NOTES ON OPERATOR VALUED KERNELS, FEATURE MAPS AND GAUSSIAN PROCESSES

HOUMAN OWHADI

Abstract. These notes serve as a short introduction to operator-valued kernels, their associated feature maps and Gaussian processes.

1. Introduction

Operator-valued kernels were introduced in [2] as a generalization of vector-valued kernels [1]. The following notes, taken almost verbatim from parts of [5], serve as a short introduction to such kernels, their associated feature maps and Gaussian processes.

2. Operator valued kernels

Let $\mathcal{X}$ and $\mathcal{Y}$ be separable Hilbert spaces endowed with the inner products $\langle\cdot,\cdot\rangle_{\mathcal{X}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{Y}}$. Write $\mathcal{L}(\mathcal{Y})$ for the set of bounded linear operators mapping $\mathcal{Y}$ to $\mathcal{Y}$. We call $K : \mathcal{X}\times\mathcal{X}\rightarrow\mathcal{L}(\mathcal{Y})$ an **operator-valued kernel** if

(1) K is Hermitian, i.e.

$$K(x,x') = K(x',x)^T \text{ for } x,x'\in\mathcal{X}\,, \tag{2.1}$$

writing A^T for the adjoint of the operator A with respect to $\langle\cdot,\cdot\rangle_{\mathcal{Y}}$, and

(2) non-negative, i.e.

$$\sum_{i,j=1}^{m}\langle y_i, K(x_i,x_j)y_j\rangle_{\mathcal{Y}} \geqslant 0 \text{ for } (x_i,y_i)\in\mathcal{X}\times\mathcal{Y},\, m\in\mathbb{N}\,. \tag{2.2}$$

We call K non-degenerate if $\sum_{i,j=1}^{m}\langle y_i, K(x_i,x_j)y_j\rangle_{\mathcal{Y}} = 0$ implies $y_i = 0$ for all i whenever $x_i \neq x_j$ for $i\neq j$.

3. Reproducing kernel Hilbert space

Each non-degenerate, locally bounded and separately continuous operator-valued kernel K (which we will refer to as a Mercer's kernel) is in one to one correspondence with a reproducing kernel Hilbert space $\mathcal{H}$ of continuous functions $f : \mathcal{X}\rightarrow\mathcal{Y}$ obtained as the closure of the linear span of functions $z\rightarrow K(z,x)y$ ($(x,y)\in\mathcal{X}\times\mathcal{Y}$) with respect to the inner product identified by the reproducing property

$$\langle f, K(\cdot,x)y\rangle_{\mathcal{H}} = \langle f(x), y\rangle_{\mathcal{Y}} \tag{3.1}$$

Date: October 18, 2021.
Caltech, MC 9-94, Pasadena, CA 91125, USA, owhadi@caltech.edu.

4. Feature maps

Let $\mathcal{F}$ be a separable Hilbert space (with inner product $\langle\cdot,\cdot\rangle_{\mathcal{F}}$ and norm $\|\cdot\|_{\mathcal{F}}$) and let $\psi : \mathcal{X} \to \mathcal{L}(\mathcal{Y},\mathcal{F})$ be a continuous function mapping $\mathcal{X}$ to the space of bounded linear operators from $\mathcal{Y}$ to $\mathcal{F}$.

Definition 4.1. *We say that $\mathcal{F}$ and $\psi : \mathcal{X} \to \mathcal{L}(\mathcal{Y},\mathcal{F})$ are a* feature space *and a* feature map *for the kernel K if, for all $(x,x',y,y') \in \mathcal{X}^2 \times \mathcal{Y}^2$,*

$$y^T K(x,x')y' = \langle \psi(x)y, \psi(x')y'\rangle_{\mathcal{F}}\,. \tag{4.1}$$

Write $\psi^T(x)$, for the adjoint of $\psi(x)$ defined as the linear function mapping $\mathcal{F}$ to $\mathcal{Y}$ satisfying

$$\langle \psi(x)y, \alpha\rangle_{\mathcal{F}} = \langle y, \psi^T(x)\alpha\rangle_{\mathcal{Y}} \tag{4.2}$$

for $x,y,\alpha \in \mathcal{X}\times\mathcal{Y}\times\mathcal{F}$. Note that $\psi^T : \mathcal{X} \to \mathcal{L}(\mathcal{F},\mathcal{Y})$ is therefore a function mapping $\mathcal{X}$ to the space of bounded linear functions from $\mathcal{F}$ to $\mathcal{Y}$. Writing $\alpha^T\alpha' := \langle\alpha,\alpha'\rangle_{\mathcal{F}}$ for the inner product in $\mathcal{F}$ we can ease our notations by writing

$$K(x,x') = \psi^T(x)\psi(x') \tag{4.3}$$

which is consistent with the finite-dimensional setting and $y^T K(x,x')y' = (\psi(x)y)^T(\psi(x')y')$ (writing y^Ty' for the inner product in $\mathcal{Y}$). For $\alpha \in \mathcal{F}$ write $\psi^T\alpha$ for the function $\mathcal{X} \to \mathcal{Y}$ mapping $x \in \mathcal{X}$ to the element $y \in \mathcal{Y}$ such that

$$\langle y', y\rangle_{\mathcal{Y}} = \langle y', \psi^T(x)\alpha\rangle_{\mathcal{Y}} = \langle \psi(x)y', \alpha\rangle_{\mathcal{F}} \text{ for all } y' \in \mathcal{Y}\,. \tag{4.4}$$

We can, without loss of generality, restrict $\mathcal{F}$ to be the range of $(x,y) \to \psi(x)y$ so that the RKHS $\mathcal{H}$ defined by K is the (closure of) linear space spanned by $\psi^T\alpha$ for $\alpha \in \mathcal{F}$. Note that the reproducing property (3.1) implies that for $\alpha \in \mathcal{F}$

$$\langle \psi^T(\cdot)\alpha, \psi^T(\cdot)\psi(x)y\rangle_{\mathcal{H}} = \langle \psi^T(x)\alpha, y\rangle_{\mathcal{Y}} = \langle \alpha, \psi(x)y\rangle_{\mathcal{F}} \tag{4.5}$$

for all $x,y \in \mathcal{X}\times\mathcal{Y}$, which leads to the following theorem.

Theorem 4.2. *The RKHS $\mathcal{H}$ defined by the kernel (4.3) is the linear span of $\psi^T\alpha$ over $\alpha \in \mathcal{F}$ such that $\|\alpha\|_{\mathcal{F}} < \infty$. Furthermore, $\langle \psi^T(\cdot)\alpha, \psi^T(\cdot)\alpha'\rangle_{\mathcal{H}} = \langle\alpha,\alpha'\rangle_{\mathcal{F}}$ and*

$$\|\psi^T(\cdot)\alpha\|_{\mathcal{H}}^2 = \|\alpha\|_{\mathcal{F}}^2 \text{ for } \alpha,\alpha' \in \mathcal{F}\,. \tag{4.6}$$

5. Interpolation

We employ the setting of supervised learning, which can be expressed as solving the following problem.

Problem 1. *Let $f^\dagger$ be an unknown continuous function mapping $\mathcal{X}$ to $\mathcal{Y}$. Given the information*[1] *$f^\dagger(X) = Y$ with the data $(X,Y) \in \mathcal{X}^N \times \mathcal{Y}^N$ approximate $f^\dagger$.*

[1] For a N-vector $X = (X_1,\ldots,X_N) \in \mathcal{X}^N$ and a function $f : \mathcal{X} \to \mathcal{Y}$, write $f(X)$ for the N vector with entries $(f(X_1),\ldots,f(X_N))$ (we will keep using this generic notation).

Using the relative error in $\|\cdot\|_{\mathcal{H}}$-norm as a loss, the minimax optimal recovery solution of Problem (1) is [6, Thm. 12.4,12.5] the minimizer (in $\mathcal{H}$) of

$$\begin{cases} \text{Minimize} & \|f\|_{\mathcal{H}}^2 \\ \text{subject to} & f(X) = Y \end{cases} \tag{5.1}$$

By the representer theorem [3], the minimizer of (5.1) is

$$f(\cdot) = \sum_{j=1}^{N} K(\cdot, X_j) Z_j \,, \tag{5.2}$$

where the coefficients $Z_j \in \mathcal{Y}$ are identified by solving the system of linear equations

$$\sum_{j=1}^{N} K(X_i, X_j) Z_j = Y_i \text{ for all } i \in \{1, \ldots, N\} \,, \tag{5.3}$$

i.e. $K(X, X)Z = Y$ where $Z = (Z_1, \ldots, Z_N)$, $Y = (Y_1, \ldots, Y_N) \in \mathcal{Y}^N$ and $K(X, X)$ is the $N \times N$ block-operator matrix[2] with entries $K(X_i, X_j)$. Therefore, writing $K(\cdot, X)$ for the vector $(K(\cdot, X_1), \ldots, K(\cdot, X_N)) \in \mathcal{H}^N$, the minimizer of (5.1) is

$$f(\cdot) = K(\cdot, X) K(X, X)^{-1} Y \,, \tag{5.4}$$

which implies that the value of (5.1) at the minimum is

$$\|f\|_{\mathcal{H}}^2 = Y^T K(X, X)^{-1} Y \,, \tag{5.5}$$

where $K(X, X)^{-1}$ is the inverse of $K(X, X)$ (whose existence is implied by the non-degeneracy of K combined with $X_i \neq X_j$ for $i \neq j$).

6. Ridge regression

Let $\lambda > 0$. A ridge regression solution (also known as Tikhonov regularizer) to Problem 1 is a minimizer of

$$\inf_{f \in \mathcal{H}} \lambda \|f\|_{\mathcal{H}}^2 + \sum_{i=1}^{N} \|Y_i' - Y_i\|_{\mathcal{Y}}^2 \,. \tag{6.1}$$

The minimizer of (6.1) is

$$f(x) = K(x, X) \big(K(X, X) + \lambda I\big)^{-1} Y \,, \tag{6.2}$$

(writing I for the identity matrix) and the value of (6.1) at the minimum is

$$\lambda Y^T \big(K(X, X) + \lambda I\big)^{-1} Y \,. \tag{6.3}$$

---

[2]For $N \geqslant 1$ let $\mathcal{Y}^N$ be the N-fold product space endowed with the inner-product $\langle Y, Z \rangle_{\mathcal{Y}^N} := \sum_{i,j=1}^{N} \langle Y_i, Z_j \rangle_{\mathcal{Y}}$ for $Y = (Y_1, \ldots, Y_N), Z = (Z_1, \ldots, Z_N) \in \mathcal{Y}^N$. $\mathbf{A} \in \mathcal{L}(\mathcal{Y}^N)$ given by $\mathbf{A} = \begin{pmatrix} A_{1,1} & \cdots & A_{1,N} \\ \vdots & & \vdots \\ A_{N,1} & \cdots & A_{N,N} \end{pmatrix}$ where $A_{i,j} \in \mathcal{L}(\mathcal{Y})$, is called a block-operator matrix. Its adjoint $\mathbf{A}^{\mathbf{T}}$ with respect to $\langle \cdot, \cdot \rangle_{\mathcal{Y}^N}$ is the block-operator matrix with entries $(A^T)_{i,j} = (A_{j,i})^T$.

7. Function-valued Gaussian processes

The following definition of function-valued Gaussian processes is a natural extension of scalar-valued Gaussian fields.

Definition 7.1. *Let $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ be an operator-valued kernel. Let m be a function mapping $\mathcal{X}$ to $\mathcal{Y}$. We call $\xi : \mathcal{X} \to \mathcal{L}(\mathcal{Y}, \mathbf{H})$ a function-valued Gaussian process if ξ is a function mapping $x \in \mathcal{X}$ to $\xi(x) \in \mathcal{L}(\mathcal{Y}, \mathbf{H})$ where $\mathbf{H}$ is a Gaussian space and $\mathcal{L}(\mathcal{Y}, \mathbf{H})$ is the space of bounded linear operators from $\mathcal{Y}$ to $\mathbf{H}$. Abusing notations we write $\langle \xi(x), y \rangle_{\mathcal{Y}}$ for $\xi(x)y$. We say that ξ has mean m and covariance kernel K and write $\xi \sim \mathcal{N}(m, K)$ if $\langle \xi(x), y \rangle_{\mathcal{Y}} \sim \mathcal{N}\big(m(x), y^T K(x,x) y\big)$ and*

$$\operatorname{Cov}\big(\langle \xi(x), y \rangle_{\mathcal{Y}}, \langle \xi(x'), y' \rangle_{\mathcal{Y}}\big) = y^T K(x,x') y' \,. \tag{7.1}$$

We say that ξ is centered if it is of zero mean.

If $K(x,x)$ is trace class ($\operatorname{Tr}[K(x,x)] < \infty$) then $\xi(x)$ defines a measure on $\mathcal{Y}$ (i.e. a $\mathcal{Y}$-valued random variable), otherwise it only defines a (weak) cylinder-measure in the sense of Gaussian fields.

Theorem 7.2. *The distribution of a function-valued Gaussian process is uniquely determined by its mean and covariance kernel K. Conversely given m and K there exists a function-valued Gaussian process having mean m and covariance kernel K. In particular if K has feature space $\mathcal{F}$ and map ψ, the e_i form an orthonormal basis of $\mathcal{F}$, and the Z_i are i.i.d. $\mathcal{N}(0,1)$ random variables, then*

$$\xi = m + \sum_i Z_i \psi^T e_i \tag{7.2}$$

is a function-valued GP with mean m and covariance kernel K.

Proof. The proof is classical, see [6, Sec. 7&17]. Note that the separability of $\mathcal{F}$ ensures the existence of the e_i. Furthermore $\mathbb{E}\big[(\xi - m)(\xi - m)^T\big] = \psi^T \psi = K$. ☐

Theorem 7.3. *Let ξ be a centered function-valued GP with covariance kernel $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$. Let $X, Y \in \mathcal{X}^N \times \mathcal{Y}^N$. Let $Z = (Z_1, \ldots, Z_N)$ be a random Gaussian vector, independent from ξ, with i.i.d. $\mathcal{N}(0, \lambda I_{\mathcal{Y}})$ entries ($\lambda \geqslant 0$ and $I_{\mathcal{Y}}$ is the identity map on $\mathcal{Y}$). Then ξ conditioned on $\xi(X) + Z$ is a function-valued GP with mean*

$$\mathbb{E}\big[\xi(x) \big| \xi(X) + Z = Y\big] = K(x,X)\big(K(X,X) + \lambda I_{\mathcal{Y}}\big)^{-1} Y = (6.2) \tag{7.3}$$

and conditional covariance operator

$$K^{\perp}(x,x') := K(x,x') - K(x,X)\big(K(X,X) + \lambda I_{\mathcal{Y}}\big)^{-1} K(X,x') \,. \tag{7.4}$$

In particular, if K is trace class, then

$$\sigma^2(x) := \mathbb{E}\Big[\big\|\xi(x) - \mathbb{E}[\xi(x)|\xi(X) + Z = Y]\big\|_{\mathcal{Y}}^2 \Big| \xi(X) + Z = Y\Big] = \operatorname{Tr}\big[K^{\perp}(x,x)\big] \,. \tag{7.5}$$

Proof. The proof is a generalization of the classical setting [6, Sec. 7&17]. Writing $\xi^T(x)y$ for $\langle \xi(x), y \rangle_{\mathcal{Y}}$ observe that $y^T \xi(x) \xi^T(x') y = y^T K(x,x') y'$ implies $\mathbb{E}[\xi(x)\xi^T(x')] = K(x,x')$. Since ξ and Z share the same Gaussian space the expectation of $\xi(x)$ conditioned on $\xi(X) + Z$ is $A\big(\xi(X) + Z\big)$ where A is a linear map identified by $0 =$

$\operatorname{Cov}\Big(\xi(x) - A\big(\xi(X)+Z\big), \xi(X)+Z\Big) = \mathbb{E}\big[\xi(x) - A\big(\xi(X)+Z\big)\big(\xi^T(X)+Z^T\big)\big] = K(x,X) - A\big(K(X,X)+\lambda I_{\mathcal{Y}}\big)$, which leads to $A = K(x,X)\big(K(X,X)+\lambda I_{\mathcal{Y}}\big)^{-1}$ and (7.3). The conditional covariance is then given by $K^{\perp}(x,x') = \mathbb{E}\Big[\Big(\xi(x) - K(x,X)\big(K(X,X) + \lambda I_{\mathcal{Y}}\big)^{-1}\big(\xi(X)+Z\big)\Big)\Big(\xi(x') - K(x',X)\big(K(X,X)+\lambda I_{\mathcal{Y}}\big)^{-1}\big(\xi(X)+Z\big)\Big)^T\Big]$ which leads to (7.4). □

8. Deterministic error estimates for function-valued Kriging

The following theorem shows that the standard deviation (7.5) provides deterministic a prior error bounds on the accuracy of the ridge regressor (7.3) to $f^\dagger$ in Problem 1. Local error estimates such as (8.1) are classical in Kriging [7] where $\sigma^2(x)$ is known as the power function/kriging variance (see also [4][Thm. 5.1] for applications to PDEs).

Theorem 8.1. *Let $f^\dagger$ be the unknown function of Problem 1 and let $f(x) =$ (7.3) $=$ (??) be its GPR/ridge regression solution. Let $\mathcal{H}$ be the RKHS associated with K and let $\mathcal{H}_\lambda$ be the RKHS associated with the kernel $K_\lambda := K + \lambda I_{\mathcal{Y}}$. It holds true that*

$$\big\|f^\dagger(x) - f(x)\big\|_{\mathcal{Y}} \leqslant \sigma(x)\|f^\dagger\|_{\mathcal{H}} \tag{8.1}$$

and

$$\big\|f^\dagger(x) - f(x)\big\|_{\mathcal{Y}} \leqslant \sqrt{\sigma^2(x) + \lambda \dim(\mathcal{Y})}\,\|f^\dagger\|_{\mathcal{H}_\lambda}\,, \tag{8.2}$$

where $\sigma(x)$ is the standard deviation (7.5).

Proof. Let $y \in \mathcal{Y}$. Using the reproducing property (3.1) and $Y = f^\dagger(X)$ we have

$$\begin{aligned} y^T\big(f^\dagger(x) - f(x)\big) &= y^T f^\dagger(x) - y^T K(x,X)\big(K(X,X)+\lambda I_{\mathcal{Y}}\big)^{-1} f^\dagger(X) \\ &= \big\langle f^\dagger, K(\cdot,x)y - K(\cdot,X)\big(K(X,X)+\lambda I_{\mathcal{Y}}\big)^{-1}K(X,x)y\big\rangle_{\mathcal{H}}\,. \end{aligned}$$

Using Cauchy-Schwartz inequality, we deduce that

$$\Big|y^T\big(f^\dagger(x) - f(x)\big)\Big|^2 \leqslant \|f^\dagger\|^2_{\mathcal{H}}\, y^T K^{\perp}(x,x)y \tag{8.3}$$

where $K^{\perp}$ is the conditional covariance (7.4). Summing over y ranging in basis of $\mathcal{Y}$ implies (8.1). The proof of (8.2) is similar, simply observe that

$$\begin{aligned} y^T\big(f^\dagger(x) - f(x)\big) &= \big\langle f^\dagger, K_\lambda(\cdot,x)y - K_\lambda(\cdot,X)\big(K(X,X)+\lambda I_{\mathcal{Y}}\big)^{-1}K(X,x)y\big\rangle_{\mathcal{H}_\lambda} \\ &\leqslant \|f^\dagger\|_{H_\lambda}\big\|K_\lambda(\cdot,x)y - K_\lambda(\cdot,X)\big(K(X,X)+\lambda I_{\mathcal{Y}}\big)^{-1}K(X,x)y\big\|_{\mathcal{H}_\lambda}\,, \end{aligned}$$

which implies

$$\Big|y^T\big(f^\dagger(x) - f(x)\big)\Big|^2 \leqslant \|f^\dagger\|^2_{\mathcal{H}_\lambda}\big(\lambda y^T y + y^T K^{\perp}(x,x)y\big)\,. \tag{8.4}$$

□

Remark 8.2. *Since Thm. 8.1 does not require $\mathcal{X}$ to be finite-dimensional, its estimates do not suffer from the curse of dimensionality but from finding a good kernel for which both $\|f^\dagger\|_{\mathcal{H}}$ and $y^T K^{\perp}(x,x)y$ are small (over x sampled from the testing distribution). Indeed both (8.1) and (8.2) provide a priori deterministic error bounds on $f^\dagger - f$ depending on the RKHS norms $\|f^\dagger\|_{\mathcal{H}}$ and $\|f^\dagger\|_{\mathcal{H}_\lambda}$. Although these norms can be controlled in the*

PDE setting [4] via compact embeddings of Sobolev spaces, there is no clear strategy for obtaining a-priori bounds on these norms for general machine learning problems.

References

[1] Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

[2] Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren. Operator-valued kernels for learning from functional response data. *The Journal of Machine Learning Research*, 17(1):613–666, 2016.

[3] Charles A Micchelli and Massimiliano Pontil. Kernels for multi–task learning. In *Advances in neural information processing systems*, pages 921–928, 2005.

[4] Houman Owhadi. Bayesian numerical homogenization. *Multiscale Modeling & Simulation*, 13(3):812–828, 2015.

[5] Houman Owhadi. Do ideas have shape? plato's theory of forms as the continuous limit of artificial neural networks. *arXiv preprint arXiv:2008.03920*, 2020.

[6] Houman Owhadi and Clint Scovel. *Operator-Adapted Wavelets, Fast Solvers, and Numerical Homogenization: From a Game Theoretic Approach to Numerical Approximation and Algorithm Design*, volume 35. Cambridge University Press, 2019.

[7] Zong-min Wu and Robert Schaback. Local error estimates for radial basis function interpolation of scattered data. *IMA journal of Numerical Analysis*, 13(1):13–27, 1993.