On estimation of surrogate models for multivariate computer experiments[‡]

Benedikt Bauer[1],Felix Heimrich[2], Michael Kohler[1] and Adam Krzyżak[3,§]

[1] *Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: bbauer@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de*

[2] *Fachbereich Maschinenbau, Technische Universität Darmstadt, Otto-Bernd-Str. 2, 64287 Darmstadt, Germany, email: heimrich@dik.tu-darmstadt.de*

[3] *Department of Computer Science and Software Engineering, Concordia University, 1455 De Maisonneuve Blvd. West, Montreal, Quebec, Canada H3G 1M8, email: krzyzak@cs.concordia.ca*

Supplementary material for the referees

Proof of Lemma 3.

In the proof we will use Proposition 3.8 in Mhaskar (1993), which we reformulate here (in a slightly different form) as Lemma 6.

Lemma 6. *Let $K \subseteq \mathbb{R}^d$ be a polytope bounded by hyperplanes $v_j \cdot x + w_j \geq 0$ $(j = 1, \ldots, L)$, where $v_1, \ldots, v_L \in \mathbb{R}^d$ and $w_1, \ldots, w_L \in \mathbb{R}$. For $\delta > 0$ set*

$$K_\delta^0 := \left\{ x \in \mathbb{R}^d \, : \, v_j \cdot x + w_j \geq \delta \quad \text{for all } j \in \{1, \ldots, L\} \right\}$$

and

$$K_\delta^c := \left\{ x \in \mathbb{R}^d \, : \, v_j \cdot x + w_j \leq -\delta \quad \text{for some } j \in \{1, \ldots, L\} \right\}.$$

Let $\sigma : \mathbb{R} \to [0,1]$ be a squashing function. Let $\varepsilon, \delta \in (0,1]$ be arbitrary. Then there exists a neural network of the form

$$f(x) = \sigma \left(\sum_{j=1}^{L} b_j \cdot \sigma \left(\sum_{k=1}^{d} a_{j,k} \cdot x^{(k)} + a_{j,0} \right) + b_0 \right)$$

satisfying

$$|f(x)| \leq 1 \text{ for } x \in \mathbb{R}^d,$$
$$|f(x) - 1| \leq \varepsilon \text{ for } x \in K_\delta^0,$$

[‡]Running title: *On estimation of surrogate models*

The online version of this article contains supplementary material.

[§]Corresponding author. Tel: +1-514-848-2424, ext. 3007, Fax: +1-514-848-2830.

$$|f(x)| \leq \varepsilon \text{ for } x \in K_\delta^c. \tag{33}$$

In case that the squashing function satisfies

$$|\sigma(y) - 1| \leq \frac{1}{y} \quad \text{if} \quad y > 0 \quad \text{and} \quad |\sigma(y)| \leq \frac{1}{|y|} \quad \text{if} \quad y < 0,$$

the weights above can be chosen such that

$$|b_j| \leq \frac{4L}{\varepsilon} \quad \text{for all } j = 0, \ldots, L$$
$$|a_{j,k}| \leq \frac{4L}{\delta} \cdot \max\{\|v_1\|_\infty, |w_1|, \ldots, \|v_L\|_\infty, |w_L|\} \quad \text{for all } j = 1, \ldots, L, k = 0, \ldots, d.$$

Proof. Follows from the proof of Proposition 3.8 in Mhaskar (1993). □

Proof of Lemma 3. We partition $\left[-a - \frac{2a}{M}, a\right]^d$ into $(M+1)^d$ equivolume cubes of side length $2a/M$. Approximating m by a piecewise constant approximant with respect to this partition yields (since m is (p, C)–smooth with $p \leq 1$) a function S satisfying

$$\|S - m\|_{\infty,[-a,a]^d} \leq \sqrt{d} \cdot C \cdot \left(\frac{2a}{M}\right)^p. \tag{34}$$

If we we choose S suitably, it can be expressed in the form

$$S(x) = m(x_{(1,\ldots,1)}) + \sum_{i \in \{1,\ldots,M+1\}^d \setminus \{(1,\ldots,1)\}} d_i \cdot \prod_{j=1}^{d} \left(x^{(j)} - x_i^{(j)}\right)_+^0,$$

where x_i are the corners of the cubes forming the above partition (indexed in ascending order per component), $0^0 := 0$, $x_+ := \max\{x, 0\}$, and d_i for $i = (i_1, \ldots, i_d)$ as above are constants satisfying

$$d_i = \sum_{J \subseteq \{1,\ldots,d\} \setminus \{k \,:\, i_k = 1\}} (-1)^{|J|} \cdot m\left(x_{i-J}\right), \tag{35}$$

where $i - J$ symbolizes the index i with i_j replaced by $i_j - 1$ for all $j \in J$. Since for a fixed set with $n > 0$ elements the number of subsets with even and uneven cardinality is 2^{n-1}, respectively, and the corners used in the above expression have a distance of at most $\sqrt{d} \cdot \frac{2a}{M}$, we can conclude from the (p, C)–smoothness of m

$$|d_i| \leq 2^{d - |\{k \,:\, i_k = 1\}| - 1} \cdot C \cdot d^{\frac{p}{2}} \cdot \left(\frac{2a}{M}\right)^p \leq c_{14} \cdot \left(\frac{2a}{M}\right)^p. \tag{36}$$

Let K_i be the polytope defined by $x^{(j)} - x_i^{(j)} \geq 0$ $(j = 1, \ldots, d)$. Set $\varepsilon = (M+1)^{-d}$, $\delta = a \cdot \eta / (2 \cdot d \cdot M)$ and apply Lemma 6 for each K_j (i.e., with $L = d$, $v_j = \mathbf{e}_j$ and $w_j = -x_i^{(j)}$, where $\mathbf{e}_j$ denotes the j-th unit vector) to obtain $f_i(x)$ satisfying (33) with K_i instead of K. Let

$$P(x) = m(x_{(1,\ldots,1)}) + \sum_{i \in \{1,\ldots,M+1\}^d \setminus \{(1,\ldots,1)\}} d_i \cdot f_i(x).$$

Then we can conclude from (33) and (36)

$$\begin{aligned}|P(x) - S(x)| &\leq \sum_{i \in \{1,\ldots,M+1\}^d \setminus \{(1,\ldots,1)\}} |d_i| \cdot \left| f_i(x) - \prod_{j=1}^{d} \left(x^{(j)} - x_i^{(j)}\right)_+^0 \right| \\ &\leq \sum_{i \in \{1,\ldots,M+1\}^d \setminus \{(1,\ldots,1)\}} |d_i| \cdot (M+1)^{-d} \\ &\leq c_{14} \cdot \left(\frac{2a}{M}\right)^p \end{aligned} \tag{37}$$

for all $x \in [-a,a]^d$ which are not contained in

$$\bigcup_{j=1,\ldots,d} \bigcup_{i \in \{1,\ldots,M+1\}^d} \left\{ x \in \mathbb{R}^d \quad : \quad |x^{(j)} - x_i^{(j)}| < a \cdot \eta/(2 \cdot d \cdot M) \right\}. \tag{38}$$

By shifting the positions of the x_i in the jth component slightly to the right (in the sense of increasing values) we can construct

$$\left\lfloor \frac{2a/M}{2\delta} \right\rfloor = \left\lfloor \frac{2a}{M} \cdot \frac{2 \cdot d \cdot M}{2 \cdot a \cdot \eta} \right\rfloor = \left\lfloor \frac{2 \cdot d}{\eta} \right\rfloor \geq d/\eta$$

different versions of P, that still satisfy (34) and (37) for all $x \in [-a,a]^d$, and corresponding disjoint versions of

$$\bigcup_{i \in \{1,\ldots,M+1\}^d} \left\{ x \in \mathbb{R}^d \quad : \quad |x^{(j)} - x_i^{(j)}| < a \cdot \eta/(2 \cdot d \cdot M) \right\},$$

and since the sum of the ν–measures of these sets is less than or equal to one, at least one of them must have measure less than or equal to η/d. Consequently we can shift the x_i such that (38) has ν–measure less than or equal to η. This together with (34) and (37) implies the first assertion of the lemma, because $P(x)$ complies with the structure of the postulated neural network $t(x)$.
In case that σ satisfies the conditions specified in the second part of the lemma, Lemma 6 allows to bound the coefficients of the neural network $t(x) := P(x)$ respecting the values of the parameters we used during the application of this lemma above. This leads to

$$|a_{i,j,k}| \leq \frac{4 \cdot d \cdot 2 \cdot d \cdot M}{a \cdot \eta} \cdot \max\left\{1, a + \frac{2a}{M}\right\} \leq 8 \cdot d^2 \cdot \frac{M}{\eta} \cdot \max\left\{\frac{1}{a}, 3\right\}$$

for all $i \in \left\{1,\ldots,(M+1)^d\right\}$, $j \in \{1,\ldots,d\}$, $k \in \{0,\ldots,d\}$ and

$$|b_{i,j}| \leq 4 \cdot d \cdot (M+1)^d$$

for all $i \in \left\{1,\ldots,(M+1)^d\right\}$, $j \in \{0,\ldots,d\}$. Furthermore, the definition of P and (35) imply

$$|d_i| \leq 2^d \cdot \|m\|_\infty$$

for all $i \in \left\{0,\ldots,(M+1)^d\right\}$, which leads to the second assertion of the lemma. □

In order to prove Lemma 5, we introduce the following technical result.

Lemma 7. *Let $l \in \mathbb{N}_0$ and let $\sigma_r : \mathbb{R} \to \mathbb{R}$ for $r = 1, \dots, l+1$ be Lipschitz continuous functions with Lipschitz constant $L \geq 1$, which satisfy*

$$|\sigma_r(x)| \leq L \cdot \max\{|x|, 1\} \quad (x \in \mathbb{R}). \tag{39}$$

Let $K_0 = d$, $K_r \in \mathbb{N}$ for $r \in \{1, \dots, l\}$ and $K_{l+1} = 1$. For $r \in \{1, \dots, l+1\}$ and $i \in \{1, \dots, K_r\}$ define recursively

$$f_i^{(r)}(x) = \sigma_r \left(\sum_{j=1}^{K_{r-1}} c_{i,j}^{(r-1)} \cdot f_j^{(r-1)}(x) + c_{i,0}^{(r-1)} \right)$$

and

$$\bar{f}_i^{(r)}(x) = \sigma_r \left(\sum_{j=1}^{K_{r-1}} \bar{c}_{i,j}^{(r-1)} \cdot \bar{f}_j^{(r-1)}(x) + \bar{c}_{i,0}^{(r-1)} \right),$$

where $c_{i,0}^{(r-1)}, \bar{c}_{i,0}^{(r-1)}, \dots, c_{i,K_{r-1}}^{(r-1)}, \bar{c}_{i,K_{r-1}}^{(r-1)} \in \mathbb{R}$, and $f_j^{(0)}(x) = \bar{f}_j^{(0)}(x) = x^{(j)}$. Furthermore, set

$$\overline{C} = \max_{\substack{r=0,\dots,l, i=1,\dots,K_{r+1}, \\ j=1,\dots,K_r}} \max\left\{ \left| c_{i,j}^{(r)} \right|, \left| \bar{c}_{i,j}^{(r)} \right|, 1 \right\}.$$

Then

$$\begin{aligned} &\left| f_1^{(l+1)}(x) - \bar{f}_1^{(l+1)}(x) \right| \\ &\leq (l+1) \cdot L^{l+1} \cdot \prod_{r=0}^{l} (K_r + 1) \cdot \overline{C}^l \cdot \max\{\|x\|_\infty, 1\} \cdot \max_{\substack{r=0,\dots,l, i=1,\dots,K_{r+1}, \\ j=0,\dots,K_r}} \left| c_{i,j}^{(r)} - \bar{c}_{i,j}^{(r)} \right| \end{aligned}$$

for any $x \in \mathbb{R}^d$.

Proof. At first, we notice that (39) implies

$$\left| \bar{f}_i^{(r)}(x) \right| \leq L \cdot (K_{r-1} + 1) \cdot \overline{C} \cdot \max_{j=1,\dots,K_{r-1}} \left\{ \left| \bar{f}_j^{(r-1)}(x) \right|, 1 \right\}$$

for $r = 1, \dots, l$ and $i = 1, \dots, K_r$, from which we can conclude

$$\left| \bar{f}_i^{(r)}(x) \right| \leq L^r \cdot \prod_{\tilde{r}=1}^{r} (K_{\tilde{r}-1} + 1) \cdot \overline{C}^r \cdot \max\{\|x\|_\infty, 1\}. \tag{40}$$

Using the Lipschitz continuity of σ_r and the triangle inequality in combination with (40) we get

$$\begin{aligned} &\left| f_i^{(r)}(x) - \bar{f}_i^{(r)}(x) \right| \\ &\leq L \cdot \left| \sum_{j=1}^{K_{r-1}} c_{i,j}^{(r-1)} \cdot f_j^{(r-1)}(x) + c_{i,0}^{(r-1)} - \sum_{j=1}^{K_{r-1}} \bar{c}_{i,j}^{(r-1)} \cdot \bar{f}_j^{(r-1)}(x) - \bar{c}_{i,0}^{(r-1)} \right| \end{aligned}$$

$$
\begin{aligned}
&\leq L \cdot \left(\sum_{j=1}^{K_{r-1}} \left| c_{i,j}^{(r-1)} \right| \cdot \left| f_j^{(r-1)}(x) - \bar{f}_j^{(r-1)}(x) \right| \right. \\
&\qquad \left. + \sum_{j=1}^{K_{r-1}} \left| c_{i,j}^{(r-1)} - \bar{c}_{i,j}^{(r-1)} \right| \cdot \left| \bar{f}_j^{(r-1)}(x) \right| + \left| c_{i,0}^{(r-1)} - \bar{c}_{i,0}^{(r-1)} \right| \right) \\
&\leq L \cdot K_{r-1} \cdot \overline{C} \cdot \max_{j=1,\dots,K_{r-1}} \left| f_j^{(r-1)}(x) - \bar{f}_j^{(r-1)}(x) \right| \\
&\qquad + L \cdot (K_{r-1}+1) \cdot L^{r-1} \cdot \prod_{\tilde{r}=1}^{r-1} (K_{\tilde{r}-1}+1) \cdot \overline{C}^{\,r-1} \cdot \max\left\{ \|x\|_\infty, 1 \right\} \\
&\qquad\qquad \cdot \max_{j=0,\dots,K_{r-1}} |c_{i,j}^{(r-1)} - \bar{c}_{i,j}^{(r-1)}| \\
&= L \cdot K_{r-1} \cdot \overline{C} \cdot \max_{j=1,\dots,K_{r-1}} \left| f_j^{(r-1)}(x) - \bar{f}_j^{(r-1)}(x) \right| \\
&\qquad + L^r \cdot \prod_{\tilde{r}=1}^{r} (K_{\tilde{r}-1}+1) \cdot \overline{C}^{\,r-1} \cdot \max\left\{ \|x\|_\infty, 1 \right\} \cdot \max_{j=0,\dots,K_{r-1}} |c_{i,j}^{(r-1)} - \bar{c}_{i,j}^{(r-1)}|
\end{aligned}
$$

for all $r = 1, \dots, l+1$. Now we start with the above inequality for $r = l+1$ and plug it in repeatedly for decreasing r in the expression $\left| f_j^{(r-1)}(x) - \bar{f}_j^{(r-1)}(x) \right|$ on the right-hand side of the inequality. Finally, the summand containing $\left| f_j^{(0)}(x) - \bar{f}_j^{(0)}(x) \right|$ vanishes in the case of $r = 1$, which implies the assertion. □

Proof of Lemma 5. At first, we notice the space $\mathcal{F}_n = \mathcal{H}^{(l)}$ (with $l > 0$) can be expressed as

$$
\begin{aligned}
\mathcal{H}^{(l)} \quad = \quad \Bigg\{ & h : \mathbb{R}^d \to \mathbb{R} \ : \ h(x) = \sum_{k=1}^{K} \sigma_{id} \left(g_k \left(\sigma_{id} \left(f_{1,k}(x) \right), \dots, \sigma_{id} \left(f_{d^*,k}(x) \right) \right) \right) \quad (x \in \mathbb{R}^d) \\
& \text{for some } g_k \in \mathcal{F}_{M_n,d^*,d^*,\alpha,\beta,\gamma}^{(\textit{neural networks})} \text{ and } f_{j,k} \in \mathcal{H}^{(l-1)} \Bigg\},
\end{aligned}
$$

where $\sigma_{id} : \mathbb{R} \to \mathbb{R}$ is the identity $\sigma_{id}(x) = x$ for all $x \in \mathbb{R}$. Furthermore, all $g \in \mathcal{F}_{M_n,d^*,d^*,\alpha,\beta,\gamma}^{(\textit{neural networks})}$ can be written as

$$
\begin{aligned}
g(x) &= \sum_{i=1}^{(M_n+1)^{d^*}} d_i \cdot \sigma \left(\sum_{j=1}^{d^*} b_{i,j} \cdot \sigma \left(\sum_{m=1}^{d^*} a_{i,j,m} \cdot x^{(m)} + a_{i,j,0} \right) + b_{i,0} \right) + d_0 \\
&= \sum_{i=1}^{(M_n+1)^{d^*}} d_i \cdot \sigma \left(\sum_{\substack{j=1,\dots,d^*, \\ \bar{i}=1,\dots,(M_n+1)^{d^*}}} b_{i,\bar{i},j} \cdot \sigma \left(\sum_{m=1}^{d^*} a_{\bar{i},j,m} \cdot x^{(m)} + a_{\bar{i},j,0} \right) + b_{i,\bar{i},0} \right) + d_0,
\end{aligned}
$$

where the new coefficients are defined by

$$b_{i,\bar{i},j} := \begin{cases} b_{i,j} & \text{if } \bar{i} = i \\ 0 & \text{otherwise} \end{cases}$$

for all $i, \bar{i} \in \left\{1, \ldots, (M_n+1)^{d^*}\right\}$ and $j \in \{0, \ldots, d^*\}$ (which works analogously for $h \in \mathcal{H}^{(0)}$). Respecting the above representations, all the functions $\sigma_{id}(h) = h$ for $h \in \mathcal{H}^{(l)}$ comply with the structure of the functions $f_1^{(l+1)}$ in Lemma 7, if we use the following specifications of the parameters in that lemma: The Lipschitz constant L is chosen as the maximum of the Lipschitz constants of σ_{id} (which is obviously 1) and the squashing function σ from Theorem 3. Thus, the property (39) is satisfied due to $\|\sigma\|_\infty \leq 1$, $L \geq 1$, and $|\sigma_{id}(x)| = |x|$. The parameter l in Lemma 7 is $4l+2$ (regarding the l in $\mathcal{H}^{(l)}$ above) and the parameters K_r with $r = 0, \ldots, l$ take repeatedly the values $\tilde{d}, d^* \cdot (M_n+1)^{d^*}, (M_n+1)^{d^*}, K$ one after another, where $\tilde{d}$ is equal to d^* except for K_0, where it is d. Since all the coefficients $c_{i,j}^{(r)}$ with $r = 0, \ldots, l$, $i = 1, \ldots, K_{r+1}$, $j = 1, \ldots, K_r$ (using $K_{l+1} = 1$ again) are 0, 1, or one of the $a_{i,j,m}, b_{i,j}, d_i$ in the definition of $\mathcal{F}_{M_n,d^*,d,\alpha,\beta,\gamma}^{(neural\ networks)}$, we can use $\overline{C} = \max\{\alpha, \beta, \gamma\}$ for n sufficiently large.
Let h and $\bar{h}$ be functions in $\mathcal{F}_n$. Since they comply with the structure of the functions in Lemma 7 according to the above argumentation, we can conclude

$$\begin{aligned} &\|h - \bar{h}\|_{\infty,[-a_n,a_n]^d} \\ &\leq (4l+3) \cdot L^{4l+3} \cdot \left(d^* \cdot (M_n+1)^{d*} + 1\right)^{4l+3} \cdot \max\{\alpha, \beta, \gamma\}^{4l+2} \\ &\qquad \cdot \max\{a_n, 1\} \cdot \max_{\substack{r=0,\ldots,\tilde{l},\ i=1,\ldots,K_{r+1},\\ j=0,\ldots,K_r}} \left|c_{i,j}^{(r)} - \bar{c}_{i,j}^{(r)}\right| \\ &\leq a_n \cdot n^{c_{11}} \cdot \max_{\substack{r=0,\ldots,\tilde{l},\ i=1,\ldots,K_{r+1},\\ j=0,\ldots,K_r}} \left|c_{i,j}^{(r)} - \bar{c}_{i,j}^{(r)}\right| \end{aligned}$$

for n sufficiently large and an adequately chosen $c_{11} > 0$. Thus, if we consider an arbitrary $h \in \mathcal{H}^{(l)}$, it suffices to choose the coefficients $\bar{c}_{i,j}^{(r)}$ of a function $\bar{h} \in \mathcal{H}^{(l)}$ such that

$$\left|c_{i,j}^{(r)} - \bar{c}_{i,j}^{(r)}\right| \leq \frac{\varepsilon_n}{a_n \cdot n^{c_{11}}} \tag{41}$$

holds for all possible indices, in order to satisfy $\|h(x) - \bar{h}(x)\|_{\infty,[-a_n,a_n]^d} \leq \varepsilon_n$. For n sufficiently large, which is assumed permanently in the following, the coefficients $c_{i,j}^{(r)}$ have to take values in $[-\max\{\alpha, \beta, \gamma\}, \max\{\alpha, \beta, \gamma\}]$ and the relations $\max\{\alpha, \beta, \gamma\} \leq \log(n) \cdot n^2 \cdot M_n^{d^*} \leq n^4$ and $a_n \leq M_n \leq n$ hold. Then due to $\varepsilon_n = \frac{a_n^p}{M_n^p}$ a number of

$$\left\lceil \frac{2 \cdot \max\{\alpha, \beta, \gamma\} \cdot a_n \cdot n^{c_{11}}}{2 \cdot \varepsilon_n} \right\rceil \leq n^{c_{12}}$$

different $\bar{c}_{i,j}^{(r)}$ suffices to guarantee, that at least one of them satisfies the relation (41) for any $c_{i,j}^{(r)}$ with fixed indices. Furthermore, the coefficients $c_{i,j}^{(r)}$, which can actually differ

regarding different $h \in \mathcal{H}^{(l)}$, are the ones originating from the coefficients $a_{i,j,m}, b_{i,j}, d_i$ in the definition of $\mathcal{F}_{M_n,d^*,d,\alpha,\beta,\gamma}^{(neural\ networks)}$. Using (22), their number can be bounded by $c_{13} \cdot M_n^{d^*}$. So the logarithm of the covering number $\mathcal{N}(\varepsilon_n, \mathcal{F}_n, \|\cdot\|_{\infty,[-a_n,a_n]^d})$ can be bounded by

$$\log\left(\mathcal{N}(\varepsilon_n, \mathcal{F}_n, \|\cdot\|_{\infty,[-a_n,a_n]^d})\right) \leq \log\left((n^{c_{12}})^{c_{13}\cdot M_n^{d^*}}\right) \leq c_{10} \cdot \log(n) \cdot M_n^{d^*},$$

which proves the assertion. □