A Unifying Representer Theorem for Inverse Problems and Machine Learning

Michael Unser

Eloïse Berthier, April 22, 2019

1 Summary of the article

1.1 Problem statement

We consider general problems of the form:

$$\min_{f \in \mathcal{X}'} E(y, \nu(f)) + \psi(\|f\|_{\mathcal{X}'}) \tag{1}$$

We are looking for a solution f in $\mathcal{X}'$, a Banach space with a predual, that matches some linear measurements $\nu = (\nu_1, ..., \nu_M)$ with data points $y = (y_1, ..., y_M)$. As such, the problem is ill-posed because the search space is possibly infinite dimensional, while we observe only M measurements. Therefore we regularize the solution by controlling its norm with a non decreasing functional ψ.

This class of optimization problems covers a wide variety of machine learning problems as well as inverse problems. For instance, choosing $\nu_m = \delta(\cdot - x_m)$, then $\nu(f) = (f(x_1), ..., f(x_M))$, and we recover the empirical risk minimization framework. ν could also represent Fourier or Radon measurements, or some convolutions, including inverse problems and compressed sensing.

When $\mathcal{X}'$ is an RKHS, $\nu = \big(\delta(\cdot - x_1), ..., \delta(\cdot - x_M)\big)$, and ψ is increasing, the representer theorem for RKHS ensures that any solution lies in the linear span of kernels centered in the observations. Similar, yet weaker, results exist for sparsity inducing regularizations and even when the optimization problem involves measures. Such theorems are valuable in practice because they transform the original problem into a finite dimensional one, enabling the use of efficient solving algorithms.

This article introduces a general reproducing theorem that holds for Banach spaces. It generalizes a number of reproducing properties already known in different research fields. The proof is relatively soft, as it uses rather high level arguments. Most of them rely on general properties of duality mappings between Banach spaces, of which we give a detailed proof in the second section.

1.2 Main results

Consider a dual pair $(\mathcal{X}, \mathcal{X}')$ of Banach spaces. Then $f \in \mathcal{X}$ and $f^* \in \mathcal{X}'$ are conjugates if they have the same norm in their respective spaces and they saturate the duality bound:

$$\begin{cases} \|f\|_{\mathcal{X}} & = \|f^*\|_{\mathcal{X}'} \\ \langle f^*, f\rangle_{\mathcal{X}'\times\mathcal{X}} & = \|f^*\|_{\mathcal{X}'}\|f\|_{\mathcal{X}} \end{cases}$$

We define the duality mapping as the set of conjugates:

$$J(f) = \left\{f^* \in \mathcal{X}' : \|f\|_{\mathcal{X}} = \|f^*\|_{\mathcal{X}'} \text{ and } \langle f^*, f\rangle_{\mathcal{X}'\times\mathcal{X}} = \|f^*\|_{\mathcal{X}'}\|f\|_{\mathcal{X}}\right\}$$

Duality mappings have some properties listed below and proved in section 2.

Theorem 1.1 (Properties of duality mappings)**.** *Let $(\mathcal{X}, \mathcal{X}')$ be a dual pair of Banach spaces. The following properties hold:*

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. *If $\mathcal{X}'$ is strictly convex, the duality mapping is single valued. If $\mathcal{X}$ is reflexive and J is single valued, then $\mathcal{X}'$ is strictly convex.*
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.*

Moreover, it can be easily shown that, when the duality mapping $J : \mathcal{X} \to \mathcal{X}'$ is single valued, it is linear if and only if $\mathcal{X}$ is a Hilbert space. Indeed, the explicit expression of an inner product in $\mathcal{X}$ can be derived from J.

Theorem 1.2 (General Banach representer theorem)**.** *Let $(\mathcal{X}, \mathcal{X}')$ a dual pair of Banach spaces, $\mathcal{N}_\nu = \operatorname{span}(\nu_m)_{m=1}^M \subset \mathcal{X}$, with the ν_m being linearly independent. Let $\nu : \mathcal{X}' \to \mathbb{R}^M, f \mapsto (\langle\nu_1, f\rangle, ..., \langle\nu_M, f\rangle)$ the linear measurement operator, $E : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}_+ \cup \{+\infty\}$ a proper and strictly convex loss functional, and $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ an increasing convex function. Then, for any $y \in \mathbb{R}^M$, the solution set of the problem*

$$S = \operatorname*{argmin}_{f\in\mathcal{X}'} E(y, \nu(f)) + \psi(\|f\|_{\mathcal{X}'})$$

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

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

If $\mathcal{X}$ is reflexive and strictly convex and $f \mapsto \psi(\|f\|_{\mathcal{X}'})$ is strictly convex, then the solution is unique with $f_0 = \nu_0^* \in \mathcal{X}'$ and $\nu_0 = f_0^* = (\nu_0^*)^*$. 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.

Sketch of the proof
From the assumptions on E and ψ, we deduce that $f \mapsto E(y, \nu(f)) + \psi(\|f\|_{\mathcal{X}'})$ is weakly lower semi-continuous, convex and coercive on $\mathcal{X}'$, which guarantees the existence of a solution, and the unicity when $f \mapsto \psi(\|f\|_{\mathcal{X}'})$ is strictly convex.
Because E is strictly convex, there can only exist one common optimal measurement $z = \nu(f_0)$ (if there were two different, both would be suboptimal compared to their mean). The optimization problem can therefore be recast as:

$$S_z = \underset{f \in \mathcal{X}'}{\operatorname{argmin}} \|f\|_{\mathcal{X}'} \text{ s.t. } \nu(f) = z$$

Because the (ν_m) are linearly independent, the above problem is well defined. Hence any $\nu \in \mathcal{N}_\nu$ is uniquely written as $a_1\nu_1 + ... + a_M\nu_M$ and we can define the linear functional for a fixed z

$$\lambda_z : \nu \mapsto \sum_{m=1}^{M} a_m z_m$$

We denote as U the set of all extensions of λ_z to $\mathcal{X}$. Among them, those with minimal norm are those that are norm preserving, which are built using the Hahn-Banach theorem. Then the optimization problem amounts to:

$$f_0 \in S_z = \arg \inf_{f \in U} \|f\|_{\mathcal{X}'} \iff \|f_0\|_{\mathcal{X}'} = \|\lambda_z\|_{\mathcal{N}_\nu'}$$

Since $\mathcal{N}_\nu$ is finite dimensional, it is reflexive and we deduce with simple computations that for any $\nu_0 \in J(\lambda_z)$, then $f_0 \in J(\nu_0)$.
Finally, we have proved that $S_z \subseteq J(\nu_0)$ for any extremal element $\nu_0 \in \mathcal{N}_\nu$ such that $\lambda(\nu_0) = \|\lambda_z\|_{\mathcal{N}_\nu'}\|\nu_0\|_{\mathcal{X}}$ and $\|\nu_0\|_{\mathcal{X}} = \|\lambda_z\|_{\mathcal{N}_\nu'}$. Hence S_z is convex and weak* compact because it is included in the closed ball of $\mathcal{X}'$ of radius $\|f_0\|_{\mathcal{X}'} < \infty$, itself weak* compact by the Banach-Alaoglu theorem.

□

1.3 Applications

When the search space $\mathcal{X}'$ is an RKHS $\mathcal{H}$, the predual space is $\mathcal{X} = \mathcal{H}' = (\mathcal{X}')'$. In the empirical risk minimization framework, we recover Schölkopf's representer theorem. Indeed, $\nu_m^*(x) = R(\delta(. - x_m))(x) = h(x, x_m)$, where h is the reproducing kernel and R the Riesz map from $\mathcal{H}'$ to $\mathcal{H}$.

One particular case is Tikhonov (or ridge) regularization:

$$f_0 = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \sum_{m=1}^{M} |y_m - \langle \nu_m, f \rangle|^2 + \lambda \|f\|_{\mathcal{H}}^2$$

Applying the general representer theorem, the solution has the parametric form:

$$f_0 = \sum_{m=1}^{M} a_m \varphi_m$$

where $\varphi_m = \nu_m^*$. The duality mapping brings the remarkable property $\langle \nu_m, \varphi_n \rangle = \langle \varphi_m, \varphi_m \rangle$, hence, in matrix notations, the same matrices are involved in both the first and second term of the objective function. Therefore we end up with the well-known simple closed form solution of the ridge regression problem.

Other applications include sparsity inducing regularizations for inverse problems and cases where the space is non-reflexive. We give another simple application that is not present in the article in section 2.2.

2 Additional work

2.1 Detailed proof of theorem 1.1

1. Every $f \in \mathcal{X}$ admits at least one conjugate $f^* \in \mathcal{X}'$.
If $f = 0$, then $0 \in J(0)$. Let $f \in \mathcal{X}, f \neq 0$. We define $U = \text{span}(f)$ which is a linear subspace of $\mathcal{X}$ with dimension one. Let $\ell \in U'$ the linear form such that $\forall \alpha \in \mathbb{R}$, $\ell(\alpha f) := \alpha \|f\|_{\mathcal{X}}$. ℓ is bounded on U:

$$\|\ell\|_{U'} = \sup_{\|f\|_{\mathcal{X}} \leq \frac{1}{\alpha}} \alpha \|f\|_{\mathcal{X}} = 1$$

By the Hahn-Banach theorem, there exists some $\tilde{\ell} \in \mathcal{X}'$ extending ℓ. It is such that: $\tilde{\ell}_{|U} = \ell$ and $\|\tilde{\ell}\|_{\mathcal{X}'} = \|\ell\|_{U'} = 1$.
Now consider $f^* := \|f\|_{\mathcal{X}} \tilde{\ell} \in \mathcal{X}'$. Then $\|f^*\|_{\mathcal{X}^p} = \|f\|_{\mathcal{X}} \|\tilde{\ell}\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}}$.
And $\langle f^*, f \rangle_{\mathcal{X}' \times \mathcal{X}} = \|f\|_{\mathcal{X}} \tilde{\ell}(f) = \|f\|_{\mathcal{X}}^2 = \|f\|_{\mathcal{X}} \|f\|_{\mathcal{X}'}$.
We have shown that $f^* \in J(f)$ and so $J(f)$ is non-empty.

2. $J(\lambda f) = \lambda J(f)$ for any $\lambda \in \mathbb{R}$.
For $\lambda = 0$, $J(0) = \{0\}$. Let $\lambda \neq 0$, $f \in \mathcal{X}$. Let $g^* \in \mathcal{X}'$. $g^* \in J(\lambda f)$ if and only if:

$$\begin{aligned} (i) \quad & \|g^*\|_{\mathcal{X}'} = \|\lambda f\|_{\mathcal{X}} \\ (ii) \quad & \langle g^*, \lambda f \rangle_{\mathcal{X}' \times X} = \|g^*\|_{\mathcal{X}'} \|\lambda f\|_{\mathcal{X}} \end{aligned}$$

$(i) \Leftrightarrow \|\frac{1}{\lambda} g^*\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}}$
$(ii) \Leftrightarrow \lambda \langle g^*, f \rangle_{\mathcal{X}' \times X} = |\lambda| \|g^*\|_{\mathcal{X}'} \|f\|_{\mathcal{X}}$.
Multiplying by $1/\lambda^2$, we get $(ii) \Leftrightarrow \langle \frac{1}{\lambda} g^*, f \rangle_{\mathcal{X}' \times X} = \|\frac{1}{\lambda} g^*\|_{\mathcal{X}'} \|f\|_{\mathcal{X}}$
Hence $g^* \in J(\lambda f) \iff \frac{g^*}{\lambda} \in J(f)$. We have proved that $J(\lambda f) = \lambda J(f)$.

3.1. For every $f \in \mathcal{X}$, the set $J(f)$ is convex.
$J(0) = \{0\}$ is convex. Let $f \in \mathcal{X}$ $(f \neq 0)$, $f_1^*, f_2^* \in J(f)$ and $\lambda \in (0, 1)$.
$\|(1-\lambda) f_1^* + \lambda f_2^*\|_{\mathcal{X}'} \leq (1-\lambda) \|f_1^*\|_{\mathcal{X}'} + \lambda \|f_2^*\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}}$, since $f_1^*, f_2^* \in J(f)$.

Also:

$$\begin{aligned}\|(1-\lambda)f_1^* + \lambda f_2^*\|_{\mathcal{X}'} &\leq \frac{1}{\|f\|_{\mathcal{X}}}\langle(1-\lambda)f_1^* + \lambda f_2^*, f\rangle_{\mathcal{X}'\times\mathcal{X}} \\ &= \frac{1}{\|f\|_{\mathcal{X}}}\Big\{(1-\lambda)\|f_1^*\|_{\mathcal{X}'}\|f\|_{\mathcal{X}} + \lambda\|f_2^*\|_{\mathcal{X}'}\|f\|_{\mathcal{X}}\Big\} \\ &= \|f\|_{\mathcal{X}}\end{aligned}$$

Hence $\|(1-\lambda)f_1^* + \lambda f_2^*\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}}$. Besides,
$\langle(1-\lambda)f_1^* + \lambda f_2^*, f\rangle_{\mathcal{X}'\times\mathcal{X}} = (1-\lambda)\|f\|_{\mathcal{X}}^2 + \lambda\|f\|_{\mathcal{X}}^2 = \|f\|_{\mathcal{X}}^2 = \|f\|_{\mathcal{X}}\|(1-\lambda)f_1^* + \lambda f_2^*\|_{\mathcal{X}'}$.
And so $(1-\lambda)f_1^* + \lambda f_2^* \in J(f)$ and $J(f)$ is convex.

3.2. For every $f \in \mathcal{X}$, the set $J(f)$ is weak* closed in $\mathcal{X}'$.
It is obvious for $f = 0$; assume $f \neq 0$. Let $(f_n^*)_{n\in\mathbb{N}}$ be a sequence in $J(f)$ such that $f_n^* \rightharpoonup g^* \in \mathcal{X}'$, where the convergence is with respect to the weak* topology. It means that $\forall h \in \mathcal{X}$, $\langle f_n^*, h\rangle_{\mathcal{X}'\times X} \xrightarrow[n\to+\infty]{} \langle g^*, h\rangle_{\mathcal{X}'\times\mathcal{X}}$.
Let us show that $g^* \in J(f)$.
For $h = f$, $\forall n \geq 0, \langle f_n^*, f\rangle_{\mathcal{X}'\times\mathcal{X}} = \|f_n\|_{\mathcal{X}'}\|f\|_{\mathcal{X}} = \|f\|_{\mathcal{X}}^2$, and so $\|f\|_{\mathcal{X}}^2 = \langle g^*, f\rangle_{\mathcal{X}'\times\mathcal{X}}$.
Hence:

$$\|g^*\|_{\mathcal{X}'} \geq \frac{\langle g^*, f\rangle_{\mathcal{X}'\times\mathcal{X}}}{\|f\|_{\mathcal{X}}} = \|f\|_{\mathcal{X}}$$

On the other hand, for any $h \in \mathcal{X}$, $\|f_n^*\|_{\mathcal{X}'}\|h\|_{\mathcal{X}} \geq \langle f_n^*, h\rangle_{\mathcal{X}'\times\mathcal{X}} \xrightarrow[n\to+\infty]{} \langle g^*, h\rangle_{\mathcal{X}'\times\mathcal{X}}$.
Since $\forall n, f_n^* \in J(f)$, $\|f_n^*\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}}$ and then $\forall h \in \mathcal{X}$, $\|f\|_{\mathcal{X}}\|h\|_{\mathcal{X}} \geq \langle g^*, h\rangle_{\mathcal{X}'\times\mathcal{X}}$.
Thus:

$$\|g^*\|_{\mathcal{X}'} = \sup_{h,\|h\|\neq 0} \frac{\langle g^*, h\rangle_{\mathcal{X}'\times\mathcal{X}}}{\|h\|_{\mathcal{X}'}} \leq \|f\|_{\mathcal{X}}$$

so $\|g^*\|_{\mathcal{X}'} = \|f\|_{\mathcal{X}}$. Besides, $\langle g^*, f\rangle_{\mathcal{X}'\times\mathcal{X}} = \|f\|_{\mathcal{X}}^2 = \|f\|_{\mathcal{X}}\|g^*\|_{\mathcal{X}'}$.
We have proved that g^*, the limit of (f_n^*), is in $J(f)$, hence $J(f)$ is weak* closed in $\mathcal{X}'$.

4.1. If $\mathcal{X}'$ is strictly convex, the duality mapping is single valued.
Let $f \in \mathcal{X}$. If $f = 0$, $J(f)$ is single valued. We now consider $f \neq 0$. Suppose there exists two $f_1^*, f_2^* \in J(f)$ with $f_1^* \neq f_2^*$. Then, using one of the previous results, $\frac{f_1^*}{\|f_1^*\|_{\mathcal{X}'}} = \frac{f_1^*}{\|f\|_{\mathcal{X}}} \in J\left(\frac{f}{\|f\|_{\mathcal{X}}}\right)$ and similarly, $\frac{f_2^*}{\|f\|_{\mathcal{X}}} \in J\left(\frac{f}{\|f\|_{\mathcal{X}}}\right)$. Both also have norm one.
Let $\lambda \in (0,1)$. $\mathcal{X}'$ being strictly convex, we have:

$$\left\|(1-\lambda)\frac{f_1^*}{\|f\|_{\mathcal{X}}} + \lambda\frac{f_2^*}{\|f\|_{\mathcal{X}}}\right\|_{\mathcal{X}'} < 1$$

But since $J\left(\frac{f}{\|f\|_{\mathcal{X}}}\right)$ is convex, $(1-\lambda)\frac{f_1^*}{\|f\|_{\mathcal{X}}} + \lambda\frac{f_2^*}{\|f\|_{\mathcal{X}}} \in J\left(\frac{f}{\|f\|_{\mathcal{X}}}\right)$, and so:

$$\left\|(1-\lambda)\frac{f_1^*}{\|f\|_{\mathcal{X}}} + \lambda\frac{f_2^*}{\|f\|_{\mathcal{X}}}\right\|_{\mathcal{X}'} = \left\|\frac{f}{\|f\|_{\mathcal{X}}}\right\|_{\mathcal{X}} = 1$$

This is a contradiction and yields $f_1^* = f_2^*$, hence $J(f)$ is single valued because at most, and at least one conjugate exists.

4.2. If $\mathcal{X}$ is reflexive and J is single valued, then $\mathcal{X}'$ is strictly convex.
Let $R : \mathcal{X}' \to \mathcal{X}'' = \mathcal{X}$ be the duality mapping from $\mathcal{X}'$ to $\mathcal{X}$. Let $g_1^* \neq g_2^* \in \mathcal{X}'$ with $\|g_1^*\|_{\mathcal{X}'} = \|g_2^*\|_{\mathcal{X}'} = 1$, let $\lambda \in (0,1)$.
Since $\mathcal{X}'$ is a Banach space, $(1-\lambda)g_1^* + \lambda g_2^* \in \mathcal{X}'$. We have proved that $R\big((1-\lambda)g_1^* + \lambda g_2^*\big)$ is non empty (using a previous result with the roles of $\mathcal{X}$ and $\mathcal{X}'$ exchanged in the reflexive case). There exists some $f \in \mathcal{X}$ such that:

$$\|f\|_{\mathcal{X}} = \|(1-\lambda)g_1^* + \lambda g_2^*\|_{\mathcal{X}'}$$
$$\langle f, (1-\lambda)g_1^* + \lambda g_2^* \rangle_{\mathcal{X}\times\mathcal{X}'} = \|f\|_{\mathcal{X}}\|(1-\lambda)g_1^* + \lambda g_2^*\|_{\mathcal{X}'}$$

We have $\|(1-\lambda)g_1^* + \lambda g_2^*\|_{\mathcal{X}'} \leq (1-\lambda)\|g_1^*\|_{\mathcal{X}'} + \lambda\|g_2^*\|_{\mathcal{X}'} = 1$ (g_1^* and g_2^* are unitary).
Suppose $\|(1-\lambda)g_1^* + \lambda g_2^*\|_{\mathcal{X}'} = 1$, we will find a contradiction.
In that case, $\|f\|_{\mathcal{X}} = 1$ and $\langle f, (1-\lambda)g_1^* + \lambda g_2^* \rangle_{\mathcal{X}\times\mathcal{X}'} = 1$, that is:

$$(1-\lambda)\langle f, g_1^* \rangle_{\mathcal{X}\times\mathcal{X}'} + \lambda\langle f, g_2^* \rangle_{\mathcal{X}\times\mathcal{X}'} = 1 \qquad (\star)$$

Since $\langle f, g_1^* \rangle_{\mathcal{X}\times\mathcal{X}'} \leq \|f\|_{\mathcal{X}}\|g_1^*\|_{\mathcal{X}'} \leq 1$ and also $\langle f, g_2^* \rangle_{\mathcal{X}\times\mathcal{X}'} \leq 1$, and $\lambda \in (0,1)$, then for $(\star)$ to hold, we have necessarily:

$$\begin{cases} \langle f, g_1^* \rangle_{\mathcal{X}\times\mathcal{X}'} &= 1 \\ \langle f, g_2^* \rangle_{\mathcal{X}\times\mathcal{X}'} &= 1 \end{cases}$$

Because $\|f\|_{\mathcal{X}} = \|g_1^*\|_{\mathcal{X}'} = \|g_2^*\|_{\mathcal{X}'} = 1$, this means that $g_1^* \in J(f)$ and $g_2^* \in J(f)$. Since $J(f)$ is supposed to be single valued and $g_1^* \neq g_2^*$, this is a contradiction with $\|(1-\lambda)g_1^* + \lambda g_2^*\|_{\mathcal{X}'} = 1$.
Finally: $\|(1-\lambda)g_1^* + \lambda g_2^*\|_{\mathcal{X}'} < 1$ and $\mathcal{X}'$ is strictly convex.

5.1. When $\mathcal{X}$ is reflexive, if the duality map is bijective, then both $\mathcal{X}$ and $\mathcal{X}'$ are strictly convex.
In this case, J is single valued and $\mathcal{X}$ is reflexive, so we have already proved that $\mathcal{X}'$ is strictly convex. We make the same reasoning on J^{-1} which is single valued and $\mathcal{X}'$ is also reflexive, so $\mathcal{X}$ is also strictly convex.

5.2. When $\mathcal{X}$ is reflexive, if both $\mathcal{X}$ and $\mathcal{X}'$ are strictly convex, then the duality map is bijective.
$\mathcal{X}'$ being strictly convex, $J : \mathcal{X} \to \mathcal{X}'$ is single valued. $\mathcal{X}$ being strictly convex, the converse duality map $R : \mathcal{X}' \to \mathcal{X}$ is also single valued. We will show that $R \circ J = Id$.
Let $f \in \mathcal{X}$. $\|f\|_X = \|J(f)\|_{\mathcal{X}'}$ and $\langle J(f), f\rangle_{\mathcal{X}'\times\mathcal{X}} = \|f\|_{\mathcal{X}}\|J(f)\|_{\mathcal{X}'}$ and so $f \in R\big(J(f)\big)$.
Since $R\big(J(f)\big)$ is single valued, $f = R \circ J(f)$.
We may also show that $\forall f^* \in \mathcal{X}'$, $f^* = J \circ R(f^*)$.
Finally $R = J^{-1}$ and J is bijective.

□

2.2 Application to convex neural networks

Empirical risk minimization for one hidden layer neural networks can be recast as a convex problem in the space of signed measures, when one does not assume a fixed number of hidden units. Indeed, we can view this neural network as a measure (the output weights) over the infinite dimensional space of parametric functions (the hidden units). More formally (the notations are taken from [3]), let $\Theta = \mathbb{R}^d$, and for $\theta \in \Theta$,

$$\phi(\theta) : x \mapsto (\sum_{i=1}^{d-1} \theta_i x_i + \theta_d)_+$$

is a function mapping an input vector $x \in \mathfrak{X} \subset \mathbb{R}^{d-1}$ to a scalar (the output of each hidden layer). We consider $\mathcal{M}(\Theta)$ the space of signed measures over Θ. Then the regularized empirical risk minimization problem can be recast as:

$$\min_{\mu \in \mathcal{M}(\Theta)} R\left(\int \phi \mathrm{d}\mu\right) + G(\mu)$$

with R a convex loss function and G a convex regularizer.

An interesting, yet well-known [2, 1] result is that solving this convex problem when $\mathcal{X}$ has finite cardinality n (finite number of samples), and G penalizes the total variation of μ, then each optimal measure is a discrete measure with a most n diracs. This means that a neural network that achieves optimality in this problem has at most n hidden units.

This result is already known, but was proved with arguments from convex optimization. Here it directly results from the "swiss knife" representer theorem. This case is very similar to another one, present in the article (super-resolution localization of spikes). Let Θ be compact domain. The problem is:

$$\min_{\mu \in \mathcal{X}'} E(y, \nu(\mu)) + \lambda \|\mu\|_{\mathcal{X}'}$$

where $\mathcal{X} = \mathcal{C}(\Theta)$ and $\mathcal{X}' = \mathcal{M}(\Theta)$.

If the ν_m are independent measurements, then the general representer theorem ensures that the set of solutions is non-empty, convex and weak*-compact, and any solution $\mu_0 \in \mathcal{X}'$ is the conjugate of a common

$$\nu_0 = \sum_{m=1}^{n} a_m \nu_m$$

The extreme points of the unit ball in $\mathcal{X}'$ being point measures $e_k = \pm\delta(\cdot - u_k)$, then there exists minimizers of the form:

$$\mu_0 = \sum_{k=1}^{K_0} a_k \delta(\cdot - u_k)$$

for some $K_0 \leq n$. In our case, it means precisely that the optimal prediction function can be written using at most K_0 hidden units with hidden weight u_k (a vector with the same size as the input features) and output weight a_k (a scalar).

Knowing that optimal solutions have less hidden units than the number of samples is a first step but is not so interesting by itself. In practice, we know that models that generalize well have much less neurons than the number of samples. We are rather interested in getting an estimate of K_0, depending on the regularity of the generative function that we want to approximate. Given a fixed neural network as a generative model, there are several ways to empirically estimate this K_0:

- incrementally add new neurons until reaching a certain accuracy. This can be achieved using Frank-Wolfe algorithm, but each step is NP hard to compute [1];
- test several overparametrizations and spot a phase transition where the error stops decreasing [3];
- start with a large number of neurons (e.g n or a fraction of it), penalize the output weights with a sparsity promoting regularizer, and compare the goodness of fit at convergence.

Conclusion

The general representer theorem extends a number of already known, but more specific representer theorems. It offers a unifying framework to study a broad class of problems, from machine learning to compressed sensing. The main tools used come from basic properties of duality mappings and enable to build a simple proof for this general result.

References

[1] F. Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

[2] Y. Bengio, N. L. Roux, P. Vincent, O. Delalleau, and P. Marcotte. Convex neural networks. In *Advances in neural information processing systems*, pages 123–130, 2006.

[3] L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*, 2018.

[4] M. Unser. A unifying representer theorem for inverse problems and machine learning. *arXiv preprint arXiv:1903.00687*, 2019.