Characterizing the Representer Theorem

Yan-Liang Yu, Hao Cheng, Dale Schuurmans, Csaba Szepesvári (ICML-2013)

Zoltán Szabó

Gatsby Unit, Tea Talk

October 3, 2013

Motivation

- Kernel methods:
  - many exciting applications (SVMC, SVMR, KPCA, ...).
- Representer theorem:
  - ∞- $\rightarrow$ finite-dimensional problem,
  - Q: Under what conditions does it hold?

- Chosen paper: *equivalent* characterization.

Reproducing kernel Hilbert space (RKHS)

- Given: $\mathcal{X} \neq \emptyset$ set (graphs, time series, distributions, ...).
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *kernel* if $\exists \varphi : \mathcal{X} \to \mathcal{H}$(ilbert) space such that

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}} \quad (\forall x, y \in \mathcal{X}). \tag{1}$$

- $\mathcal{H}$ is not unique, but $\exists! \, \mathcal{H} = \mathcal{H}(k)$ RKHS such that

$$k(x, \cdot) \in \mathcal{H} \qquad (\forall x \in \mathcal{X}), \tag{2}$$

$$\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x) \qquad (\forall x \in \mathcal{X}, \forall f \in \mathcal{H}). \tag{3}$$

Problem setup

- Objective function (regularized empirical risk, $\lambda > 0$; $R :=$):

$$J(f) := L\left(f(x_1), \ldots, f(x_n)\right) + \lambda\Omega\left(f\right) \rightarrow \min_{f \in \mathcal{H}}. \tag{4}$$

- Example (SVMR):

$$J(f) = \frac{1}{T}\sum_{i=1}^{n} \left|y_i - f(x_i)\right|_{\epsilon} + \lambda \left\|f\right\|^2_{\mathcal{H}(k)} \rightarrow \min_{f \in \mathcal{H}(k)}. \tag{5}$$

- Representer theorem [Kimeldorf & Wahba '71]: Solutions of (4) for $\mathcal{H} = \mathcal{H}(k)$ and $\Omega(f) = \|f\|^2_{\mathcal{H}}$ take the form

$$f(\cdot) = \sum_{i=1}^{n} a_i k(x_i, \cdot) \quad (\mathbf{a} \in \mathbb{R}^n). \tag{6}$$

Consequence ($\Omega(f) = \|f\|^2_{\mathcal{H}(k)}$)

$$f(x_j) = \sum_{i=1}^{n} a_i k(x_i, x_j) = (\mathbf{Ka})_j, \quad \mathbf{K} = [k(x_i, x_j)] \in \mathbb{R}^{n\times n} \tag{7}$$

$$\|f\|^2_{\mathcal{H}(k)} = \left\langle \sum_{i=1}^{n} a_i k(x_i, \cdot), \sum_{j=1}^{n} a_j k(x_j, \cdot) \right\rangle_{\mathcal{H}(k)} \tag{8}$$

$$= \sum_{i,j=1}^{n} a_i a_j \left\langle k(x_i, \cdot), k(x_j, \cdot)\right\rangle_{\mathcal{H}(k)} = \sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) \tag{9}$$

$$= \mathbf{a}^T \mathbf{Ka}. \tag{10}$$

Plugging the obtained formulas to the objective:

$$J(\mathbf{a}) = L\left((\mathbf{Ka})_1, \ldots, (\mathbf{Ka})_n\right) + \lambda \mathbf{a}^T \mathbf{Ka} \to \min_{\mathbf{a}\in\mathbb{R}^n}. \tag{11}$$

Schölkopf et al., '01: *Sufficient* condition (RKHS)

- Result:
  - Let $\Omega(f) = h\left(\|f\|_{\mathcal{H}}\right)$, $h : \mathbb{R}^{\geq 0} \to \bar{\mathbb{R}}$.
  - If h is
    - monotonically increasing: then $\exists$,
    - strictly monotonically increasing: then $\forall$

    f minimizers admit the required form.
- *Necessary* conditions?

Necessary conditions (Hilbert space)

If Ω is

- Gateaux differentiable (=directional) [Argyriou et al. '09], or
- lower semi-continuous [Dinuzzo & Schölkopf '12],

then the *sufficient* condition is also *necessary*.

Chosen paper: idea – interpolation

- Let us consider the interpolation problem ($I :=$):

$$\min_{f \in \mathcal{H}: \langle f, f_i \rangle_{\mathcal{H}} = y_i, i=1,\ldots n} \Omega(f). \tag{12}$$

- $\Omega : \mathcal{H} \to \bar{\mathbb{R}}$ is
  - *admissible*: if $\exists$,
  - *strictly admissible*: if $\forall$

  f minimizers of task I take the form

$$f = \sum_{i=1}^{n} a_i f_i \quad (\mathbf{a} \in \mathbb{R}^n). \tag{13}$$

Chosen paper: continued – $R \Leftrightarrow I$

- If $f_i = k(x_i, \cdot)$, then $f(x_i) = y_i$ (RKHS).
- Advantage: loss L no longer appears in the formulation.
- Still: from representer theorem point of view
  - $I \Rightarrow R$,
  - $R \Leftarrow I$: under mild conditions on L.
- Inner product constraints: Euclidean spaces.

- Similarly to [Argyriou et al. '09]: Ω is
  - *admissible* iff:

$$\forall f, g \in \mathcal{H}, \langle f, g\rangle_{\mathcal{H}} = 0 \Rightarrow \Omega(f+g) \geq \Omega(f). \qquad (14)$$

  - *strictly admissible* iff:

$$\forall f, g \in \mathcal{H}, \langle f, g\rangle_{\mathcal{H}} = 0, g \neq 0 \Rightarrow \Omega(f+g) > \Omega(f). \qquad (15)$$

- Intuitively, the contours of Ω are spheres.

Chosen paper: result

Ω is

- *admissible* iff it is *weakly*-

$$\forall f, g \in \mathcal{H}, \|g\| > \|f\| \Rightarrow \Omega(g) \geq \Omega(f), \tag{16}$$

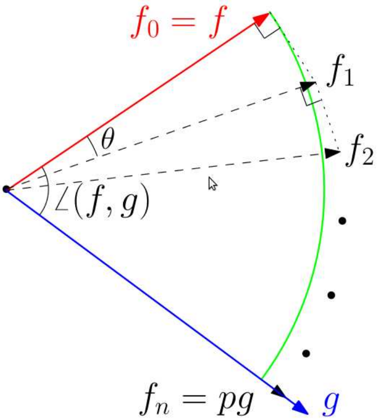
- *strictly admissible* iff it is *strictly increasing*

$$\forall f, g \in \mathcal{H}, \|g\| > \|f\| \Rightarrow \Omega(g) > \Omega(f) \tag{17}$$

function of the norm of its argument.

Chosen paper: proof – intuition

1. Admissibility $\Rightarrow \Omega$ is increasing along any ray $R_g = \{tg : t \geq 0, g \neq 0\}$ ($\Leftarrow$ contour result 2x).
2. If $\|f\| < \|g\|$, then f can be subtly, on each segment perpendicularly rotated to pg ($p < 1$):



Conclusion

- From representer theorem point of view: $R \Leftrightarrow I$.
- Contributions:
  - differentiability/semi-continuous assumption: relaxed.
  - Hilbert $\rightarrow$ Euclidean space.
- (Strict) admissibility of Ω-s can be characterized.

Thank you for the attention!

Definitions: l.s.c., u.s.c.

$f: \mathcal{H} \to \bar{\mathbb{R}}$ is

- l.s.c., if
  - As $x \to x_0$, $f(x)$ is either close to, or larger than $f(x_0)$. Formally,

$$f(x_0) \le \liminf_{x \to x_0} f(x) \quad (\forall x_0). \tag{18}$$

  - $epi(f) := \{(x,t) \in \mathcal{H} \times \bar{\mathbb{R}} : f(x) \le t\}$ is closed.
  - $\{x \in \mathcal{H} : f(x) > \alpha\}$ open ($\forall \alpha \in \mathbb{R}$),
  - $\{x \in \mathcal{H} : f(x) \le \alpha\}$ closed ($\forall \alpha \in \mathbb{R}$).
- u.s.c, if $-f$ is l.s.c. Continuous: l.s.c. and u.s.c.

Example: $f(x) = \lceil x \rceil$ ($\lfloor x \rfloor$) is l.s.c. (u.s.c.).

Fatou lemma

- $\mathcal{H} := L^+(\mathcal{X}, \mathcal{A}, \mu)$:
  - non-negative measureable functions,
  - topology: convergence in μ.
- Fatou lemma:
  - $\int : L^+(\mathcal{X}, \mathcal{A}, \mu) \to \bar{\mathbb{R}}$ is l.s.c., i.e.,
  - $\int_{\mathcal{X}} \liminf_{n\to\infty} f_n \mathrm{d}\mu \le \liminf_{n\to\infty} \int_{\mathcal{X}} f_n \mathrm{d}\mu$.

Gateaux differentiability

$\Omega : \mathcal{H} \to \mathbb{R}$ is said to be Gateaux differentiable

- at $f \in \mathcal{H}$, if for $\forall h$ (direction) $\in \mathcal{H}\ \exists\ \Omega'_f(h) \in \mathbb{R}$ such that

$$\Omega(f + th) - \Omega(f) = t\Omega'_f(h) + o(t), \text{ as } t \to 0. \qquad (19)$$

- if it is Gateaux differentiable at $\forall f \in \mathcal{H}$.