The Representer Theorem

David S. Rosenberg

New York University

February 13, 2018
Contents

1 Inner Product Spaces and Projections (Hilbert Spaces)

2 Representer Theorem
Inner Product Spaces and Projections (Hilbert Spaces)
An **inner product space** (over reals) is a vector space $\mathcal{V}$ and an **inner product**, which is a mapping

$$\langle \cdot , \cdot \rangle : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$$

that has the following properties $\forall x, y, z \in \mathcal{V}$ and $a, b \in \mathbb{R}$:

- **Symmetry:** $\langle x, y \rangle = \langle y, x \rangle$
- **Linearity:** $\langle ax + by, z \rangle = a \langle x, z \rangle + b \langle y, z \rangle$
- **Positive-definiteness:** $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \iff x = 0$. 
For an inner product space, we define a norm as 

$$\|x\| = \sqrt{\langle x, x \rangle}.$$ 

Example 

$\mathbb{R}^d$ with standard Euclidean inner product is an inner product space: 

$$\langle x, y \rangle := x^T y \quad \forall x, y \in \mathbb{R}^d.$$ 

Norm is 

$$\|x\| = \sqrt{x^T x}.$$
What norms can we get from an inner product?

**Theorem (Parallelogram Law)**

A norm $\| \cdot \|$ can be written in terms of an inner product on $\mathcal{V}$ iff $\forall x, x' \in \mathcal{V}$

$$2\|x\|^2 + 2\|x'\|^2 = \|x + x'\|^2 + \|x - x'\|^2,$$

and if it can, the inner product is given by the **polarization identity**

$$\langle x, x' \rangle = \frac{\|x\|^2 + \|x'\|^2 - \|x - x'\|^2}{2}.$$

**Example**

$\ell_1$ norm on $\mathbb{R}^d$ is NOT generated by an inner product. [Exercise]

Is $\ell_2$ norm on $\mathbb{R}^d$ generated by an inner product?
Definition

Two vectors are **orthogonal** if \( \langle x, x' \rangle = 0 \). We denote this by \( x \perp x' \).

Definition

\( x \) is orthogonal to a set \( S \), i.e. \( x \perp S \), if \( x \perp s \) for all \( x \in S \).
Proof.
We have
\[
\|x + x'|^2 = \langle x + x', x + x' \rangle \\
= \langle x, x \rangle + \langle x, x' \rangle + \langle x', x \rangle + \langle x', x' \rangle \\
= \|x\|^2 + \|x'\|^2.
\]
Choose some $x \in \mathcal{V}$.

Let $M$ be a subspace of inner product space $\mathcal{V}$.

Then $m_0$ is the **projection of $x$ onto $M$**, if $m_0 \in M$ and is the closest point to $x$ in $M$.

In math: For all $m \in M$, 
\[
\|x - m_0\| \leq \|x - m\|.
\]
Projections exist for all finite-dimensional inner product spaces.
We want to allow infinite-dimensional spaces.
Need an extra condition called completeness.
A space is complete if all Cauchy sequences in the space converge.

Definition

A Hilbert space is a complete inner product space.

Example

Any finite dimensional inner product space is a Hilbert space.
The Projection Theorem

Theorem (Classical Projection Theorem)

- $\mathcal{H}$ a Hilbert space
- $M$ a closed subspace of $\mathcal{H}$ (picture a hyperplane through the origin)
- For any $x \in \mathcal{H}$, there exists a unique $m_0 \in M$ for which
  \[ \|x - m_0\| \leq \|x - m\| \quad \forall m \in M. \]
- This $m_0$ is called the [orthogonal] projection of $x$ onto $M$.
- Furthermore, $m_0 \in M$ is the projection of $x$ onto $M$ iff
  \[ x - m_0 \perp M. \]
Projection Reduces Norm

Theorem

Let $M$ be a closed subspace of $\mathcal{H}$. For any $x \in \mathcal{H}$, let $m_0 = \text{Proj}_M x$ be the projection of $x$ onto $M$. Then

$$\|m_0\| \leq \|x\|,$$

with equality only when $m_0 = x$.

Proof.

\[
\|x\|^2 = \|m_0 + (x - m_0)\|^2 \quad \text{(note: $x - m_0 \perp m_0$ by Projection theorem)}
\]
\[
= \|m_0\|^2 + \|x - m_0\|^2 \quad \text{by Pythagorean theorem}
\]
\[
\|m_0\|^2 = \|x\|^2 - \|x - m_0\|^2
\]

Then $\|x - m_0\|^2 \geq 0$ implies $\|m_0\|^2 \leq \|x\|^2$. If $\|x - m_0\|^2 = 0$, then $x = m_0$, by definition of norm.
Representer Theorem
Generalize from SVM Objective

- **SVM objective:**
  \[
  \min_{w \in \mathbb{R}^d} \frac{1}{2} \|w\|^2 + c \sum_{i=1}^{n} \max(0, 1 - y_i \langle w, x_i \rangle).
  \]

- **Generalized objective:**
  \[
  \min_{w \in \mathcal{H}} R(\|w\|) + L(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle),
  \]

  where
  - \( R : [0, \infty) \rightarrow \mathbb{R} \) is nondecreasing (Regularization term)
  - and \( L : \mathbb{R}^n \rightarrow \mathbb{R} \) is arbitrary. (Loss term)
General Objective Function for Linear Hypothesis Space (Details)

- **Generalized objective:**

\[
\min_{w \in H} R(\|w\|) + L(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle),
\]

where

- \( w, x_1, \ldots, x_n \in \mathcal{H} \) for some Hilbert space \( \mathcal{H} \). (We typically have \( \mathcal{H} = \mathbb{R}^d \).)
- \( \| \cdot \| \) is the norm corresponding to the inner product of \( \mathcal{H} \). (i.e. \( \| w \| = \sqrt{\langle w, w \rangle} \))
- \( R : [0, \infty) \to \mathbb{R} \) is nondecreasing (Regularization term), and
- \( L : \mathbb{R}^n \to \mathbb{R} \) is arbitrary (Loss term).
General Objective Function for Linear Hypothesis Space (Details)

- **Generalized objective**: 
  \[ \min_{w \in H} R(\|w\|) + L(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle) , \]

- **What’s “linear”?**
- The prediction/score function \( x \mapsto \langle w, x_i \rangle \) is linear – in what?
  - in parameter vector \( w \), and
  - in the feature vector \( x_i \).

- **Why?** [Real-valued] inner products are linear in each argument.

- **The important part is the linearity in the parameter \( w \).**
General Objective Function for Linear Hypothesis Space (Details)

- **Generalized objective:**

  \[
  \min_{w \in H} R(\|w\|) + L(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle),
  \]

- Ridge regression and SVM are of this form.

- What if we penalize with \(\lambda \|w\|_2\) instead of \(\lambda \|w\|_2^2\)? Yes!

- What if we use lasso regression? No! \(\ell_1\) norm does not correspond to an inner product.
The Representer Theorem

Theorem (Representer Theorem)
Let

\[ J(w) = R(\|w\|) + L(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle), \]

where

- \( w, x_1, \ldots, x_n \in \mathcal{H} \) for some Hilbert space \( \mathcal{H} \). (We typically have \( \mathcal{H} = \mathbb{R}^d \).)
- \( \| \cdot \| \) is the norm corresponding to the inner product of \( \mathcal{H} \). (i.e. \( \|w\| = \sqrt{\langle w, w \rangle} \))
- \( R : [0, \infty) \rightarrow \mathbb{R} \) is nondecreasing (Regularization term), and
- \( L : \mathbb{R}^n \rightarrow \mathbb{R} \) is arbitrary (Loss term).

Then

- If \( M = \text{span}(x_1, \ldots, x_n) \), then \( J(\text{Proj}_M w) \leq J(w) \) for any \( w \in \mathcal{H} \).
- If \( J(w) \) has a minimizer, then it has a minimizer of the form \( w^* = \sum_{i=1}^{n} \alpha_i x_i \).
- If \( R \) is strictly increasing, then all minimizers have this form. (Proof in homework.)
Fix any \( w \in \mathcal{H} \).

Let \( w_M = \text{Proj}_M w \).

Then \( w_M^\perp := w - w_M \) is orthogonal to \( M \).

So \( \langle w, x_i \rangle = \langle w_M + w_M^\perp, x_i \rangle = \langle w_M, x_i \rangle \) \( \forall i \), and

\[
L(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle) = L(\langle w_M, x_1 \rangle, \ldots, \langle w_M, x_n \rangle).
\]

Projections decrease norms: \( \|w_M\| \leq \|w\| \).

Since \( R \) is nondecreasing, \( R(\|w_M\|) \leq R(\|w\|) \).

\( J(w_M) \leq J(w) \). [Proves first result.]

If \( w^* \) minimizes \( J(w) \), then \( w^*_M = \text{Proj}_M w^* \) is also a minimizer, since \( J(w^*_M) \leq J(w^*) \).

So \( \exists \alpha \text{ s.t. } w^*_M = \sum_{i=1}^n \alpha_i x_i \) is a minimizer of \( J(w) \).

Q.E.D.
Theorem

\[ J(w) = R(\|w\|) + L(\langle w, x_1 \rangle, \ldots, \langle w, x_n \rangle), \]

and let \( M = \text{span}(x_1, \ldots, x_n) \). Then under the same conditions given in the Representer theorem, if \( w^*_M \) minimizes \( J(w) \) over the set \( M \), then \( w^*_M \) minimizes \( J(w) \) over all \( \mathcal{H} \).

\(^a\)Thanks to Mingsi Long for suggesting this nice theorem and proof.

- One consequence of the Representer theorem only applies if \( J(w) \) has a minimizer over \( \mathcal{H} \). This theorem tells us that it’s sufficient to check for a constrained minimizer of \( J(w) \) over \( M \). If one exists, then it’s also an unconstrained minimizer of \( J(w) \) over \( \mathcal{H} \). If there is no constrained minimizer over \( M \), then \( J(w) \) has no minimizer over \( \mathcal{H} \) (by the Representer theorem).

- Bottom Line: We can jump straight to minimizing over \( M \), the “span of the data”.
Let $w_M^* \in \arg \min_{w \in M} J(w)$. [the constrained minimizer]

Consider any $w \in \mathcal{H}$.

Let $w_M = \text{Proj}_M w$.

By the Representer theorem, $J(w_M) \leq J(w)$.

$J(w_M^*) \leq J(w_M)$ by definition of $w_M^*$.

Thus for any $w \in \mathcal{H}$, $J(w_M^*) \leq J(w)$.

Therefore $w_M^*$ minimizes $J(w)$ over $\mathcal{H}$.

QED