The mathematics of imaging

A unifying representer theorem for inverse problems and machine learning

Quentin Duchemin

April 2019

Introduction

Michael Unser in his paper *A unifying representer theorem for inverse problems and machine learning* adopts a general framework to deal with a particular form of optimization problem. He is able to give a specific formulation of the solution(s) of this problem and gives some conditions providing the uniqueness of the solution.
Michael Unser presents the depth of his result making the link with several widely used machine learning techniques like compressed sensing, super resolution or kernel methods. Thus, M. Unser offers us :

- a higher understanding of the link between different machine learning methods.
- a general framework allowing to write a really clever proof, getting rid of notion of Gâteaux differentiability required in some particular cases.

In a first part, I will present the article adding some personal comments to fully understand the theorem. In a second and last part, I will describe my personal work on the article. I will give detailed proofs (that are not present in the article) and I will discuss some points of the theorem providing examples.

Contents

1 Presentation of the research paper

1.1 The general representer theorem

We consider $\mathcal{X}$ a Banach space. M. Unser is interested in wide class optimization problems, which contains in particular the classical following one :

$$\underset{f\in\mathcal{X}'}{\text{argmin}} \quad \|f\|_{\mathcal{X}'} \qquad s.t.\ \langle \nu_m, f\rangle = y_m,\ \ m=1,\dots,M,$$

which admits the equivalent form

$$\underset{f\in\mathcal{X}'}{\text{argmin}} \quad \sum_{m=1}^{M} |y_m - \langle \nu_m, f\rangle|^2 + \lambda\|f\|^p_{\mathcal{X}'},$$

for an adequate choice of hyper-parameters $\lambda \in \mathbb{R}_+$ and $p \geq 1$.
The proof of the main result of the paper is based on the notion of conjugate pair $(f, f^*) \in \mathcal{X}\times\mathcal{X}'$. We say that (f, f^*) forms a conjugate pair if $\|f^*\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}}$ and if $\langle f^*, f\rangle = \|f^*\|_{\mathcal{X}'}\|f\|_{\mathcal{X}}$. M. Unser also introduces the notion of duality mapping which is defined for all $f\in\mathcal{X}$ as $J(f) = \{f^*\in\mathcal{X}' : (f,f^*) \text{ is a conjugate pair}\}$.

The notions of conjugate and duality mapping (and their properties) are the keys for the proof of the general representer theorem of the paper which takes the following form :

Theorem 1.1. *We consider the following settings :*

- $(\mathcal{X}, \mathcal{X}')$ *is a dual pair of Banach spaces*
- $\mathcal{N}_\nu = span\{\nu_m\}_{m=1}^M \subset \mathcal{X}$ *with the* ν_m *being linearly independent*
- $\nu : \mathcal{X}' \to \mathbb{R}^M\ :\ f \mapsto (\langle\nu_1, f\rangle, \dots, \langle\nu_M, f\rangle)$ *is the linear measurement operator*
- $E : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}^+ \cup \{+\infty\}$ *is a proper and strictly convex loss functional*
- $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ *is an increasing convex function*

Then, for any $y \in \mathbb{R}^M$*, the solution set of the generic optimization problem*

$$S = \underset{f\in\mathcal{X}'}{arg\ min}\, E(y, \nu(f)) + \phi\left(\|f\|_{\mathcal{X}'}\right) \tag{1}$$

is non-empty, convex and weak- compact and such that any solution* $f_0 \in S \subset \mathcal{X}'$ *is a* $(\mathcal{X}', \mathcal{X})$*-conjugate of a common*

$$\nu_0 = \sum_{m=1}^{M} a_m \nu_m \in \mathcal{N}_\nu \subset \mathcal{X},$$

with a suitable set of weights $a \in \mathbb{R}^M$*; i.e.* $S \subset J(\nu_0)$.
If $\mathcal{X}$ *is reflexive and strictly convex and* $f \mapsto \phi(\|f\|_{\mathcal{X}'})$ *is strictly convex, then the solution is unique with* $f_0 = \nu_0^* \in \mathcal{X}'$ *(Banach conjugate of* ν_0*) and* $\nu_0 = f_0^* = (\nu_0^*)^* \in \mathcal{X}$*. In particular, if* $\mathcal{X}$ *is a Hilbert space, then* $f_0 = \sum_{m=1}^M a_m \nu_m^*$ *where* ν_m^* *is the Riesz conjugate of* ν_m*.*

1.2 Comments on the general representer theorem

• Finding the solution when it is unique: How to deal with the weights $a \in \mathbb{R}^M$ of the general representer theorem ?
The theorem 1.1 provides the general form of the solution(s) of the optimization problem (1), however it does not give any clue on how to find the right weights $a \in \mathbb{R}^M$. This is due to the fact that we are considering a really general framework. When we know that the solution of (1) is unique, the example of applications of the theorem in the final section of the paper allow us to understand different situations that can be found :

- In the framework of Kernel Methods, the theorem 1.1 allows us to perform a stunning reduction of the dimension of the problem. Indeed, we can go from an optimization problem on an infinite dimension set to an optimization problem over a finite dimensional set which can thus be solved using standard techniques.
- In the Tikhonov regularization problem, a can be computed with a closed form expression.

• **When uniqueness is not guaranteed, the general representer theorem is a tool to characterize the set of solutions S.**
In the two previous cases, we knew that there existed a unique solution to (1). When we are considering a framework where the uniqueness is not guaranteed, using the Krein Milman theorem and the theorem 1.3 [1] allows us to describe the set of solutions S. Indeed :

- applying the Krein Milmamn theorem with $Y = \mathcal{X}'$ (endowed with the weak-* topology) and $K = S$ ensures that S is the convex hull of its extreme points. This is valid since S is weak-* compact and convex (from theorem 1.1).
- and the extreme points of the set S are given by the theorem 1.3.

Theorem 1.2. *Krein Milman theorem*
Let Y be a locally convex topological vector space (assumed to be Hausdorff), and let K be a compact convex subset of Y. Then, K is the closed convex hull of its extreme points.

Theorem 1.3. *All extremal points f_0 of the solution set S of Problem (1) can be expressed as*

$$f_0 = \sum_{k=1}^{K_0} a_k e_k,$$

for some $1 \leq K_0 \leq M$ where the e_k are some extremal points of the unit "regularization" ball $B_{\mathcal{X}} = \{f \in \mathcal{X} : \|f\|_{\mathcal{X}'} \leq 1\}$ and $a = (a_1, \dots, a_{K_0}) \in \mathbb{R}^{K_0}$ is a vector of appropriate weights.

The general representer theorem can allow to understand why a l_1 regularization term can induce sparsity in the solution of the optimization problem (1). Indeed, when $\mathcal{X}' = l_1(\mathbb{Z})$, the extreme vectors are intrinsically sparse since there are equals to $e_k = (\pm\delta[n - n_k])_{n\in\mathbb{N}}$ for some fixed offset $n_k \in \mathbb{Z}$. Here, $\delta[.]$ denotes the Kronecker impulse which is such that $\delta[0] = 1$ and $\delta[n] = 0$ for $n \neq 0$.

Michael Unser also takes the example of the super resolution framework where the theorem 1.3 can also be used to show that the extreme points of the set S are sparse.

2 Personal work

2.1 Proofs for the duality mapping properties

The proof of the general representer theorem relies deeply on the notion of conjugate and duality mapping. The paper lists the properties of the duality mapping that are relevant for the theorem 1.1 but no proofs are given. I have decided to find the proofs of those essential properties given by the theorem 2.1 in order to have a strong understanding of the theoretical concepts necessary to build the proof of the general representer theorem.

Theorem 2.1. *Duality mapping*
Let $(\mathcal{X}, \mathcal{X}')$ be a dual pair of Banach spaces. Then, the following holds:

1. *Every $f \in \mathcal{X}$ admits at least one conjugate $f^* \in \mathcal{X}'$.*
2. *$J(\lambda f) = \lambda J(f)$ for any $\lambda \in \mathbb{R}$.*
3. *For every $f \in \mathcal{X}$, the set $J(f)$ is convex and weak-* closed in $\mathcal{X}'$.*
4. *The duality mapping is single-valued if $\mathcal{X}'$ is strictly convex; the latter condition is also necessary if $\mathcal{X}$ is reflexive.*
5. *When $\mathcal{X}$ is reflexive, then the duality map is bijective if and only if both $\mathcal{X}$ and $\mathcal{X}'$ are strictly convex.*

[1] which is a consequence of the work of Boyer et al. 2018

Proof.

1. Let $f \in \mathcal{X}$. We consider the point $g = f\|f\|_{\mathcal{X}}$. Using the Hahn-Banach theorem, there exists $g^* \in \mathcal{X}'$ with $\|g^*\|_{\mathcal{X}'} = 1$ $(*)$ and $\langle g^*, g\rangle = \|g\|_{\mathcal{X}}$ $(*)$. Now, if we define $h := g^*\|f\|_{\mathcal{X}} \in \mathcal{X}'$, we have that h satisfies $\|h\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}}$ (using $(*)$) and $\langle h, f\rangle = \langle g^*, g\rangle = \|g\|_{\mathcal{X}} = \|f\|^2_{\mathcal{X}}$ (using $(*)$). Hence, $h \in J(f)$.

2. • First, we prove that $\forall f \in \mathcal{X},\ \forall \lambda > 0,\ J(\lambda f) = \lambda J(f)$.

   Let $f \in \mathcal{X}$ and $\lambda > 0$.

   - Let $g \in J(f)$. We have that

   $$\|\lambda g\|_{\mathcal{X}'} = \lambda\|g\|_{\mathcal{X}'} \underbrace{=}_{\text{since } g\in J(f)} \lambda\|f\|_{\mathcal{X}} = \|\lambda f\|_{\mathcal{X}}, \quad \text{and} \quad \langle \lambda g, \lambda f\rangle = \lambda^2\langle g, f\rangle \underbrace{=}_{\text{since } g\in J(f)} \lambda^2\|g\|_{\mathcal{X}'}\|f\|_{\mathcal{X}} = \|\lambda g\|_{\mathcal{X}'}\|\lambda f\|_{\mathcal{X}}.$$

   We deduce that $\lambda g \in J(\lambda f)$. Hence, $\lambda J(f) \subset J(\lambda f)$.

   - Let $g \in J(\lambda f)$. We define $h := \frac{g}{\lambda}$.

   $$\|h\|_{\mathcal{X}'} = \frac{1}{\lambda}\|g\|_{\mathcal{X}'} \underbrace{=}_{\text{since } g\in J(\lambda f)} \frac{1}{\lambda}\|\lambda f\|_{\mathcal{X}} = \|f\|_{\mathcal{X}}, \quad \text{and} \quad \langle h, f\rangle = \frac{1}{\lambda^2}\langle g, \lambda f\rangle \underbrace{=}_{\text{since } g\in J(\lambda f)} \frac{1}{\lambda^2}\|g\|_{\mathcal{X}'}\|\lambda f\|_{\mathcal{X}} = \|h\|_{\mathcal{X}'}\|f\|_{\mathcal{X}}.$$

   We deduce that $h = \frac{g}{\lambda} \in J(f)$. Hence, $J(\lambda f) \subset \lambda J(f)$.

   • Now, we prove that $\forall f \in \mathcal{X},\ J(-f) = -J(f)$.

   Let $f \in \mathcal{X}$ and $g \in J(f)$.

   $$\|-g\|_{\mathcal{X}'} = \|g\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}} = \|-f\|_{\mathcal{X}}, \quad \textit{and} \quad \langle -g, -f\rangle = (-1)^2\langle g, f\rangle = \|g\|_{\mathcal{X}'}\|f\|_{\mathcal{X}}.$$

   We deduce that $-g \in J(-f)$. Hence, $-J(f) \subset J(-f)$. The converse is also true and can be again proved with basic manipulations.

3. <u>Convexity of $J(f)$ for $f \in \mathcal{X}$</u>

   Let $f \in \mathcal{X}$, $g, h \in J(f)$ and $\alpha \in [0,1]$. We have

   - $\|\alpha g + (1-\alpha)h\|_{\mathcal{X}'} \underbrace{\leq}_{\text{using the triangle inequality}} \alpha\|g\|_{\mathcal{X}'} + (1-\alpha)\|h\|_{\mathcal{X}'} = \alpha\|f\|_{\mathcal{X}} + (1-\alpha)\|f\|_{\mathcal{X}} = \|f\|_{\mathcal{X}}$
   - $\|\alpha g + (1-\alpha)h\|_{\mathcal{X}'} = \sup_{l\in\mathcal{X},\ \|l\|_{\mathcal{X}}=1} \langle \alpha g + (1-\alpha)h, l\rangle \geq \langle \alpha g + (1-\alpha)h, \frac{f}{\|f\|_{\mathcal{X}}}\rangle$

     $= \frac{\alpha}{\|f\|_{\mathcal{X}}}\langle g, f\rangle + \frac{1-\alpha}{\|f\|_{\mathcal{X}}}\langle h, f\rangle$ and using the fact that $g, h \in J(f)$ we have $\langle g, f\rangle = \langle h, f\rangle = \|f\|^2_{\mathcal{X}}$ $(*)$

     $= \frac{\alpha}{\|f\|_{\mathcal{X}}}\|f\|^2_{\mathcal{X}} + \frac{1-\alpha}{\|f\|_{\mathcal{X}}}\|f\|^2_{\mathcal{X}} = \|f\|_{\mathcal{X}}$

   Hence, $\|\alpha g + (1-\alpha)h\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}}$.

   Moreover, $\langle \alpha g + (1-\alpha)h, f\rangle = \alpha\langle g, f\rangle + (1-\alpha)\langle h, f\rangle \underbrace{=}_{\text{using } (*)} \alpha\|f\|^2_{\mathcal{X}} + (1-\alpha)\|f\|^2_{\mathcal{X}} = \|f\|^2_{\mathcal{X}} = \|\alpha g + (1-\alpha)h\|_{\mathcal{X}'}\|f\|_{\mathcal{X}}$ where we used what we have just shown for the last equality.

   We deduce that $J(f)$ is a convex subset of $\mathcal{X}'$.

   <u>Weak-* closeness of $J(f)$ for $f \in \mathcal{X}$</u>

   Now we consider $(g_n)_{n\in\mathbb{N}} \in J(f)^{\mathbb{N}}$ weakly-* converging towards $g \in \mathcal{X}'$. Let's prove that $g \in J(f)$.

- Let's prove that $\|g\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}}$.
  For all $n \in \mathbb{N}$ and for all $s \in \mathcal{X}$ we have $\langle g_n, s\rangle \leq \sup_{l\in\mathcal{X}\ \|l\|_{\mathcal{X}}=1} \langle g_n, l\rangle = \|g_n\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}}$ since $g_n \in J(f)$. Taking the limit when $n \to +\infty$, we get that $\lim_{n\to+\infty} \langle g_n, s\rangle \leq \|f\|_{\mathcal{X}}$. By taking the supremum over $s \in \mathcal{X}$ such that $\|s\|_{\mathcal{X}} = 1$ we finally have

$$\|g\|_{\mathcal{X}'} = \sup_{l\in\mathcal{X}\ \|l\|_{\mathcal{X}}=1} \langle g, l\rangle = \sup_{s\in\mathcal{X}\ \|s\|_{\mathcal{X}}=1} \lim_{n\to+\infty} \langle g_n, s\rangle \leq \|f\|_{\mathcal{X}}. \tag{2}$$

  The other inequality can be computed more easily. First we remark that for all $n \in \mathbb{N}$, since $g_n \in J(f)$, we have that $\langle g_n, f\rangle = \|f\|^2_{\mathcal{X}}$. This remark allows us to find the desired lower bound on $\|g\|_{\mathcal{X}'}$.

$$\|g\|_{\mathcal{X}'} = \sup_{l\in\mathcal{X}\ \|l\|_{\mathcal{X}}=1} \langle g, l\rangle \geq \langle g, \frac{f}{\|f\|_{\mathcal{X}}}\rangle = \lim_{n\to+\infty} \langle g_n, \frac{f}{\|f\|_{\mathcal{X}}}\rangle = \frac{1}{\|f\|_{\mathcal{X}}} \lim_{n\to+\infty} \langle g_n, f\rangle = \|f\|_{\mathcal{X}}. \tag{3}$$

  From (2) and (3), we can conclude that $\|g\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}}$.
- $\langle g, f\rangle = \lim_{n\to+\infty} \langle g_n, f\rangle = \|f\|^2_{\mathcal{X}}$ since for all $n \in \mathbb{N}$, $g_n \in J(f)$. Using the fact that we proved $\|g\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}}$, we have $\langle g, f\rangle = \|g\|_{\mathcal{X}'}\|f\|_{\mathcal{X}}$.

We deduce from the two previous items that $g \in J(f)$. We have just proved that $J(f)$ is weak-* closed in $\mathcal{X}'$.

4. We suppose that $\mathcal{X}'$ is strictly convex.

   Let $f \in \mathcal{X}$, $\alpha \in (0,1)$ and $g, h \in J(f)$.

   Without loss of generality, we suppose that $\|f\|_{\mathcal{X}} = 1$. Using the point 3., we know that $\alpha g + (1-\alpha)h \in J(f)$ since $J(f)$ is convex and $g, h \in J(f)$. We then deduce that $\|\alpha g + (1-\alpha)h\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}} = 1$ $(*)$.

   Since $g, h \in J(f)$, we also have that $\|g\|_{\mathcal{X}'} = \|h\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}} = 1$. Moreover $\mathcal{X}'$ is assumed strictly convex. We deduce that if g is different from h, we would have that $\|\alpha g + (1-\alpha)h\|_{\mathcal{X}'} < 1$, which would be in contradiction with the equality $(*)$.

   Hence, we deduce that $\forall f \in \mathcal{X}$, $J(f)$ contains at most one element. Using the point 1., we conclude the J is single-valued.

   We suppose that $\mathcal{X}$ is reflexive and that J is single-valued

Lemma 2.1. *Let $\mathcal{F}$ be a Banach space. If any $g \in \mathcal{F}'$ assumes its supremum at most in one point of the unit ball, then $\mathcal{F}$ is strictly convex.*

Proof. We will proof the contrapositive. Suppose that $\mathcal{F}$ is not strictly convex. Then there exists $f_1, f_2 \in \mathcal{X}$ and $\alpha \in (0,1)$ so that $f_1 \neq f_2$, $\|f_1\|_{\mathcal{X}} = \|f_2\|_{\mathcal{X}} = 1$ and $\|\alpha f_1 + (1-\alpha)f_2\|_{\mathcal{X}} = 1$.

- First we prove that we can consider $\alpha = \frac{1}{2}$ by proving that the segment line $[f_1, f_2]$ is on the unit ball.
  We consider α_0 such that $\alpha < \alpha_0 < 1$. Since we have $\alpha f_1 + (1-\alpha)f_2 = \frac{\alpha}{\alpha_0}(\alpha_0 f_1 + (1-\alpha_0)f_2) + (1 - \frac{\alpha}{\alpha_0})f_2$, we have the following using Cauchy Schwarz inequality :

$$1 = \|\alpha_0 f_1 + (1-\alpha_0)f_2\|_{\mathcal{F}} \leq \frac{\alpha}{\alpha_0}\|\alpha_0 f_1 + (1-\alpha_0)f_2\|_{\mathcal{F}} + 1 - \frac{\alpha}{\alpha_0}.$$

  We deduce that $\|\alpha_0 f_1 + (1-\alpha_0)f_2\|_{\mathcal{F}} \geq 1$ which means that $\|\alpha_0 f_1 + (1-\alpha_0)f_2\|_{\mathcal{F}} = 1$ (since $\|f_1\|_{\mathcal{F}} = \|f_2\|_{\mathcal{F}} = 1$).

  We could be similar computations that for α_0 such that $0 < \alpha_0 < \alpha$, $\alpha_0 f_1 + (1-\alpha_0)f_2$ belongs to the unit ball. We can conclude that the segment $[f_1, f_2]$ is on the unit ball. In particular, we have $\|f_1 + f_2\|_{\mathcal{F}} = 2$ (with $\alpha_0 = \frac{1}{2}$).
- Using the Hahn-Banach theorem, there exists an element $g \in \mathcal{X}'$ such that $\|g\|_{\mathcal{X}'} = 1$ and $\langle g, \frac{f_1+f_2}{2}\rangle = \left\|\frac{f_1+f_2}{2}\right\|_{\mathcal{X}} = 1$. Hence, $\langle g, f_1\rangle + \langle g, f_2\rangle = 2$.
  As $\langle g, f_1\rangle \leq 1$ and $\langle g, f_2\rangle \leq 1$, it follows that $\langle g, f_1\rangle = \langle g, f_2\rangle = \|g\|_{\mathcal{X}'} = 1$. This means that $g \in \mathcal{F}'$ assumes its supremum at least in two different points of the unit ball, which concludes the proof.

□

We go back to the proof of the second part of the item 4. If we suppose by the absurd that $\mathcal{X}'$ is not strictly convex, the lemma 2.1 applied with $\mathcal{F} = \mathcal{X}'$ gives us that there exists $f \in \mathcal{X}''$ and $g, h \in \mathcal{X}'$ with $g \neq h$ and

$$\|g\|_{\mathcal{X}'} = \|h\|_{\mathcal{X}'} = 1 \text{ and } \langle f, g\rangle = \langle f, h\rangle = \|f\|_{\mathcal{X}''}. \tag{4}$$

But, since $\mathcal{X}$ is reflexive, we have that $f \in \mathcal{X}$. Then, the two relations of (4) give us that $\overline{g} := g\|f\|_{\mathcal{X}}$ and $\overline{h} := h\|f\|_{\mathcal{X}}$ both belong to $J(f)$. Since $\overline{g} \neq \overline{h}$ (since $g \neq h$), we find a contradiction with the fact that J is single-valued.

Hence, $\mathcal{X}'$ must be strictly convex.

2.2 Proof of the general representer theorem

The existence of the solution in the general case is obtained using a very usual result on convex optimization. Similarly, the uniqueness of the solution in the case where $\mathcal{X}$ is reflexive and strictly convex and $f \mapsto \phi(\|f\|_{\mathcal{X}'})$ is strictly convex comes from a usual result of convex optimization.

The trickier part of the proof lies in the general expression of the solution(s) (in the general case).

1. A key point of the proof is to describe all the solutions of the optimization problem (1) as the solutions of another optimization problem parametrized by some $z_0 \in \mathbb{R}^M$:

$$S_{z_0} = \underset{f \in \mathcal{X}'}{\arg\min} \|f\|_{\mathcal{X}'} \text{ s.t. } \nu(f) = z. \tag{5}$$

   Stated otherwise, there exists $z_0 \in \mathbb{R}^M$ (unique) such that all the set of solutions of the optimization problem (2) coincides with S_{z_0}. M. Unser briefly justifies this fact explaining that the contrary would lead to a contradiction using strict convexity. I propose to detail here this point which constitutes the cornerstone of the proof.

   Suppose that there exists f_1 and f_2 with $f_1 \neq f_2$ two solutions of the problem (1) such that $z_1 := \nu(f_1) \neq \nu(f_2) := z_2$. We then consider $\alpha \in (0,1)$ and $f := \alpha f_1 + (1-\alpha) f_2$. We will show that the f leads to a value for the objective function of the problem (1) which is strictly lower than the one associated with f_1 and f_2. This will be a contradiction since f_1 and f_2 are minimizers.
   The objective function F of the problem (1) expressed with the element $f \in \mathcal{X}$ is

$$F(f) := E(y, f) + \phi(\|f\|_{\mathcal{X}}) = E\left(y, \alpha f_1 + (1-\alpha) f_2\right) + \phi(\|\alpha f_1 + (1-\alpha) f_2\|_{\mathcal{X}'}).$$

   • Using the convexity of the norm $\|.\|_{\mathcal{X}'}$ and the fact that ϕ is increasing, we have

$$\phi(\|\alpha f_1 + (1-\alpha) f_2\|_{\mathcal{X}'}) \leq \alpha\phi(\|f_1\|_{\mathcal{X}'}) + (1-\alpha)\phi(\|f_2\|_{\mathcal{X}'}).$$

   • Using the strict convexity of the function E, we also have (since $f_1 \neq f_2$ and since $\alpha \in (0,1)$):

$$E(y, \alpha f_1 + (1-\alpha) f_2) < \alpha E(y, f_1) + (1-\alpha) E(y, f_2).$$

   Gathering this two points leads to

$$F(f) = E(y,f) + \phi(\|f\|_{\mathcal{X}'}) < \alpha \underbrace{\left(E(y,f_1) + \phi(\|f_1\|_{\mathcal{X}'})\right)}_{=F(f_1)} + (1-\alpha) \underbrace{\left(E(y,f_2) + \phi(\|f_2\|_{\mathcal{X}'})\right)}_{=F(f_2)}.$$

   Since f_1 and f_2 are minimizers of F, we have $F(f_1) = F(f_2)$. We finally conclude that $F(f) < F(f_1) = F(f_2)$ which contradicts the fact that f_1 and f_2 are minimizers of F.

2. Now we consider the element $z \in \mathbb{R}^M$ such that all solution f of (1) satisfies $\nu(f) = z$. The second step of the proof consists in showing that $S_z \subset J(\nu_0)$ for any extremal element $\nu_0 \in \{g \in \mathcal{N}_\nu \;:\; \lambda(g) = \|\lambda\| \times \|g\|_{\mathcal{X}} \text{ and } \|\lambda\| = \|g\|_{\mathcal{X}}\}$.

2.3 The "predual space assumption" is necessary

In the general representer theorem (GRT), we are looking for a function $f \in \mathcal{X}'$ which minimizes the objective function. This formulation supposes that the space $\mathcal{X}'$ on which we are looking for solution(s) admits a pre-dual space. M. Unser has insisted during the class on this assumption, claiming that it was necessary. Here I propose to give a counter example to confirm his claim. I dig into the literature and I found in Schlegel 2018 a case where the problem doesn't admit a solution of the form given by the GRT when we look for a solution in a space $\mathcal{Y}$ that does not admit a predual space.

We consider $\mathcal{Y} = l_1$ and we define x and w in l_∞ by :

$$\forall i \in \mathbb{N},\ x_i = \begin{cases} \frac{i}{i+1} & \text{if } i \text{ is odd} \\ 0 & \text{otherwise.} \end{cases} \quad \text{i.e } x = \left(\frac{1}{2}, 0, \frac{3}{4}, 0, \dots\right)$$

$$\forall i \in \mathbb{N},\ w_i = \begin{cases} \frac{i}{i+1} & \text{if } i \text{ is even} \\ 0 & \text{otherwise.} \end{cases} \quad \text{i.e } w = \left(0, \frac{2}{3}, 0, \frac{4}{5}, \dots\right)$$

We then consider ν_x and ν_w in $(l_1)' \cong l_\infty$ associated with x and y, i.e. for all $z \in l_1$, $\langle \nu_x, z\rangle = \sum_{n\in\mathbb{N}} z_n x_n$ and $\langle \nu_w, z\rangle = \sum_{n\in\mathbb{N}} w_n z_n$.
We are interesting in the following optimization problem for any $y \in \mathbb{R}^2$:

$$S = \arg\min_{z\in\mathcal{Y}} (y_1 - \langle \nu_x, z\rangle)^2 + (y_2 - \langle \nu_w, z\rangle)^2 + ||z||_{l_1}^2 \tag{6}$$

We are satisfying all the conditions of the GRT, except the fact that $\mathcal{Y}$ admits a predual space. Indeed :

- ν_w and ν_x are linearly independent
- $\nu : l_1 \to \mathbb{R}^2 \ :\ z \mapsto (\langle \nu_x, z\rangle, \langle \nu_w, z\rangle)$ is the linear measurement operator
- $E : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^+ \cup \{+\infty\}$ is such that $E(a,b) = (a_1 - b_1)^2 + (a_2 - b_2)^2$: It is a proper and strictly convex loss functional
- $\phi = ||.||_{l_1}^2 : \mathbb{R}^+ \to \mathbb{R}^+$ is an increasing convex function

However, we have that $||\nu_x||_{l_1'} = ||\nu_y||_{l_1'} = 1$ $(*_1)$, but it is not possible to find a vector $z \in l_1$ with $||\text{z}||_{l_1} = 1$ such that $\langle \nu_x, z\rangle = 1$ or $\langle \nu_w, z\rangle = 1$ $(*_2)$ (by construction of x and w). This implies that **for any $\mathbf{z} \in \mathbf{l_1}$, $\nu_{\mathbf{x}} \notin \mathbf{J(z)}$ and $\nu_{\mathbf{y}} \notin \mathbf{J(z)}$** $(*)$.
(Indeed, with the definition of the duality mapping, we have $J(z) = \{\nu \in (l_1)' \ :\ \langle \nu, z\rangle = ||\nu|| \times ||z|| \text{ and } ||\nu|| = ||z||\}$. Thus using $(*_1)$ we have that $\nu_x \in J(z)$ would imply that $1 = ||\nu_x|| = ||z||$ and so $\langle \nu_x, z\rangle = 1$ which is not possible by $(*_2)$).

By construction of x and w, the result is also true for any linear combination of ν_x and ν_w, i.e.

$$\text{Span}(\nu_x, \nu_w) \cap J(z) = \varnothing, \quad \forall z \in l_1.$$

Let me justify in details this point. We suppose by contradiction that there exists $z \in l_1$ such that there exist $\lambda, \mu \in \mathbb{R}^2$ with $\nu := \lambda\nu_x + \mu\nu_w \in J(z)$.

- First, I compute the norm of ν.

$$\begin{aligned} ||\nu||_{l_1'} &= \sup_{c\in l_1\ ||c||_{l_1}=1} |\langle \nu, c\rangle| \\ &= \sup_{c\in l_1\ ||c||_{l_1}=1} |\lambda\langle \nu_x, c\rangle + \mu\langle \nu_w, c\rangle| \text{ and since } x \text{ and } w \text{ have disjoint supports :} \\ &= \sup_{c\in l_1\ ||c||_{l_1}=1} |\lambda\langle \nu_x, c\rangle| + \sup_{d\in l_1\ ||d||_{l_1}=1} |\mu\langle \nu_w, d\rangle| \\ &= |\lambda| \underbrace{||\nu_x||_{l_1'}}_{=1} + |\mu| \underbrace{||\nu_w||_{l_1'}}_{=1} \\ &= |\lambda| + |\mu| \end{aligned}$$

- Now, we can finally get a contradiction.

  We consider $\tilde{z} = \frac{z}{\|z\|_{l_1}}$.

  Using (∗) (or simply looking at the definition of x and w), we have, since $\|\tilde{z}\|_{l_1} = 1$, that

$$|\langle \nu_x, \tilde{z} \rangle| < 1 \text{ and } |\langle \nu_w, \tilde{z} \rangle| < 1.$$

$$\text{Thus : } |\langle \nu, \tilde{z} \rangle| \underbrace{=}_{\text{since } x \text{ and } w \text{ have disjoint supports.}} |\lambda| \times |\langle \nu_x, \tilde{z} \rangle| + |\mu| \times |\langle \nu_w, \tilde{z} \rangle| < |\lambda| + |\mu|.$$

  Multiplying by $\|z\|_{l_1}$, the last inequality can also be written as

$$|\langle \nu, z \rangle| < \underbrace{(|\lambda| + |\mu|)}_{\|\nu\|'_{l_1}} \|z\|_{l_1} = \|\nu\|_{l_1'} \|z\|_{l_1}.$$

  This strict inequality contradicts the fact that $\nu \in J(z)$. We conclude that

$$\mathrm{Span}(\nu_x, \nu_w) \cap J(z) = \varnothing, \quad \forall z \in l_1.$$

If the GRT was true, it would exist $\nu \in \mathrm{Span}(\nu_x, \nu_w)$ such that for any solution z of the problem (6), (z, ν) would be a conjugate pair. This would imply that $\nu \in \mathrm{Span}(\nu_x, \nu_w) \cap J(z)$ which is absurd since it is the empty set.
Hence, the space in which a solution is searched needs to admit a predual space if we hope the result of the GRT to hold.

Conclusion

The result given by the studied paper is not constructive in the sense that the form of the solution(s) is given without providing a way to compute it. However, the GRT allows us to adopt a more general point of view and gives us the key to make links between different essential theorems of machine learning and inverse problems. The proof of the GRT is very elegant using only properties on Banach spaces and duality.

This work leads me to think about some possible extensions. With the other courses that I followed with this MVA class, I naturally wonder what can be said if we are not optimizing over Banach spaces but over manifolds. M. Unser also concludes giving other directions for future research like the way to deal with pseudo-norm.

References

Boyer, Claire et al. (2018). "On Representer Theorems and Convex Regularization". working paper or preprint. URL: `https://hal.archives-ouvertes.fr/hal-01823135`.

Schlegel, Kevin (2018). *When is there a Representer Theorem? Reflexive Banach spaces.*