2} \to \bar{\mathbb{R}}$ be convex functions of u_1 and u_2 respectively. Let $f : \mathbb{R}^{d_1+d_2} \to \bar{\mathbb{R}}$. One has:

$$f(u_1, u_2) = f_1(u_1) + f_2(u_2) \quad \Rightarrow \quad \text{prox}_f(u_1, u_2) = (\text{prox}_{f_1}(u_1), \text{prox}_{f_2}(u_2)). \tag{23}$$

Proof: Setting $u = (u_1, u_2)$ and $v = (v_1, v_2)$ (both elements of $\mathbb{R}^{d_1+d_2}$), one has

$$\begin{aligned}
&\arg\min_v \frac{1}{2}\|u - v\|^2 + f(v) \\
&\qquad = \arg\min_{v_1, v_2} \frac{1}{2}\|(u_1, u_2) - (v_1, v_2)\|_2^2 + f_1(v_1) + f_2(v_2) \\
&\qquad = \arg\min_{v_1, v_2} \frac{1}{2}\|u_1 - v_1\|_2^2 + f_1(v_1) + \frac{1}{2}\|u_2 - v_2\|_2^2 + f_2(v_2) \\
&\qquad = (\text{prox}_{f_1}(u_1), \text{prox}_{f_2}(u_2))
\end{aligned}$$

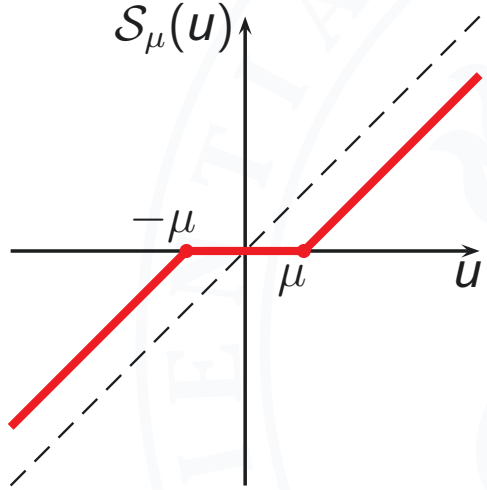
□

Proximal operator: Example

Example 1:

$$f = \mu\|u\|_1 \quad \Rightarrow \quad \text{prox}_f(u)_i = \begin{cases} 0 & |u_i| \leq \mu \\ u_i - \text{sgn}(u_i)\mu & |u_i| \geq \mu. \end{cases} \tag{24}$$

Proximity operator of $\mu\|u\|_1$ is component-wise "soft-thresholding": $\text{prox}_f(u) = \mathcal{S}_\mu(u)$



Example 2: $f(u) = \mu \sum_i \sqrt{u_{i,1}^2 + u_{i,2}^2}$

$$(\text{prox}_f(u))_i = \begin{cases} (0,0) & \text{if} \quad \sqrt{u_{i,1}^2 + u_{i,2}^2} \leq \mu \\ (u_{i,1}, u_{i,2}) - \mu \frac{(u_{i,1}, u_{i,2})}{\sqrt{u_{i,1}^2 + u_{i,2}^2}} & \text{if} \quad \sqrt{u_{i,1}^2 + u_{i,2}^2} \geq \mu. \end{cases} \tag{25}$$

using formula (21).

Proximal operator: Properties

Property

Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be a convex function. If $t^+ = \text{prox}_f(t^- + \Delta)$ then:

$$\|t^+ - t\|_2^2 \leq \|t^- - t\|_2^2 - \|t^+ - t^-\|_2^2 + 2\langle t^+ - t, \Delta\rangle + 2f(t) - 2f(t^+) \quad (26)$$

for all t.

Proof: $t^+ = \text{prox}_f(t^- + \Delta)$

$$\begin{aligned}
&\Leftrightarrow \quad t^+ = \arg\min_t \frac{1}{2}\|t - (t^- + \Delta)\|_2^2 + f(t) \\
&\Leftrightarrow \quad 0 \in t^+ - t^- - \Delta + \partial f(t^+) \\
&\Leftrightarrow \quad t^- + \Delta - t^+ \in \partial f(t^+) \\
&\Leftrightarrow \quad f(t) \geq f(t^+) + \langle t^- + \Delta - t^+, t - t^+\rangle \\
&\Leftrightarrow \quad 0 \leq 2\langle t^- - t^+, t^+ - t\rangle + 2\langle t^+ - t, \Delta\rangle + 2f(t) - 2f(t^+) \\
&\Leftrightarrow \quad \|t^+ - t\|_2^2 \leq \|t^- - t\|_2^2 - \|t^+ - t^-\|_2^2 + 2\langle t^+ - t, \Delta\rangle + 2f(t) - 2f(t^+)
\end{aligned}$$

Firmly non-expansive operators

Definition

A map $T : \mathbb{R}^d \to \mathbb{R}^d$ is said to be firmly non-expansive if

$$\|Tu - Tv\|_2^2 \leq \langle Tu - Tv, u - v \rangle \qquad \forall u, v \in \mathbb{R}^d \tag{27}$$

Property

prox_f *is firmly non-expansive.*

Proof: If $u^+ = \text{prox}_f(u)$, eq. (26) implies $(t^+ = u^+, t^- = u, \Delta = 0, t = v^+)$:

$$\|u^+ - v^+\|_2^2 \leq \|u - v^+\|_2^2 - \|u^+ - u\|_2^2 + 2f(v^+) - 2f(u^+)$$

If $v^+ = \text{prox}_f(v)$, eq. (26) implies $(t^+ = v^+, t^- = v, \Delta = 0, t = u^+)$:

$$\|v^+ - u^+\|_2^2 \leq \|v - u^+\|_2^2 - \|v^+ - v\|_2^2 + 2f(u^+) - 2f(v^+)$$

$$\begin{aligned} \Rightarrow 2\|u^+ - v^+\|_2^2 &\leq \|u - v^+\|_2^2 - \|u^+ - u\|_2^2 + \|v - u^+\|_2^2 - \|v^+ - v\|_2^2 \\ &= -2uv^+ + 2u^+u - 2vu^+ + 2v^+v \\ &= 2\langle u^+ - v^+, u - v \rangle \end{aligned}$$

□

Proximal operator: Continuity

Property

$\text{prox}_f(u)$ *is Lipschitz continuous in* u*, with Lipschitz constant equal to* 1*:*

$$\|\text{prox}_f(u) - \text{prox}_f(v)\|_2 \leq \|u - v\|_2 \qquad \forall u, v \in \mathbb{R}^d. \tag{28}$$

Proof: prox_f is firmly non-expansive:

$$\|\text{prox}_f(u) - \text{prox}_f(v)\|_2^2 \leq \langle \text{prox}_f(u) - \text{prox}_f(v), u - v \rangle$$

Thus:

$$\|\text{prox}_f(u) - \text{prox}_f(v)\|_2^2 \leq \|\text{prox}_f(u) - \text{prox}_f(v)\|_2 \, \|u - v\|_2$$

and hence:

$$\|\text{prox}_f(u) - \text{prox}_f(v)\|_2 \leq \|u - v\|_2 \qquad \square$$

Proximal operator: Continuity

Property

For fixed u, $\text{prox}_{\alpha f}(u)$ is continuous with respect to α ($\alpha > 0$).

Proof: Let $u_\alpha = \text{prox}_{\alpha f}(u)$ and $u_{\alpha_0} = \text{prox}_{\alpha_0 f}(u)$. We need to prove that $\|u_\alpha - u_{\alpha_0}\|_2 \overset{\alpha \to \alpha_0}{\longrightarrow} 0$.

Eq. (26) with $(t^+ = u_\alpha, t^- = u, \Delta = 0, t = u_{\alpha_0})$ implies:

$$\|u_\alpha - u_{\alpha_0}\|_2^2 \leq \|u - u_{\alpha_0}\|_2 - \|u_\alpha - u\|_2^2 + 2\alpha f(u_{\alpha_0}) - 2\alpha f(u_\alpha)$$

Eq. (26) with $(t^+ = u_{\alpha_0}, t^- = u, \Delta = 0, t = u_\alpha)$ implies:

$$\|u_{\alpha_0} - u_\alpha\|_2^2 \leq \|u - u_\alpha\|_2^2 - \|u_{\alpha_0} - u\|_2^2 + 2\alpha_0 f(u_\alpha) - 2\alpha_0 f(u_{\alpha_0})$$

Together:

$$2\|u_\alpha - u_{\alpha_0}\|_2^2 \leq 2(\alpha - \alpha_0)(f(u_{\alpha_0}) - f(u_\alpha))$$

We treat the left and right limit ($\alpha \to \alpha_0$) separately:

Proximal operator: Continuity

- Case $\alpha > \alpha_0$ and $\alpha \to \alpha_0$: Choose $\alpha_0 < \alpha \leq 2\alpha_0$ (with $2\alpha_0 > \alpha_0$). Then $f(u_{2\alpha_0}) \leq f(u_\alpha)$:

$$\begin{aligned} \|u_\alpha - u_{\alpha_0}\|_2^2 &\leq (\alpha - \alpha_0)(f(u_{\alpha_0}) - f(u_\alpha)) \\ &\leq (\alpha - \alpha_0)\left(f(u_{\alpha_0}) - f(u_{2\alpha_0})\right) \overset{\alpha \to \alpha_0}{\longrightarrow} 0 \end{aligned}$$

  NB: if $\alpha_0 > 0$ then $u_{\alpha_0}, u_{2\alpha_0} \in \mathrm{dom}(f) \neq \emptyset$ and rhs is finite

- Case $\alpha < \alpha_0$ and $\alpha \to \alpha_0$: Choose $\alpha_0/2 \leq \alpha < \alpha_0$ (with $\alpha_0/2 < \alpha_0$). Then $f(u_\alpha) \leq f(u_{\alpha_0/2})$.

$$\begin{aligned} \|u_\alpha - u_{\alpha_0}\|_2^2 &\leq (\alpha_0 - \alpha)\left(f(u_\alpha) - f(u_{\alpha_0})\right) \\ &\leq (\alpha_0 - \alpha)\left(f(u_{\alpha_0/2}) - f(u_{\alpha_0})\right) \overset{\alpha \to \alpha_0}{\longrightarrow} 0 \end{aligned}$$

□

We have used:

Property

Let $u_\alpha = \mathrm{prox}_{\alpha f}(u) = \arg\min_v \frac{1}{2}\|v - u\|_2^2 + \alpha f(v)$. *If* $\alpha \geq \beta$ *then* $\|u_\alpha - u\|_2 \geq \|u_\beta - u\|_2$ *and* $f(u_\alpha) \leq f(u_\beta)$. *(Proof: exercise).*

Proximal operator: Continuity

Property

$\text{prox}_{\alpha f}(u)$ *is continuous with respect to* (u, α) *for* $u \in \mathbb{R}^d$ *and* $\alpha > 0$.

Proof:

$$
\begin{aligned}
&\|\text{prox}_{\alpha f}(u) - \text{prox}_{\alpha_0 f}(u_0)\|_2 \\
&\quad = \|\text{prox}_{\alpha f}(u) - \text{prox}_{\alpha f}(u_0) + \text{prox}_{\alpha f}(u_0) - \text{prox}_{\alpha_0 f}(u_0)\|_2 \\
&\quad \leq \|\text{prox}_{\alpha f}(u) - \text{prox}_{\alpha f}(u_0)\|_2 + \|\text{prox}_{\alpha f}(u_0) - \text{prox}_{\alpha_0 f}(u_0)\|_2 \\
&\quad \leq \|u - u_0\|_2 + \|\text{prox}_{\alpha f}(u_0) - \text{prox}_{\alpha_0 f}(u_0)\|_2 \\
&\quad \leq \epsilon/2 + \epsilon/2 \qquad \text{if} \quad \|(u, \alpha) - (u_0, \alpha_0)\| < \delta
\end{aligned}
$$

□

NB: This means that if $\alpha_n \overset{n\to\infty}{\longrightarrow} \alpha$ and $u_n \overset{n\to\infty}{\longrightarrow} u$ then:

$$\text{prox}_{\alpha_n f}(u_n) \overset{n\to\infty}{\longrightarrow} \text{prox}_{\alpha f}(u) \tag{29}$$

Characterization of subdifferential

Property

Let $g : \mathbb{R}^d \to \bar{\mathbb{R}}$ be a convex function. One has:

$$w \in \partial g(u) \qquad \Leftrightarrow \qquad u = \text{prox}_g(u + w) \tag{30}$$

Proof: $w \in \partial g(u)$

$$\begin{aligned}
&\Leftrightarrow 0 \in -w + \partial g(u) \\
&\Leftrightarrow 0 \in v - (u + w) + \partial g(v) \quad \text{at} \quad v = u \\
&\Leftrightarrow 0 \in \partial_v \left[\tfrac{1}{2}\|v - (u + w)\|_2^2 + g(v) \right] \quad \text{at} \quad v = u \\
&u = \arg\min_v \tfrac{1}{2}\|v - (u + w)\|_2^2 + g(v) = \text{prox}_g(u + w) \qquad \square
\end{aligned}$$

Characterization of minimizers of $f + g$

Property

Let $f : \mathbb{R}^d \to \mathbb{R}$ convex differentiable and $g : \mathbb{R}^d \to \bar{\mathbb{R}}$ convex functions. The following are equivalent:

1. *$\hat{u}$ is a minimizer of $f(u) + g(u)$.*
2. *There exist $\hat{w} \in \partial g(\hat{u})$ such that $\hat{w} = -\nabla f(\hat{u})$*
3. *$\hat{u}$ that satisfies the equation: $\hat{u} = \text{prox}_{\alpha g}(\hat{u} - \alpha \nabla f(\hat{u}))$ for $\alpha > 0$.*

Proof: (2) $\Leftrightarrow \quad \hat{u} = \text{prox}_g(\hat{u} + \hat{w}) \quad \text{s.t.} \quad \hat{w} = -\nabla f(\hat{u})$

$$\Leftrightarrow \quad \hat{u} = \text{prox}_g(\hat{u} - \nabla f(\hat{u}))$$

and minimizer of $f(u) + g(u)$ is same as minimizer of $\alpha f(u) + \alpha g(u)$. $\square$

NB: The last equation is a fixed-point equation (Ansatz for writing an iterative algorithm).

Fixed-point iterations

- $\hat{u}$ is minimizer of $f(u) + g(u)$
- $\hat{u}$ satisfies $\hat{u} = \text{prox}_{\alpha g}(\hat{u} - \alpha \nabla f(\hat{u}))$ for some $\alpha > 0$
- Iterative algorithm could be:

$$u_{n+1} = \text{prox}_{\alpha g}(u_n - \alpha \nabla f(u_n))$$

or

$$u_{n+1} = \text{prox}_{\alpha_n g}(u_n - \alpha_n \nabla f(u_n)).$$

- We will study convergence of:

$$\begin{cases} \tilde{u}_{n+1} &= \text{prox}_{\alpha_n g}(u_n - \alpha_n \nabla f(u_n)) \\ u_{n+1} &= (1 - \lambda_n) u_n + \lambda_n \tilde{u}_{n+1} \end{cases}$$

with $0 < \lambda_n \leq 1$.

Fixed-point iterations

- $\hat{u}$ is minimizer of $f(u) + g(u)$
- $\hat{u}$ satisfies $\hat{u} = \text{prox}_{\alpha g}(\hat{u} - \alpha \nabla f(\hat{u}))$ for some $\alpha > 0$
- Iterative algorithm could be:

$$u_{n+1} = \text{prox}_{\alpha g}(u_n - \alpha \nabla f(u_n))$$

or

$$u_{n+1} = \text{prox}_{\alpha_n g}(u_n - \alpha_n \nabla f(u_n)).$$

- We will study convergence of:

$$\begin{cases} \tilde{u}_{n+1} &= \text{prox}_{\alpha_n g}(u_n - \alpha_n \nabla f(u_n)) \\ u_{n+1} &= (1 - \lambda_n) u_n + \lambda_n \tilde{u}_{n+1} \end{cases}$$

with $0 < \lambda_n \leq 1$.

Fixed-point iterations

- $\hat{u}$ is minimizer of $f(u) + g(u)$
- $\hat{u}$ satisfies $\hat{u} = \text{prox}_{\alpha g}(\hat{u} - \alpha \nabla f(\hat{u}))$ for some $\alpha > 0$
- Iterative algorithm could be:

$$u_{n+1} = \text{prox}_{\alpha g}(u_n - \alpha \nabla f(u_n))$$

or

$$u_{n+1} = \text{prox}_{\alpha_n g}(u_n - \alpha_n \nabla f(u_n)).$$

- We will study convergence of:

$$\begin{cases} \tilde{u}_{n+1} &= \text{prox}_{\alpha_n g}(u_n - \alpha_n \nabla f(u_n)) \\ u_{n+1} &= (1 - \lambda_n) u_n + \lambda_n \tilde{u}_{n+1} \end{cases}$$

with $0 < \lambda_n \leq 1$.

Fixed-point iterations

- $\hat{u}$ is minimizer of $f(u) + g(u)$
- $\hat{u}$ satisfies $\hat{u} = \text{prox}_{\alpha g}(\hat{u} - \alpha \nabla f(\hat{u}))$ for some $\alpha > 0$
- Iterative algorithm could be:

$$u_{n+1} = \text{prox}_{\alpha g}(u_n - \alpha \nabla f(u_n))$$

or

$$u_{n+1} = \text{prox}_{\alpha_n g}(u_n - \alpha_n \nabla f(u_n)).$$

- We will study convergence of:

$$\begin{cases} \tilde{u}_{n+1} &= \text{prox}_{\alpha_n g}\left(u_n - \alpha_n \nabla f(u_n)\right) \\ u_{n+1} &= (1 - \lambda_n) u_n + \lambda_n \tilde{u}_{n+1} \end{cases}$$

with $0 < \lambda_n \leq 1$.

Lemma: $\frac{1}{L}\nabla f$ is firmly non-expansive

Property

If $f : \mathbb{R}^d \to \mathbb{R}$ is convex with Lipschitz continuous gradient (L) then $\frac{1}{L}\nabla f$ is firmly non-expansive:

$$\|\nabla f(u) - \nabla f(v)\|_2^2 \leq L\langle \nabla f(u) - \nabla f(v), u - v\rangle \qquad \forall u, v \in \mathbb{R}^d \tag{31}$$

Proof: see [7, Part 2, Chapter X, Th. 4.2.2]. Here we give a proof for $f(u) = \frac{1}{2}\|Ku - y\|_2^2$. In this case $\nabla f(u) = K^T(Ku - y)$ and $L = \sigma_{\max}(K)^2$, such that:

$$\begin{aligned}
\|\nabla f(u) - \nabla f(v)\|_2^2 &= \|K^T(Ku - y) - K^T(Ku - y)\|_2^2 \\
&= \|K^T(Ku - v)\|_2^2 \\
&\leq L\|K(u - v)\|_2^2 \\
&= L\langle K(u - v), K(u - v)\rangle \\
&= L\langle K^T K(u - v), u - v\rangle \\
&= L\langle \nabla f(u) - \nabla f(v), u - v\rangle
\end{aligned}$$

□

Proximal gradient algorithm

Theorem (proximal gradient algorithm [4])

Let $\epsilon > 0$. IF $f : \mathbb{R}^d \to \mathbb{R}$ is convex with Lipschitz continuous gradient (L), $g : \mathbb{R}^d \to \mathbb{R}$ is convex, proper, lower semi-continuous, and a minimizer of $F(u) = f(u) + g(u)$ exists, THEN the proximal gradient algorithm:

$$\begin{cases} \tilde{u}_{n+1} &= \text{prox}_{\alpha_n g}\left(u_n - \alpha_n \nabla f(u_n)\right) \\ u_{n+1} &= (1-\lambda_n)u_n + \lambda_n \tilde{u}_{n+1} \end{cases} \tag{32}$$

with $u_0 =$ arbitrary, $\epsilon \leq \alpha_n \leq 2/L - \epsilon$ and $\epsilon \leq \lambda_n \leq 1$ converges to a minimizer of $F(u)$.

Proof: Let $\hat{u} \in \arg\min_u f(u) + g(u)$, i.e. $\hat{u} = \text{prox}_{\alpha g}\left(\hat{u} - \alpha \nabla f(\hat{u})\right)$.
One has:

$$\|u_{n+1} - \hat{u}\|_2^2 = (1-\lambda_n)\|u_n - \hat{u}\|_2^2 + \lambda_n \|\tilde{u}_{n+1} - \hat{u}\|_2^2 - \lambda_n(1-\lambda_n)\|u_n - \tilde{u}_{n+1}\|_2^2 \tag{33}$$

Recall property of proximal operators

If $t^+ = \text{prox}_g(t^- + \Delta)$ then:

$$\|t^+ - t\|_2^2 \leq \|t^- - t\|_2^2 - \|t^+ - t^-\|_2^2 + 2\langle t^+ - t, \Delta\rangle + 2g(t) - 2g(t^+) \quad (34)$$

for all t.

We will use this property on the (iteration) relation:

$$\tilde{u}_{n+1} = \text{prox}_{\alpha_n g}\left(u_n - \alpha_n \nabla f(u_n)\right)$$

with $t^+ = \tilde{u}_{n+1}, t^- = u_n, t = \hat{u}, \Delta = -\alpha_n \nabla f(u_n)$,

and on the fixed-point relation:

$$\hat{u} = \text{prox}_{\alpha_n g}\left(\hat{u} - \alpha_n \nabla f(\hat{u})\right)$$

with $t^+ = \hat{u}, t^- = \hat{u}, t = \tilde{u}_{n+1}, \Delta = -\alpha_n \nabla f(\hat{u})$

Proximal gradient algorithm: proof of convergence

It follows from $\tilde{u}_{n+1} = \text{prox}_{\alpha_n g}\left(u_n - \alpha_n \nabla f(u_n)\right)$ and eq. (26) with $t^+ = \tilde{u}_{n+1}, t^- = u_n, t = \hat{u}, \Delta = -\alpha_n \nabla f(u_n)$ that:

$$\begin{aligned}\|\tilde{u}_{n+1} - \hat{u}\|_2^2 \quad \leq \quad & \|u_n - \hat{u}\|_2^2 - \|\tilde{u}_{n+1} - u_n\|_2^2 + 2\langle \tilde{u}_{n+1} - \hat{u}, -\alpha_n \nabla f(u_n)\rangle \\ & + 2\alpha_n g(\hat{u}) - 2\alpha_n g(\tilde{u}_{n+1})\end{aligned}$$

It follows from $\hat{u} = \text{prox}_{\alpha_n g}\left(\hat{u} - \alpha_n \nabla f(\hat{u})\right)$ and eq. (26) with $t^+ = \hat{u}, t^- = \hat{u}, t = \tilde{u}_{n+1}, \Delta = -\alpha_n \nabla f(\hat{u})$ that:

$$\begin{aligned}\cancel{\|\hat{u} - \tilde{u}_{n+1}\|_2^2} \quad \leq \quad & \cancel{\|\hat{u} - \tilde{u}_{n+1}\|_2^2} - \cancel{\|\hat{u} - \hat{u}\|_2^2} + 2\langle \hat{u} - \tilde{u}_{n+1}, -\alpha_n \nabla f(\hat{u})\rangle \\ & + 2\alpha_n g(\tilde{u}_{n+1}) - 2\alpha_n g(\hat{u})\end{aligned}$$

Together:

$$\begin{aligned}\|\tilde{u}_{n+1} - \hat{u}\|_2^2 \quad \leq \quad & \|u_n - \hat{u}\|_2^2 - \|\tilde{u}_{n+1} - u_n\|_2^2 \\ & + 2\alpha_n \langle \hat{u} - \tilde{u}_{n+1}, \nabla f(u_n) - \nabla f(\hat{u})\rangle\end{aligned} \tag{35}$$

The inner product can be bounded by:

Proximal gradient algorithm: proof of convergence

$$
\begin{aligned}
\langle \hat{u} - \tilde{u}_{n+1}, \nabla f(u_n) - \nabla f(\hat{u}) \rangle = \quad & \langle \hat{u} - u_n, \nabla f(u_n) - \nabla f(\hat{u}) \rangle \\
& + \langle u_n - \tilde{u}_{n+1}, \nabla f(u_n) - \nabla f(\hat{u}) \rangle \\
\overset{(31)}{\leq} \quad & \frac{-1}{L} \|\nabla f(u_n) - \nabla f(\hat{u})\|_2^2 \\
& + \langle u_n - \tilde{u}_{n+1}, \nabla f(u_n) - \nabla f(\hat{u}) \rangle \\
= \quad & \langle \sqrt{L}(u_n - \tilde{u}_{n+1}) - \frac{1}{\sqrt{L}}(\nabla f(u_n) - \nabla f(\hat{u})), \\
& \frac{1}{\sqrt{L}}(\nabla f(u_n) - \nabla f(\hat{u})) \rangle \\
\overset{\langle a,b \rangle = \frac{\|a+b\|_2^2 - \|a-b\|_2^2}{4}}{=} \quad & \frac{L\|u_n - \tilde{u}_{n+1} + 0\|_2^2 - \|u_n - \tilde{u}_{n+1} - \frac{2}{\sqrt{L}} \dots \|_2^2}{4} \\
\leq \quad & \frac{L}{4} \|u_n - \tilde{u}_{n+1}\|_2^2
\end{aligned}
$$

Proximal gradient algorithm: proof of convergence

The latter inequality combined with expression (35) yields:

$$\begin{aligned}\|\tilde{u}_{n+1} - \hat{u}\|_2^2 &\leq \|u_n - \hat{u}\|_2^2 - \|\tilde{u}_{n+1} - u_n\|_2^2 + 2\alpha_n \tfrac{L}{4}\|u_n - \tilde{u}_{n+1}\|_2^2 \\ &= \|u_n - \hat{u}\|_2^2 - (1 - \tfrac{\alpha_n L}{2})\|\tilde{u}_{n+1} - u_n\|_2^2\end{aligned}$$

This can be inserted in expression (33) to yield:

$$\begin{aligned}\|u_{n+1} - \hat{u}\|_2^2 &\overset{(33)}{=} (1-\lambda_n)\|u_n - \hat{u}\|_2^2 + \lambda_n\|\tilde{u}_{n+1} - \hat{u}\|_2^2 - \lambda_n(1-\lambda_n)\|u_n - \tilde{u}_{n+1}\|_2^2 \\ &\leq (1-\lambda_n)\|u_n - \hat{u}\|_2^2 + \lambda_n\left[\|u_n - \hat{u}\|_2^2 - (1 - \tfrac{\alpha_n L}{2})\|\tilde{u}_{n+1} - u_n\|_2^2\right] \\ &\qquad - \lambda_n(1-\lambda_n)\|u_n - \tilde{u}_{n+1}\|_2^2 \\ &= \|u_n - \hat{u}\|_2^2 - \lambda_n\left[(1 - \tfrac{\alpha_n L}{2} + 1 - \lambda_n)\|\tilde{u}_{n+1} - u_n\|_2^2\right] \\ &\leq \|u_n - \hat{u}\|_2^2 - \epsilon(\epsilon L/2 + 0)\|\tilde{u}_{n+1} - u_n\|_2^2\end{aligned}$$

as $\lambda_n \geq \epsilon, 1 - \lambda_n \geq 0$ and $1 - \frac{\alpha_n L}{2} \geq \epsilon L/2$.

Proximal gradient algorithm: proof of convergence

One therefore has (with $c = \epsilon^2 L/2 > 0$):

$$\|u_{n+1} - \hat{u}\|_2^2 \leq \|u_n - \hat{u}\|_2^2 - c\,\|\tilde{u}_{n+1} - u_n\|_2^2 \tag{36}$$

- Eq. (36) implies: $\|u_{n+1} - \hat{u}\|_2 \leq \|u_n - \hat{u}\|_2$, i.e. $(u_n)_n$ is bounded. As $(\alpha_n)_n$ and $(\lambda_n)_n$ are also bounded, there exists a common converging subsequence:

$$u_{n_j} \overset{j\to\infty}{\longrightarrow} u^\dagger, \quad \alpha_{n_j} \overset{j\to\infty}{\longrightarrow} \alpha > 0, \quad \lambda_{n_j} \overset{j\to\infty}{\longrightarrow} \lambda \quad (\text{with } 0 < \lambda \leq 1)$$

- Eq. (36) also implies ($N > M$):

$$\begin{aligned} c\sum_{n=M}^{N-1} \|\tilde{u}_{n+1} - u_n\|_2^2 &\leq \sum_{n=M}^{N-1} \|u_n - \hat{u}\|_2^2 - \|u_{n+1} - \hat{u}\|_2^2 \\ &= \|u_M - \hat{u}\|_2^2 - \|u_N - \hat{u}\|_2^2 \\ &\leq \|u_M - \hat{u}\|_2^2 \ (= \text{independent of } N) \end{aligned} \tag{37}$$

This means that $\|\tilde{u}_{n+1} - u_n\|_2 \overset{n\to\infty}{\longrightarrow} 0$ and thus: $\tilde{u}_{n_j+1} \overset{j\to\infty}{\longrightarrow} u^\dagger$

Proximal gradient algorithm: proof of convergence

- But as $\tilde{u}_{n_j+1} = \text{prox}_{\alpha_{n_j} g}\left(u_{n_j} - \alpha_{n_j}\nabla f(u_{n_j})\right)$, one finds ($j \to \infty$) that:

$$u^\dagger = \text{prox}_{\alpha g}\left(u^\dagger - \alpha\nabla f(u^\dagger)\right)$$

  i.e. $u^\dagger$ is a minimizer of $f + g$.
- Finally, choosing $\hat{u} = u^\dagger$, inequality (37) implies that:

$$\|u_N - u^\dagger\|_2^2 \leq \|u_M - u^\dagger\|_2^2 \qquad \text{for} \quad N > M$$

  As $u_{n_j} \overset{j\to\infty}{\longrightarrow} u^\dagger$, the rhs can be made as small as one likes. This shows that the whole sequence $(u_n)_n$ converges to $u^\dagger$.

One also has:

$$\begin{aligned}
\|\tilde{u}_{n+1} - u^\dagger\|_2 &= \|\text{prox}_{\alpha_n g}\left(u_n - \alpha_n\nabla f(u_n)\right) - \text{prox}_{\alpha_n g}\left(u^\dagger - \alpha_n\nabla f(u^\dagger)\right)\|_2 \\
&\leq \|u_n - \alpha_n\nabla f(u_n) - u^\dagger + \alpha_n\nabla f(u^\dagger)\|_2 \\
&\leq (1 + \alpha_n L)\|u_n - u^\dagger\|_2 \overset{n\to\infty}{\longrightarrow} 0
\end{aligned}$$

Hence $\tilde{u}_n$ also converges to $u^\dagger$. □

Proximal gradient algorithm: remarks

1) It is possible to introduce error terms at each step:

Theorem (proximal gradient algorithm [4])

Let $\epsilon > 0$. IF $f : \mathbb{R}^d \to \mathbb{R}$ is convex with Lipschitz continuous gradient (L), $g : \mathbb{R}^d \to \bar{\mathbb{R}}$ is convex, proper, lower semi-continuous, and a minimizer of $F(u) = f(u) + g(u)$ exists, THEN the proximal gradient algorithm:

$$\begin{cases} \tilde{u}_{n+1} &= \text{prox}_{\alpha_n g}\left(u_n - \alpha_n(\nabla f(u_n) + \delta_n)\right) + \epsilon_n \\ u_{n+1} &= (1-\lambda_n)u_n + \lambda_n \tilde{u}_{n+1} \end{cases} \tag{38}$$

with $u_0 =$ arbitrary, $\epsilon \leq \alpha_n \leq 2/L - \epsilon$, $\epsilon \leq \lambda_n \leq 1$, $\sum_n \|\delta_n\|_2 < \infty$ and $\sum_n \|\epsilon_n\|_2 < \infty$ converges to a minimizer of $F(u)$.

Proof: exercise. □

2) Theorem also holds in (infinite-dimensional) Hilbert space, see [4].

Special case: sparse recovery

Choose $f(u) = \frac{1}{2}\|Ku - y\|_2^2$ and $g(u) = \mu\|u\|_1$, i.e.:

$$\hat{u} = \arg\min_u \frac{1}{2}\|Ku - y\|_2^2 + \mu\|u\|_1$$

- $\nabla f(u) = K^T(Ku - y)$, with $L = \sigma_{\max}(K)^2 = \|K\|^2$
- $\text{prox}_{\alpha g} = S_{\mu\alpha}$, i.e. $S_{\mu\alpha}(u)_i = \begin{cases} 0 & |u_i| \leq \mu\alpha \\ u_i - \mu\alpha\,\text{sgn}(u_i) & |u_i| \geq \mu\alpha. \end{cases}$
- The proximal gradient algorithm reduces to:

$$u_{n+1} = (1 - \lambda_n)u_n + \lambda_n S_{\alpha_n\mu}\left(u_n - \alpha_n K^T(Ku_n - y)\right) \tag{39}$$

- E.g. $\lambda_n = 1$ and $\alpha_n = \alpha$ with $0 < \alpha < 2/L$:

$$u_{n+1} = S_{\alpha\mu}\left(u_n - \alpha K^T(Ku_n - y)\right) \tag{40}$$

So-called “iterative soft-thresholding algorithm” (ISTA) [5]

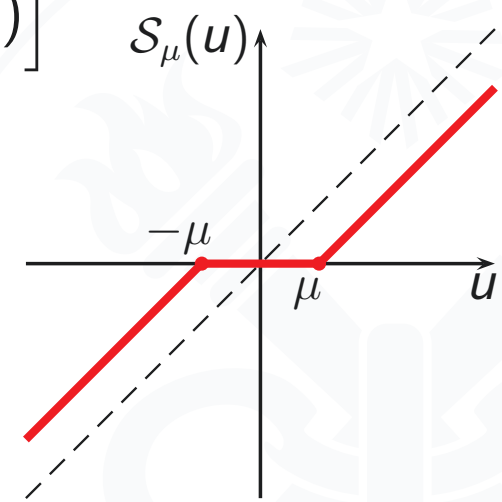
Iterative soft-thresholding algorithm (1)

- Iterative algorithm for finding minimizer of $F(u) \equiv \frac{1}{2}\|Ku - y\|_2^2 + \mu\|u\|_1$:

$$u_{n+1} = \mathcal{S}_{\alpha\mu}\left[u_n + \alpha K^T(y - Ku_n)\right]$$

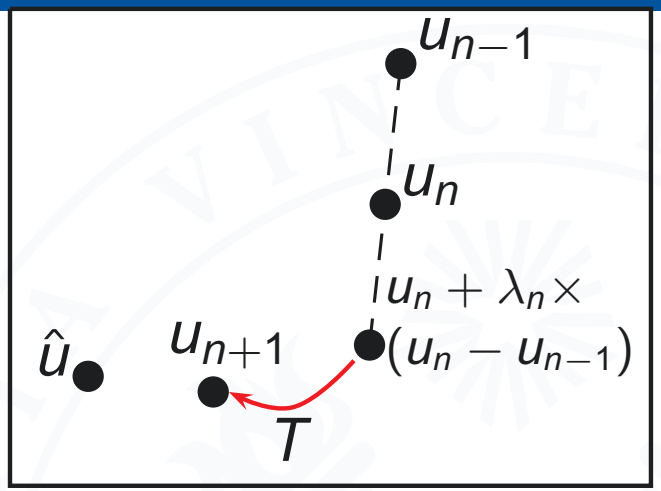
with $\mathcal{S}_\mu =$ component-wise soft-thresholding:

$$\mathcal{S}_\mu(u) = \begin{cases} u - \mu & u \geq \mu \\ 0 & |u| \leq \mu \\ u + \mu & u \leq -\mu \end{cases}$$



- Properties:
  1. Simple
  2. Converges for $\alpha < 2/\|K\|^2$
  3. Soft-thresholding guarantees sparsity of u_n at every iteration
  4. $F(u_n) - F(\hat{u}) \leq \dfrac{\|u_0 - \hat{u}\|_2^2}{2n} \qquad \forall n > 0$
- (Other algorithms exist as well)
- [5, 4]

Iterative soft-thresholding algorithm (2)



- ISTA can be slow
- Improvement (FISTA):

$$u_{n+1} = T(u_n + \lambda_n(u_n - u_{n-1}))$$

with same

$$T(u) \equiv \mathcal{S}_{\alpha\mu}\left[u + \alpha K^T(y - Ku)\right] \quad \text{and} \quad \lambda_n = \frac{n-1}{n+2}.$$

NB: $u_n + \lambda_n(u_n - u_{n-1})$ is not a convex combination of u_n and u_{n-1}

- Advantages:
  1. Simple
  2. Works for $\alpha < 1/\|K\|^2$
  3. $F(u_n) - F(\hat{u}) \leq \dfrac{4\|u_0 - \hat{u}\|_2^2}{(n+1)^2} \qquad \forall n > 0$
  4. Optimal (in some sense)
- See [1, 8]

Special case: Gradient projection algorithm

- $g(u) = I_C(u)$ (indicator function of a closed convex set C) for constrained optimization problem

$$\hat{u} = \arg\min_{u \in C} f(u) = \arg\min_{u} f(u) + g(u)$$

- $\text{prox}_{\alpha_n g} = P_C$ (projection)
- The proximal gradient algorithm reduces to:

$$u_{n+1} = (1 - \lambda_n)u_n + \lambda_n P_C\left(u_n - \alpha_n \nabla f(u_n)\right) \qquad (41)$$

  $u_0 =$ arbitrary, $\epsilon \leq \alpha_n \leq 2/L - \epsilon$, $\epsilon \leq \lambda_n \leq 1$
- Other step-length selection schemes are possible. E.g. one can show that $\forall \alpha_n > 0$ in iteration (41) there exists λ_n such that:

$$f(u_n) - f(u_{n+1}) \geq -\sigma\lambda_n \left[\langle \nabla f(u_n), u_n - \tilde{u}_{n+1}\rangle\right] > 0 \qquad (0 < \sigma < 1)$$

  ("Armijo step-length selection rule").
  In this way, α_n can be chosen freely to accelerate convergence, while λ_n is chosen to guarantee convergence [2, 11, 10]

Special case: Gradient projection algorithm

- $g(u) = I_C(u)$ (indicator function of a closed convex set C) for constrained optimization problem

$$\hat{u} = \arg\min_{u \in C} f(u) = \arg\min_{u} f(u) + g(u)$$

- $\text{prox}_{\alpha_n g} = P_C$ (projection)
- The proximal gradient algorithm reduces to:

$$u_{n+1} = (1 - \lambda_n)u_n + \lambda_n P_C\left(u_n - \alpha_n \nabla f(u_n)\right) \tag{41}$$

$u_0 =$ arbitrary, $\epsilon \leq \alpha_n \leq 2/L - \epsilon$, $\epsilon \leq \lambda_n \leq 1$

- Other step-length selection schemes are possible. E.g. one can show that $\forall \alpha_n > 0$ in iteration (41) there exists λ_n such that:

$$f(u_n) - f(u_{n+1}) \geq -\sigma\lambda_n \left[\langle \nabla f(u_n), u_n - \tilde{u}_{n+1}\rangle\right] > 0 \qquad (0 < \sigma < 1)$$

("Armijo step-length selection rule").
In this way, α_n can be chosen freely to accelerate convergence, while λ_n is chosen to guarantee convergence [2, 11, 10]

Special case: Gradient projection algorithm

- $g(u) = I_C(u)$ (indicator function of a closed convex set C) for constrained optimization problem

$$\hat{u} = \arg\min_{u \in C} f(u) = \arg\min_{u} f(u) + g(u)$$

- $\text{prox}_{\alpha_n g} = P_C$ (projection)
- The proximal gradient algorithm reduces to:

$$u_{n+1} = (1 - \lambda_n)u_n + \lambda_n P_C\left(u_n - \alpha_n \nabla f(u_n)\right) \tag{41}$$

  $u_0 =$ arbitrary, $\epsilon \leq \alpha_n \leq 2/L - \epsilon$, $\epsilon \leq \lambda_n \leq 1$
- Other step-length selection schemes are possible. E.g. one can show that $\forall \alpha_n > 0$ in iteration (41) there exists λ_n such that:

$$f(u_n) - f(u_{n+1}) \geq -\sigma\lambda_n \left[\langle \nabla f(u_n), u_n - \tilde{u}_{n+1}\rangle\right] > 0 \qquad (0 < \sigma < 1)$$

  ("Armijo step-length selection rule").
  In this way, α_n can be chosen freely to accelerate convergence, while λ_n is chosen to guarantee convergence [2, 11, 10]

Second part

- Discuss iterative algorithm for the problem

$$f(u) + g(Au) \tag{42}$$

  where f is convex with Lipschitz continuous gradient, g is convex and A is a linear map,

- using only knowledge of ∇f, A and prox_g,

- but without knowledge of $\text{prox}_{g(A\cdot)}$!

Acknowledgements

- Thanks to organizers.
- Thanks to collaborators
  - Hoan-Phung Bui,
  - Federica Porta,
  - Caroline Verhoeven.

Bibliography I

[1] Amir Beck and Marc Teboulle.
A fast iterative shrinkage-threshold algorithm for linear inverse problems.
SIAM Journal on Imaging Sciences, 2:183–202, 2009.

[2] Dimitri P. Bertsekas.
Nonlinear programming.
Athena Scientific, second edition, 1999.

[3] Alfred M. Bruckstein, David L. Donoho, and Michael Elad.
From sparse solutions of systems of equations to sparse modeling of signals and images.
SIAM Review, 51(1):34–81, 2009.

[4] Patrick L. Combettes and Valerie R. Wajs.
Signal recovery by proximal forward-backward splitting.
Multiscale Model. Simul., 4(4):1168–1200, January 2005.

[5] I. Daubechies, M. Defrise, and C. De Mol.
An iterative thresholding algorithm for linear inverse problems with a sparsity constraint.
Communications On Pure And Applied Mathematics, 57(11):1413–1457, November 2004.

[6] D. L. Donoho.
For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution.
Comm. Pure Appl. Math., 59:797–829, 2006.

[7] J. B. Hiriart-Urruty and C. Lemarechal.
Convex analysis and minimization algorithms.
Springer, 1993.

[8] Yu E. Nesterov.
A method for solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$.
Soviet Math. Dokl., 27:372–376, 1983.

Bibliography II

[9] Neal Parikh and Stephen Boyd.
Proximal algorithms.
Foundations and Trends in Optimization, 1:123–231, 2014.

[10] Federica Porta and Ignace Loris.
On some steplength approaches for proximal algorithms.
2014.

[11] P. Tseng and S. Yun.
A coordinate gradient descent method for nonsmooth separable minimization.
Math. Program. Ser. B, 117:387–423, 2009.