Universität Stuttgart

SimTech
Cluster of Excellence

D. Wittwar[a] · **B. Haasdonk**[a]

Greedy Algorithms for Matrix-Valued Kernels

Stuttgart, January 2018

[a] Institute for Applied Analysis and Numerical Simulation, University of Stuttgart,
Pfaffenwaldring 57, 70569 Stuttgart/ Germany
{dominik.wittwar,bernard.haasdonk}@mathematik.uni-stuttgart.de
www.agh.ians.uni-stuttgart.de

Abstract We are interested in approximating vector-valued functions on a compact set $\Omega \subset \mathbb{R}^d$. We consider reproducing kernel Hilbert spaces of $\mathbb{R}^m$-valued functions which each admit a unique matrix-valued reproducing kernel k. These spaces seem promising, when modelling correlations between the target function components. The approximation of a function is a linear combination of matrix-valued kernel evaluations multiplied with coefficient vectors. To guarantee a fast evaluation of the approximant the expansion size, i.e. the number of centers n is desired to be small. We thus present three different greedy algorithms by which a suitable set of centers is chosen in an incremental fashion: First, the P-Greedy which requires no function evaluations, second and third, the f-Greedy and f/P-Greedy which require function evaluations but produce centers tailored to the target function. The efficiency of the approaches is investigated on some data from an artificial model.

Stuttgart Research Centre for Simulation Technology (SRC SimTech)
SimTech – Cluster of Excellence
Pfaffenwaldring 5a
70569 Stuttgart
publications@simtech.uni-stuttgart.de
www.simtech.uni-stuttgart.de

Greedy Algorithms for Matrix-Valued Kernels

Dominik Wittwar[1] and Bernard Haasdonk[2]

[1] University of Stuttgart, Institute for Applied Analysis and Numerical Simulation, dominik.wittwar@mathematik.uni-stuttgart.de
[2] University of Stuttgart, Institute for Applied Analysis and Numerical Simulation, bernard.haasdonk@mathematik.uni-stuttgart.de

Abstract. We are interested in approximating vector-valued functions on a compact set $\Omega \subset \mathbb{R}^d$. We consider reproducing kernel Hilbert spaces of $\mathbb{R}^m$-valued functions which each admit a unique matriv-valued reproducing kernel k. These spaces seem promising, when modelling correlations between the target function components. The approximation of a function is a linear combination of matrix-valued kernel evaluations multiplied with coefficient vectors. To guarantee a fast evaluation of the approximant the expansion size, i.e. the number of centers n is desired to be small. We thus present three different greedy algorithms by which a suitable set of centers is chosen in an incremental fashion: First, the P–Greedy which requires no function evaluations, second and third, the f–Greedy and f/P–Greedy which require function evaluations but produce centers tailored to the target function. The efficiency of the approaches is investigated on some data from an artificial model.

1 Matrix-Valued Kernels

We will give a short overview on the theory of matrix-valued kernels and how they can be applied in the context of approximation/surrogate modelling. For further information and a more thorough introduction, we refer to literature, e.g. [1,3].

For a compact set $\Omega \subset \mathbb{R}^d$ a bivariate function $k : \Omega \times \Omega \to \mathbb{R}^{m\times m}$ is called a *matrix-valued kernel* if $k(x, y) = k(y, x)^T$. It is further denoted as *(strictly) positive definite* if for any finite set $X = \{x_1, \dots, x_n\} \subset \Omega$ of pairwise distinct points the associated *block Gramian matrix* $k(X, X) := (k(x_i, x_j))_{i,j} \in \mathbb{R}^{mn\times mn}$ is positive (semi-)definite. In this case, there exists a unique Hilbert space, the so called *native space* $\mathcal{H}_k$ of $\mathbb{R}^m$-valued functions over the domain Ω such that the kernel k satisfies

$$k(\cdot, x)\alpha \in \mathcal{H}_k, \quad \forall\, x \in \Omega, \forall\, \alpha \in \mathbb{R}^m \tag{1}$$

$$\langle f, k(\cdot, x)\alpha\rangle_{\mathcal{H}_k} = f(x)^T\alpha, \quad \forall\, x \in \Omega, \forall\, f \in \mathcal{H}_k, \forall\, \alpha \in \mathbb{R}^m \tag{2}$$

where (2) is called the *reproducing property.*

It follows that the directional kernel evaluations $k(\cdot, x)\alpha$ are the Riesz representers of the directional point evaluation functionals $\delta_x^\alpha : \mathcal{H}_k \to \mathbb{R}$, $\delta_x^\alpha(f) := f(x)^T\alpha$. With the Cauchy-Schwarz inequality these δ_x^α are bounded. Vice versa, if for a Hilbert space $\mathcal{H}$ of functions $f : \Omega \to \mathbb{R}^d$ all directional

point evaluation functionals are bounded, there exists a unique positive definite kernel which satisfies (1)–(2). Hence, such a Hilbert space is referred to as *reproducing kernel Hilbert space*, or RKHS for short, and k is called its *reproducing kernel*.

Given a function $f : \Omega \to \mathbb{R}^m$ and a set of n pairwise distinct points $X := \{x_1, \dots, x_n\}$, the kernel interpolant s_X^f of f on the *centers* X can be defined via

$$s_X^f(x) := \sum_{i=1}^{n} k(x, x_i)\alpha_i, \tag{3}$$

where the *coefficient vectors* $\alpha_i \in \mathbb{R}^m$ solve the linear system

$$\sum_{i=1}^{p} k(x_j, x_i)\alpha_i = f(x_j), \quad \text{for } j = 1, \dots, n. \tag{4}$$

We assume in the following, that k is a (not necessarily strictly) positive definite kernel and $f \in \mathcal{H}_k$. In this case, (4) may have non-unique coefficient solutions, which however all represent the unique interpolant s_X^f. In the case of strictly positive definite kernels, even the coefficient vectors in (4) are unique for arbitrary finite sets of pairwise distinct points $X \subset \Omega$, and interpolation is also well-posed if $f \notin \mathcal{H}_k$. Moreover, as a consequence of the reproducing property (2), the interpolant s_X^f can be identified as the best approximation of f in the subspace $\mathcal{N}(X) \subset \mathcal{H}_k$ given by

$$\mathcal{N}(X) := \operatorname{span}\{k(\cdot, x)\alpha |\, x \in X, \alpha \in \mathbb{R}^m\}. \tag{5}$$

Since $\mathcal{N}(X)$ is a finite dimensional closed subspace, it is also an RKHS, since the directional point evaluation functionals δ_x^α restricted to $\mathcal{N}(X)$ are still bounded. Hence, $\mathcal{N}(X)$ admits its own unique reproducing kernel $k_{\mathcal{N}(X)}$. It can be shown, c.f. [6], that this reproducing kernel is given by

$$k_{\mathcal{N}(X)}(x, y) = k(x, X)k(X, X)^+ k(X, y), \tag{6}$$

where $k(X, X)^+$ denotes the Moore-Penrose pseudoinverse of the Gramian matrix $k(X, X)$. Furthermore, the orthogonal projection operator $\Pi_{\mathcal{N}(X)} : \mathcal{H}_k \to \mathcal{N}(X)$ is well defined and, therefore, we are able to define the *Power-Function* $\mathcal{P}_X : \mathcal{H}_k^* \to \mathbb{R}$ via

$$\mathcal{P}_X(\lambda) = \sup_{f \in \mathcal{H}_k \setminus \{0\}} \frac{|\lambda(f) - \lambda(\Pi_{\mathcal{N}(X)}(f))|}{\|f\|_{\mathcal{H}_k}}, \quad \text{for } \lambda \in \mathcal{H}_k^*. \tag{7}$$

Using the Cauchy-Schwarz inequality and the fact that $\Pi_{\mathcal{N}(X)}$ is self-adjoint, it can be shown, see [6], that for the directional point evaluation functional δ_x^α the Power-Function is given by

$$\mathcal{P}_X(\delta_x^\alpha)^2 = \alpha^T (k(x, x) - k_{\mathcal{N}(X)}(x, x))\alpha. \tag{8}$$

For notational convenience, we denote as

$$\mathbf{P}_X(x) := k(x,x) - k_{\mathcal{N}(X)}(x,x) \tag{9}$$

the *Power-Function matrix*, which in general is positive semidefinite.

Combining (7)–(9) we get the following directional error bound

$$|(s_X^f(x) - f(x))^T \alpha|^2 \leq \alpha^T \mathbf{P}_X(x) \alpha \|f\|_{\mathcal{H}_k}^2. \tag{10}$$

2 Greedy Algorithm

For a given kernel and target function, the quality of the interpolant is dependent on the choice of centers. For the selection of these centers we employ the kernel greedy algorithm, whose pseudo code is given in Algorithm 1, and which works as follows: We assume to have a given finite sampling $\Omega_N \subset \Omega$ of the input space, an initial set of centers $X \subset \Omega$, this may be empty, a tolerance $\varepsilon > 0$ and an error indicator function E. Now, we iteratively select a point maximizing E, add it to the set of centers and compute the next approximant by interpolation on the small set of chosen centers. This is repeated until the tolerance ε is reached.

Algorithm 1 General Kernel Greedy Algorithm

Require: finite sampling of the input domain $\Omega_N \subset \Omega$, kernel $k : \Omega \times \Omega \to \mathbb{R}^{m \times m}$, target function $f : \Omega \to \mathbb{R}^m$, initial set of centers X, error indicator function E, tolerance $\varepsilon > 0$.

1: **while** $\max_{x \in \Omega_N} E(k, f, X, x) \geq \varepsilon$ **do**
2: $\quad x^* = \arg\max_{x \in \Omega_N} E(k, f, X, x)$
3: $\quad X = X \cup \{x^*\}$
4: **end while**
5: **return** X

In the following we consider three instantiations by different choices of E, resulting in the P–Greedy, f–Greedy and f/P–Greedy algorithms.

2.1 *P*–Greedy

The P–Greedy uses an error indicator that depends on the Power-function but is independent of the target function. Hence, it results in a set of centers which are suitable for variety of different target functions. Furthermore, no (expensive) function evaluations are necessary in the selection process which can thus be performed in a rapid manner. Using the directional error bound

(10) and taking the supremum over all directions $\alpha \in \mathbb{R}^m$ of length one, we end up with the bound

$$\|s_X^f(x) - f(x)\|_2^2 \leq \|\mathbf{P}_X(x)\|_2 \|f\|_{\mathcal{H}_k}^2, \tag{11}$$

where $\|\cdot\|_2$ denotes the Euclidean norm for vectors and the spectral norm for matrices, respectively. The error indicator function E_1 is then given as

$$E_1(k, f, X, x) := E_1(k, X, x) := \|\mathbf{P}_X(x)\|_2. \tag{12}$$

We note, that by (6) we have $k_{\mathcal{N}(X)}(x, x) = k(x, x)$ for all $x \in X$ and thus $E_1(k, X, x) = 0$ for all $x \in X$. Moreover, as a direct consequence of (7) we have $E_1(k, X, x) \leq E_1(k, Y, x)$ for all $x \in \Omega$ and $Y \subset X$. In particular the algorithm terminates after a finite number of steps and no point is chosen a second time.

In the scalar-valued case, see [2], this algorithm has recently been shown to result in quasi-optimal approximations for kernels of Sobolev spaces [4] and asymptotically uniformly distributed point sets.

2.2 f–Greedy

For the f–Greedy the error indicator function E_2 is given by

$$E_2(k, f, X, x) := \|s_X^f(x) - f(x)\|_2^2. \tag{13}$$

One can see, that the indicator relies on the evaluation of the target function and should therefore select a set of centers that is tailored to the target function. This is expected to lead to a smaller number of centers when compared to the P–Greedy. However, it involves all target values $f(x)$, $x \in \Omega_N$ which may not be cheaply available, and the resulting set of centers is individually suited to this particular target function. In contrast to the indicator E_1 the indicator E_2 is in general not decreasing, i.e. the inequality $E_2(k, f, X, x) \leq E_2(k, f, Y, x)$ for $x \in \Omega$ and $Y \subset X$ does not necessarily hold. Nonetheless, we still have $E_2(k, f, X, x) = 0$ for all $x \in X$ and no point is selected twice.

2.3 f/P–Greedy

By the reproducing property (2) we obtain

$$\|f(x) - s_X^f(x)\|_2 \leq \|k(x, x)\|_2 \|f - s_X^f\|_{\mathcal{H}_k}, \tag{14}$$

thus, the error in the Euclidean norm can be bounded by a kernel dependent constant and the error in the Hilbert space norm. Hence, it seems reasonable to choose an indicator function E_3 in such a way, that the error in the Hilbert

space norm is minimized. Since s_X^f is the best approximation of f in $\mathcal{N}(X)$ we have

$$\|f - s_X^f\|_{\mathcal{H}_k}^2 = \|f\|_{\mathcal{H}_k}^2 - \|s_X^f\|_{\mathcal{H}_k}^2 \tag{15}$$

and, therefore, minimization of the left hand side is equivalent to maximizing the Hilbert space norm of the interpolant s_X^f. For this purpose the error idicator function E_3 is chosen as

$$E_3(k, f, X, x) := (s_X^f(x) - f(x))^T \mathbf{P}_X(x)^+ (s_X^f(x) - f(x)). \tag{16}$$

For scalar strictly positive definite kernels this is equal to

$$E_3(k, f, X, x) := \frac{|s_X^f(x) - f(x)|}{\mathbf{P}_X(x)}$$

thus a fraction of an "f" and "P" dependent term motivating the notion "f/P"–Greedy. The following lemma, which extends the results in [5] to matrix-valued kernels, shows that the right hand side in (16) is equal to the gain in the square of the Hilbert space norm of the interpolant $\|s_X^f\|_{\mathcal{H}_k}^2$, when the set of centers X is enriched by x:

Lemma 1 (Local optimality of the f/P–Greedy selection rule). *Let $k : \Omega \times \Omega \to \mathbb{R}^{m\times m}$ be a positive definite matrix-valued kernel, $f \in \mathcal{H}_k$ and $X = \{x_1, \ldots, x_n\} \subset \Omega$ a finite set of pairwise distinct points. Let $s_X^f \in \mathcal{N}(X)$ denote the unique interpolant of f on the centers X. Then it holds for all $x \in \Omega$:*

$$\|s_{X\cup\{x\}}^f\|_{\mathcal{H}_k}^2 = \|s_X^f\|_{\mathcal{H}_k}^2 + (s_X^f(x) - f(x))^T \mathbf{P}_X(x)^+ (s_X^f(x) - f(x)). \tag{17}$$

Proof. We restrict ourselves to the strictly p.d. case. For the non-strictly p.d. case technical consideration of the null spaces of the kernel matrices is required without major change of the main arguments. For suitable coefficients the interpolants can be expressed as

$$s_{X\cup\{x\}}^f = k(\cdot, X)\boldsymbol{\alpha} + k(\cdot, x)\alpha_{n+1}, \quad \boldsymbol{\alpha} \in \mathbb{R}^{mn}, \alpha_{n+1} \in \mathbb{R}^m$$

and

$$s_X^f = k(\cdot, X)\boldsymbol{\beta}, \quad \boldsymbol{\beta} \in \mathbb{R}^{mn}.$$

Furthermore, let $A := k(X, X)$ and $B := k(X, x)$. Since both $s_{X\cup\{x\}}^f$ and s_X^f interpolate f on X we have

$$A\boldsymbol{\alpha} + B\alpha_{n+1} = A\boldsymbol{\beta} \quad \Longleftrightarrow \quad \boldsymbol{\beta} = \boldsymbol{\alpha} + A^{-1}B\alpha_{n+1}. \tag{18}$$

For the norm of $s_{X\cup\{x\}}^f$ it holds

$$\|s_{X\cup\{x\}}^f\|_{\mathcal{H}_k}^2 = \boldsymbol{\alpha}^T A\boldsymbol{\alpha} + 2\boldsymbol{\alpha}^T B\alpha_{n+1} + \alpha_{n+1}^T k(x, x)\alpha_{n+1}$$

and using (18) we get

$$\begin{aligned}\|s_X^f\|_{\mathcal{H}_k}^2 &= \boldsymbol{\beta}^T A \boldsymbol{\beta} \\ &= \boldsymbol{\alpha}^T A \boldsymbol{\alpha} + 2\boldsymbol{\alpha}^T B \alpha_{n+1} + \alpha_{n+1}^T B^T A^{-1} B \alpha_{n+1} \\ &= \|s_{X\cup\{x\}}^f\|_{\mathcal{H}_k}^2 - \alpha_{n+1}^T \left(k(x,x) - B^T A^{-1} B\right) \alpha_{n+1} \\ &= \|s_{X\cup\{x\}}^f\|_{\mathcal{H}_k}^2 - \alpha_{n+1}^T \mathbf{P}_X(x) \alpha_{n+1}. \end{aligned} \tag{19}$$

For the difference between the target function and the interpolant on X we again have via (18)

$$\begin{aligned} f(x) - s_X^f(x) = s_{X\cup\{x\}}^f(x) - s_X^f(x) &= B^T \boldsymbol{\alpha} + k(x,x)\alpha_{n+1} - B^T \boldsymbol{\beta} \\ &= \left(k(x,x) - B^T A^{-1} B\right) \alpha_{n+1} = \mathbf{P}_X(x) \alpha_{n+1}. \end{aligned} \tag{20}$$

Combining (19) and (20) concludes the proof as

$$\begin{aligned}\|s_X^f\|_{\mathcal{H}_k}^2 &= \|s_{X\cup\{x\}}^f\|_{\mathcal{H}_k}^2 - \alpha_{n+1}^T \mathbf{P}_X(x) \alpha_{n+1} \\ &= \|s_X^f\|_{\mathcal{H}_k}^2 + (s_X^f(x) - f(x))^T \mathbf{P}_X(x)^{-1} (s_X^f(x) - f(x)).\end{aligned}$$

□

Similar to the f–Greedy, the f/P–Greedy is more expensive than the P–Greedy and in general the indicator is not monotically decreasing. However, due to the interpolation property we have $E_3(k, f, X, x) = 0$ for all $x \in X$ and thus the algorithm again terminates with a finite number of centers.

3 Numerical Example

In this section we want to investigate the effect of the different error indicator function on the quality of the approximation and the placement of the centers. For this purpose we consider the unit disc segment $\Omega = \{x = (r\cos(\varphi), r\sin(\varphi))^T \in \mathbb{R}^2 | \, (r,\varphi)^T \in \tilde{\Omega}\}$ with $\tilde{\Omega} = [0,1] \times [\frac{1}{3}\pi, \frac{5}{3}\pi]$, the target function $f = (f_i)_{i=1}^8 : \Omega \to \mathbb{R}^8$ given by

$$f_i(x) := \sum_{j=1}^{10} e^{-\lfloor (i+1)/2 \rfloor \|x - x_j\|^2}, \quad i = 1, \ldots, 8,$$

with $x_1 = (0,0)^T$ and $x_j = 0.1(\cos(\frac{j}{6}\pi), \sin(\frac{j}{6}\pi))^T, j = 2, \ldots, 10$ and the kernel $k : \Omega \times \Omega \to \mathbb{R}^{8\times 8}$ given by a diagonal Gaussian with decaying widths

$$k_{i,j}(x,y) := \begin{cases} e^{-\lfloor (i+1)/2 \rfloor \|x-y\|^2}, & i = j \\ 0, & i \neq j \end{cases}$$

By straightforward computation one can see that $f(x) = k(x, Y)\mathbf{1}$ where $Y = \{x_1, \ldots, x_{10}\}$ and $\mathbf{1} \in \mathbb{R}^{80}$ is the vector containing only ones. In particular we

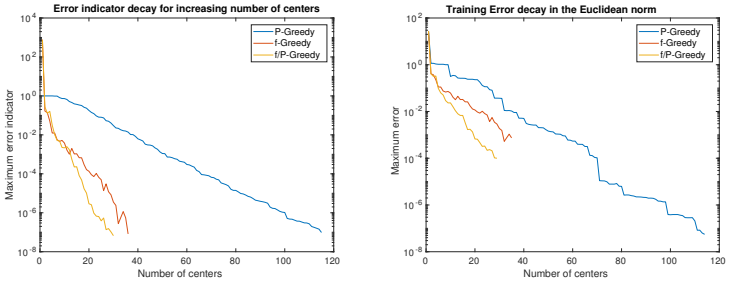

Fig. 1. Error indicator decay (left) and maximum training error decay in the Euclidean norm (right) for increasing number of centers.

have $\|f\|^2_{\mathcal{H}_k} = \mathbf{1}^T k(Y,Y)\mathbf{1} \approx 768.295$. For the Greedy algorithm we choose Ω_N by transforming 50×50 uniformly distributed points in $\tilde{\Omega}$ to rectangular coordinates, which results in 2451 sample points and use the tolerance $\varepsilon = 10^{-7}$. The sets of centers which are generated are denoted by X_i, $i = 1, 2, 3$ where the index corresponds to the index of the respective error indicator function E_i, $i = 1, 2, 3$ from Section 2. In Figure 1 the decay of the error indicator (maximum E_i over the training set Ω_N) and maximum training error for an increasing number of centers are depicted. As we can see, it takes 114 iterations for the P–Greedy algorithm to terminate, where only 35 (f–Greedy) and 29 (f/P–Greedy) are required for the other algorithms. This is caused by the slow decay in the Power-function which in itself is caused by the narrow Gaussians which model the last target function components. As we mentioned before in Section 2.1, we can see in Figure 2 that the set X_1 is somewhat uniformly distributed while X_3 is clearly not space filling.

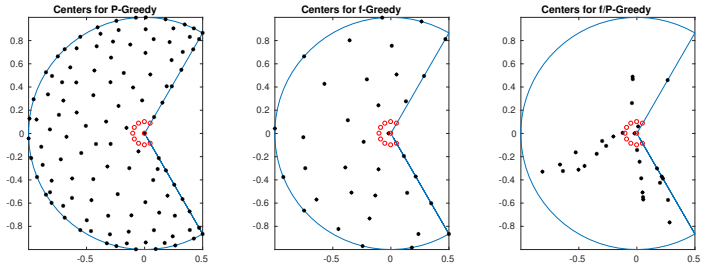

Fig. 2. Distributions of the centers X_1, X_2 and X_3.

In Figure 3 the decay for the maximum test error in the Euclidean norm err_i^2 on the test set Ω_M generated by transforming 100×100 uniformly distributed points in $\tilde{\Omega}$, and Hilbert space norm err_i^k, $i = 1, 2, 3$ are shown. While

the error in the Hilbert space norm is monotically decaying for any choice of E_i. This is not the case for the Euclidean norm. However, in both cases the f/P–Greedy generates the best sets with regards to the number of centers that are used in the interpolant expansion. For example, the P-Greedy algorithm takes about 70 iteration to reach a Euclidean error of order 10^{-4}, while the f/P–Greedy requires 29 to reach the same result. Overall, all three variants generate very sparse kernel-based models. Future work will aim at surrogate modelling for engineering applications with those techniques.

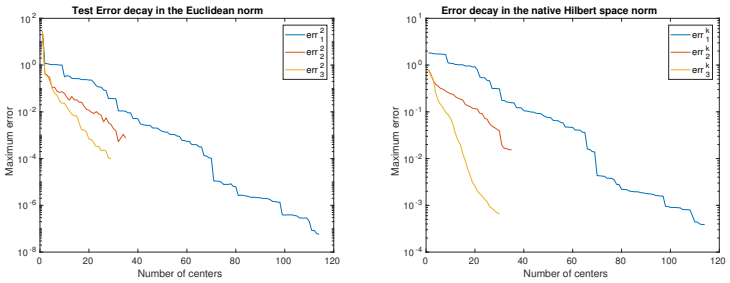

Fig. 3. Test error decay in the Euclidean norm (left) and in the Hilbert space norm (right) for increasing number of centers.

Acknowledgements

We thank Gabriele Santin for fruitful discussions.

References

1. M. Alvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: a review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012.
2. Stefano De Marchi, Robert Schaback, and Holger Wendland. Near-optimal data-independent point locations for radial basis function interpolation. *Adv. Comput. Math.*, 23(3):317–330, 2005.
3. C. A. Micchelli and M. Pontil. Kernels for multi-task learning. *Advances in Neural Information Processing Systems*, 2004.
4. G. Santin and B. Haasdonk. Convergence rate of the data-independent P-greedy algorithm in kernel-based approximation. *Dolomites Research Notes on Approximation*, 10:68–78, 2017.
5. R. Schaback and J. Werner. Linearly constrained reconstruction of functions by kernes with applications to machine learning. *Advances in Computational Mathematics*, (25):237–258, 2006.
6. D. Wittwar, G. Santin, and B. Haasdonk. Interpolation with uncoupled separable matrix-valued kernels. Technical report, University of Stuttgart, 2017. In preparation.