MA4199 Project – Bias Variance Tradeoff

Ng Wei Le

1 Kernels

Notation: we use the symbol $\mathbb{K}$ when it can refer to both $\mathbb{R}$ or $\mathbb{C}$. Also, let z^* or $(z)^*$ denote the conjugate of z for any $z \in \mathbb{C}$. The sections covering Kernels and reproducing kernel Hilbert spaces are heavily referenced using Steinwart, Christman [1].

Definition 1. For a non-empty set X, let $k : X \times X \to \mathbb{K}$ be known as a kernel if there exists a function $\phi : X \to \mathcal{H}$ (known as a feature map of k) where $\mathcal{H}$ is a $\mathbb{K}$-Hilbert space (known as a feature space of k) such that

$$k(x_1, x_2) = \langle \phi(x_2), \phi(x_1) \rangle_{\mathcal{H}}. \tag{1}$$

Lemma 1. *For any kernel k on X, $k(x_1, x_2) = k(x_2, x_1)^*$.*

From the properties of the inner product, we know that $k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle^* = k(x_2, x_1)^*$. Therefore, for kernels on $\mathbb{R}$, the symmetric property: $k(x_1, x_2) = k(x_2, x_1)$ holds.

Below, we define the Gaussian RBF kernel:

Definition 2. Let the complex Gaussian RBF kernel be:

$$k_{\gamma, \mathbb{C}^d}(z, z') := e^{-\gamma^{-2} \sum_{i=1}^{d} (z_i - z_i'^*)^2}.$$

We then define the real Gaussian RBF kernel (or simply the Gaussian RBF kernel for short) as:

$$k_\gamma(x, x') = e^{-\gamma^{-2} \|x - x'\|_2^2}$$

Lemma 2. *Let k_1, k_2 be kernels on a non-empty set X. Then $k_1 + k_2$ and $ak_1, a \in \mathbb{R}^+ \cup \{0\}$ are kernels.*

Definition 3. For a non-empty set X, a function $k : X \times X \to \mathbb{R}$ is said to be a positive definite if, for any $m \in \mathbb{Z}^+ \cup \{0\}$ and all $x_1, ..., x_n \in X$, we have the following matrix (called the Gram matrix) being positive semi-definite:

$$K := (k(x_i, x_j))_{i,j}.$$

Equivalently: for all $a_1, ..., a_n \in \mathbb{R}$, we have:

$$\sum_{j=1}^{n} \sum_{i=1}^{n} a_j a_i k(x_j, x_i).$$

Definition 4. The positive definite function $k : X \times X \to \mathbb{R}$ is said to be symmetric if $k(x_1, x_2) = k(x_2, x_1)$ for all $x_1, x_2 \in X$

Theorem 1. *A real function $k : X \times X \to \mathbb{R}$ is a kernel if and only if k is a positive definite symmetric function (also known as a positive definite kernel).*

2 Reproducing Kernel Hilbert Spaces

Initially introduced by Stanislaw Zaremba, reproducing kernel Hilbert spaces have many applications in the fields such as Statistical Learning and complex analysis. An RKHS is a $\mathbb{K}$-Hilbert function space where point evaluation is continuous linear funcitonal.

Definition 5. (RKHS). Let $\mathcal{H}$ be a $\mathbb{K}$-Hilbert space of functions over a non-empty set X. $\mathcal{H}$ is called an RKHS over X if the Dirac function $\delta_x : \mathcal{H} \to \mathbb{K}$ defined as:

$$\delta_x(f) := f(x), \ x \in X, \ f \in \mathcal{H}$$

is continuous. Equivalently, there exists $0 < M_x < \infty$ such that

$$\delta_x(f) \leq M_x \|f\|_{\mathcal{H}}, \text{ for all} f \in \mathcal{H}.$$

δ_x is called a bounded operator on $\mathcal{H}$.

This is not easy to put into practice, hence the reproducing kernel is defined.

Definition 6. (Reproducing Kernel). For a non-empty set X and a function $k : X \times X \to \mathbb{K}$ where $k(\cdot, x) \in \mathcal{H}$ for all $x \in X$ and the following property hold for all $xinX$ and $f \in \mathcal{H}$:

$$f(x) = \langle f, k(\cdot, x) \rangle \tag{2}$$

The condition in equation (2) is also known as the reproducing property.

Definition 7. (Canonical Feature Maps). Let $\mathcal{H}$ be an RKHS over X with reproducing kernel k. Let the function $\Phi : X \to \mathcal{H}$ be defined such that for all $x \in X$,

$$\Phi(x) = k(\cdot, x).$$

We call Φ the canonical feature map of k.

Lemma 3. ***(A reproducing kernel of an RKHS is a kernel).*** *Let $\mathcal{H}$ be an RKHS over X with reproducing kernel k. Then k is a kernel.*

Proof. We simply proof that Φ is a feature map of k.

$$\begin{aligned} \langle \Phi(x_2), \Phi(x_1) \rangle &= \langle k(\cdot, x_2), k(\cdot, x_1) \rangle \\ &= k(x_1, x_2) \ (\because \text{Reproducing Property (2)}) \end{aligned}$$

So $\mathcal{H} is also a feature space of k.$ □

Lemma 4. *Let $\mathcal{H}$ be an $\mathbb{K}$-Hilbert functional RKHS over X with reproducing kernel k. Then H is a Reproducing Kernel Hilbert Space.*

Proof. Recall the Dirac functional $\delta_x : H \to \mathbb{K}$ where:

$$\delta_x(f) = f(x),\ x \in X,\ f \in H.$$

Then we have:

$$\begin{aligned}|\delta_x(f)| &= |f(x)| \\ &= |\langle f, k(\cdot, x)\rangle| \ (\because \text{Reproducing Property (2)}) \\ &\leq \|k(\cdot, x)\|_{\mathcal{H}} \, \|f\|_{\mathcal{H}} \ \ (\because \text{Cauchy-Schwarz Inequality})\end{aligned}$$

This shows that the Dirac functionals are continuous. □

2.1 Representer Theorem

Representer Theorem ensures that the *argmin* of an empirical risk expression involving a function over an RKHS can be expressed as a linear combination of kernels applied on the training data points as proven in [7].

Theorem 2. *Given a non-empty set X, training data $\{(x_1, y_1), ...(x_n, y_n)\} \in X \times \mathbb{R}$, and RKHS $\mathcal{H}$ be an $\mathbb{R}$-Hilbert function space over X with reproducing kernel $k : X \times X \to \mathbb{R}$. Let g be a strictly increasing function $g : [o, \infty] \to \mathbb{R}$, and l be an arbitrary loss function, where $l : (X \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$.*
We want to minimize the following empirical risk term:

$$E(f, (x_1, y_1), ..., (x_n, y_n)) \ := \ l((x_1, y_1, f(x_1)), ..., (x_n, y_n, f(x_n))) + g(\|f\|).$$

For $\hat{f} = argmin_{f \in \mathcal{H}} E(f, (x_1, y_1), ..., (x_n, y_n))$, $\hat{f}$ can be represented in the form:

$$\hat{f}(\cdot) = \sum_{i=1}^{n} a_i k(\cdot, x_i)$$

with $a_i \in \mathbb{R}$ for all i.

Proof. First we let Φ be the canonical feature map of k as defined in 7. Recall: function $\Phi : X \to \mathcal{H}$ where $\Phi(x)(\cdot) = k(\cdot, x)$. Due to the reproducing property where $\Phi(x)(x') = \langle \Phi(x), k(\cdot, x')\rangle$, we have:

$$\begin{aligned}\Phi(x)(x') &= k(x', x) \\ &= \langle \Phi(x), k(\cdot, x')\rangle \\ &= \langle \Phi(x), \Phi(x')\rangle.\end{aligned}$$

So Φ is a feature space of k. Using orthogonal decomposition, we decompose $f \in \mathcal{H}$ into a component projected onto the span of $\Phi(x_i), ..., \Phi(x_n)$, and the other component orthogonal to this span. We will then prove this orthogonal component is 0 for any f that reduces the empirical risk term, hence completing the prove.

$$f = \sum_{i=1}^{n} a_i \Phi(x_i) + \gamma,$$

where $\gamma \in \mathcal{H}$, $\langle \Phi(x_i), \gamma \rangle = 0$ for all i.
Next, applying the reproducing property again,

$$\begin{aligned} f(x_j) &= \langle f, k(\cdot, x_j) \rangle \\ &= \langle \sum_{i=1}^{n} a_i \Phi(x_i) + \gamma, \Phi(x_j) \rangle \\ &= \langle \sum_{i=1}^{n} a_i \Phi(x_i), \Phi(x_j) \rangle + \langle \gamma, \Phi(x_j) \rangle \\ &= \sum_{i=1}^{n} a_i \langle \Phi(x_i), \Phi(x_j) \rangle. \end{aligned}$$

Now, consider:

$$\begin{aligned} \|f\|^2 &= \left\| \sum_{i=1}^{n} a_i \Phi(x_i) + \gamma \right\|^2 \quad (\because \text{orthogonality}) \\ &= \left\| \sum_{i=1}^{n} a_i \Phi(x_i) \right\|^2 + \|\gamma\|^2 \\ &\geq \left\| \sum_{i=1}^{n} a_i \Phi(x_i) \right\|^2 \\ \implies g(\|f\|) &\geq g(\left\| \sum_{i=1}^{n} a_i \Phi(x_i) \right\|) \end{aligned}$$

Therefore, if we have $\gamma = 0$, since $f(x_i)$ is unaffected by this for all i, $l((x_1, y_1, f(x_1)), ..., (x_n, y_n, f(x_n)))$ is also unaffected by γ. For the term $g(\|f\|)$, it decreases if we have $\gamma = 0$. Hence, $\hat{f} = argmin_{f \in \mathcal{H}} E(f, (x_1, y_1), ..., (x_n, y_n))$, $\hat{f}$ must have $\gamma = 0$, and

$$\begin{aligned} \hat{f} &= \sum_{i=1}^{n} a_i \Phi(x_i) \\ &= \sum_{i=1}^{n} a_i k(\cdot, x_i) \end{aligned}$$

$\square$

3 Approximation Theorem

Definition 8. The fill distance for a set of points $X = \{x_1, ..., x_N\} \subseteq \Omega$ for a bounded domain Ω is defined to be

$$h_{X,\Omega} := \sup_{x \in \Omega} \min_{1 \le j \le N} \|x - x_j\|_2$$

.

The below theorem gives us some justification as to why the minimum norm interpolating function was chosen, though this only works under noiseless conditions:

Theorem 3. *Fix $h^* \in \mathcal{H}_\infty$. Let $(x_1, y_1), ..., (x_n, y_n)$ be i.i.d. random variables where x_i drawn randomly from a compact cube $\Omega \subseteq \mathbb{R}^d$, $y_i = h^*(x_i)\ \forall i$. There exists $A, B > 0$ such that for any interpolating $h \in \mathcal{H}_\infty$ with high probability*

$$\sup_{x \in \Omega} |h(x) - h^*(x)| < Ae^{-B(n/log\, n)^{1/d}}(\|h^*\|_{\mathcal{H}_\infty} + \|h\|_{\mathcal{H}_\infty})$$

Theorem 11.22 in [8]:

Let Ω be a cube in $\mathbb{R}^d$. Suppose ... There exists a constant $c > 0$ such that the error between a function $f \in N(\Omega)$ and its interpolant $s_{f,X}$ can be bounded by:

$$\|f - s_{f,X}\|_{L_\infty(\Omega)} \le \exp(-c/h_{X,\Omega})|f|_N(\Omega)$$

for all data sites X with sufficiently small $h_{X,\Omega}$.

With $h_{X,\Omega}$ as the fill on the order of $O(n/logn)^{-1/d}$ (using the theorem S1 in Belkin's paper which wasn't proved). We consider $f(x) := h(x) - h*(x)$. Since h is interpolating, we have $f(x_i) = 0$ for all x_i. We then let $s_{f,X}$ be the zero function, since it is an interpolant of f. Thus, we have: $s_{f,X}$ can be bounded by:

$$\|f\|_{L_\infty(\Omega)} = \sup_{x \in \Omega} |h(x) - h^*(x)| < \exp(-c(n/logn)^{1/d})|f|_N(\Omega)$$

$$\le \exp(-c(n/logn)^{1/d})(\|h^*\|_{\mathcal{H}_\infty} + \|h\|_{\mathcal{H}_\infty})$$

Another form we can have is using proposition 14.1 in [8]:

Proposition 1. *Let $\Omega \subseteq \mathbb{R}^d$ be bounded and measurable. Suppose $X = \{x_1, ..., x_N\} \subseteq \Omega$ is quasi-uniform with respect to $c_{qu} > 0$. Then there exists constants $c_1, c_2 > 0$ depending only on space dimension d, on Ω and on c_{qu} such that:*

$$c_1 N^{-1/d} \le h_{X,\Omega} \le c_2 N^{-1/d}$$

.

With the definition of quasi-uniformness being:

Definition 9. For the separation distance of $X = \{x_1, ..., x_N\}$ being defined as $q_x := \frac{1}{2} \min_{i \ne j} \|x_i - x_j\|_2$.

We can then use the above proposition with n replacing $n/\log n$.

In either case, by choosing a the smallest norm for h, we can see that it corresponds to the smallest upperbound for $|h(x) - h^*(x)|$.

4 Existing Bounds Provide No Guarantees for Interpolated Kernel Classifiers

Steps are:

- Find lower bound on function norm of t-overfitted classifiers in RKHS corresponding to Gaussian Kernels.
- Show loss for available bounds for kernel methods based on function norm (can perhaps use this to explain approximation theorem as well?)

Interpolation: 0 regression error. Overfitting: 0 classification error. Interpolation implies overfitting.

Definition 10. We say $h \in H$ t-overfits data, if it achieves zero classification loss (overfits) and $\forall_i y_i h(x_i) > t > 0$.

The below shows a theorem on how the function norm changes with respect to t-overfitting.

Theorem 4. *Let* $(\mathbf{x}_i, y_i)$ *be data sampled from* P *on* $\Omega \times \{-1, 1\}$ *for* $i = 1, ..., n$. *Assume that* y *is not a deterministic function of* x *on a subset of non-zero measure. Then, with high probability, any* h *that t-overfits the data, satisfies*

$$\|h\|_H > Ae^{Bn^{1/d}}$$

for some constants $A, B > 0$ *depending on* t.

We define the γ-shattering and fat-shattering dimension below:

Definition 11. Let F be a set of functions mapping from a domain X to $\mathbb{R}$. Suppose $S = \{x_1, x_2, ..., x_m\} \subseteq X$. Suppose also that γ is a positive real number. Then S is γ-shattered by F if there are real numbers $r_1, r_2, ..., r_m$, such that for each $b \in \{0, 1\}^m$ there is a function f_b in F with

$$f_b(x_i) \geq r_i + \gamma \text{ if } b_i = 1, \text{ and } f_b(x_i) \leq r_i - \gamma \text{ if } b_i = 0, \text{ for } 1 \leq i \leq m.$$

We say $r = (r_1, r_2, ..., r_m)$ witnesses the shattering. Suppose that F is a set of functions from a domain X to $\mathbb{R}$ and that $\gamma > 0$. Then F has γ-dimension d if d is the maximum cardinality of a subset S of X that is γ-shattered by F. If no such maximum exists, we say that F has infinite γ-dimension. The γ-dimension of F is denoted $fat_F(\gamma)$. This defines a function $fat_F : \mathbb{R} \to N \cup \{0, \infty\}$, which we call the fat-shattering dimension of F.

Proof. Let $B_R = \{f \in \mathcal{H}, \|f\|_{\mathcal{H}} < R\}$ be a ball of radius R in RKHS $\mathcal{H}$. Suppose the data is γ-overfitted, [4] gives us a high probability of a bound of

$$L(f) < O(\frac{\ln(n)^2}{\sqrt{n}} \sqrt{fat_{B_R}(\gamma/8)})$$

for $L(f)$ the expected classification error. Also, from [2] we have

$$fat_{B_R}(\gamma) < O((log(R/\gamma))^d)$$

. We then have B_R containing no function that γ overfits the data unless

$$(log(R/\gamma))^d > O(n) \implies R > c_1 \exp(c_2(\frac{n}{\ln n})^{1/d})$$

for some positive constants c_1, c_2. $\square$

Classical bounds for kernel methods ([3]) are in the form:

$$|\frac{1}{n}\sum_i l(f(x_i), y_i) - L(f)| \leq C\frac{\|f\|_{\mathcal{H}}^a}{n^b}, \quad C, a, b \geq 0$$

The right side on this will tend to infinity for bigger $\|f\|_{\mathcal{H}}$, which is suggested by Theorem 4.

5 Random Fourier Features

For a feature map $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$ the kernel trick allows easy computation for positive definite kernel k where $k(x, y) =< \phi(x), \phi(y) >$. We want to find a randomized feature map $z : \mathbb{R}^d \to \mathbb{R}^{\bar{d}}$ such that

$$k(x, y) =< \phi(x), \phi(y) >\approx< z^{\mathrm{T}}(x), z(y) >$$

. As suggested by [5], for a shift-invariant kernel k: $k(x, y) = k(x - y))$, we consider the mapping $z(x) = cos(w^{\mathrm{T}}x + b)$, where w is drawn from the probability distribution p:

$$p(w) = \frac{1}{2\pi}\int k(h) \exp(-iw^{\mathrm{T}}h) \, \mathrm{d}h \tag{3}$$

when we compute the Fourier transform of the kernel k, and b is drawn from the uniform distribution on $[0, 2\pi]$.

We know that the fourier transform of $k(\cdot)$ is a probability distribution from Bochner's theorem:

Theorem 5. *(Bochner [6]).For a continuous kernel $k(x - y)$ it is a positive definite kernel if and only if $k(\cdot)$ is the fourier transform of a non-negative measure.*

We now have:

$$k(x - y) = \int_{\mathbb{R}^d} p(w)\exp(iw^{\mathrm{T}}(x - y)) \, \mathrm{d}w = \mathbb{E}_w[e^{iw^{\mathrm{T}}x}(e^{iw^{\mathrm{T}}y})^*]$$

. Therefore, we can use $e^{iw^{\mathrm{T}}x}(e^{iw^{\mathrm{T}}y})^*$ as an estimate (unbiased) of $k(x, y)$. Let $\phi_w(x) = e^{iw^{\mathrm{T}}x}$ We can also use $z_w(x) = \sqrt{2}cos(w^{\mathrm{T}}x + b)$ instead of $\phi_w(x)$, as suggested by [5].

Proposition 2. *For $z_w(x) = \sqrt{2}cos(w^Tx + b)$, where w is drawn from probability distribution p in* (3) *and b drawn from a uniform random variable on $[0, 2\pi]$.*

$$E(z_w(x))z_w(y) = k(x, y)$$

Proof.

$$
\begin{aligned}
z_w(x) &= 2\ \frac{\sqrt{2}}{2} cos(w^{\mathrm{T}}x + b) \\
&= \frac{1}{\sqrt{2}}\ (e^{i(w^{\mathrm{T}}x+b)} + e^{-i(w^{\mathrm{T}}x+b)}) \\
&= \frac{1}{\sqrt{2}}\ (\phi_w(x)e^{ib} + \phi_w(x)^* e^{-ib})
\end{aligned}
$$

Where $\phi_w(x) = e^{iw^{\mathrm{T}}x}$.

$$
\begin{aligned}
z_w(x)z_y(y) &= \frac{1}{2}[\phi_w(x)\phi_w(y)e^{i2b} + \phi_w(x)^*\phi_w(y)^*e^{-i2b} + \phi_w(x)\phi_w(y)^* + \phi_w(x)^*\phi_w(y)] \\
\mathbb{E}[z_w(x)z_y(y)] &= \frac{1}{2}\mathbb{E}[\phi_w(x)\phi_w(y)e^{i2b} + \phi_w(x)^*\phi_w(y)^*e^{-i2b}] + \frac{1}{2}\mathbb{E}[\phi_w(x)\phi_w(y)^*] + \frac{1}{2}\mathbb{E}[\phi_w(x)^*\phi_w(y)]
\end{aligned}
$$

As mentioned earlier in Theorem 5, $\mathbb{E}_w[\phi_w(x)\phi_w(y)^*] = k(x-y)$. Also $\phi_w(x)\phi_w(y)^* = (\phi_w(x)^*\phi_w(y))^*$.

$$
\begin{aligned}
\mathbb{E}[z_w(x)z_y(y)] &= \frac{1}{2}\mathbb{E}[\phi_w(x)\phi_w(y)e^{i2b} + \phi_w(x)^*\phi_w(y)^*e^{-i2b}] + \frac{1}{2}k(x-y) + \frac{1}{2}[k(x-y)]^* \\
&= \frac{1}{2}\mathbb{E}[\phi_w(x)\phi_w(y)e^{i2b} + \phi_w(x)^*\phi_w(y)^*e^{-i2b}] + k(x-y)
\end{aligned}
$$

For real kernel , $k(x-y) = (k(x-y))^*$.

$$
\begin{aligned}
\mathbb{E}_{w,b}[\phi_w(x)\phi_w(y)e^{i2b}] &= \frac{1}{2\pi}\int_{\mathbb{R}^d}\int_0^{2\pi} p(w)\phi_w(x)\phi_w(y)e^{i2b}\mathrm{d}b\ \mathrm{d}w \\
&= \frac{1}{2\pi}\int_{\mathbb{R}^d} p(w)\phi_w(x)\phi_w(y)\int_0^{2\pi} e^{i2b}\mathrm{d}b\ \mathrm{d}w \\
&= 0
\end{aligned}
$$

Since $\int_0^{2\pi} e^{i2b}\mathrm{d}b = 0$. Similarly, $\mathbb{E}_{w,b}[\phi_w(x)^*\phi_w(y)^*e^{-i2b}] = 0$.

$$\therefore \mathbb{E}[z_w(x)z_y(y)] = k(x-y).$$

□

As suggested by [5], the variance of the estimate is decreased by using z, a D dimensional vector by concatenating D of z_w and normalizing by a constant $\sqrt{D}$. We let:

$$z(x) = \sqrt{\frac{2}{D}}[cos(w_1^{\mathrm{T}}x + b_1)...cos(w_D^{\mathrm{T}}x + b_D)]$$

with randomly drawn w_i and b_i as described previously.

Theorem 6. *For N the number of random features, and $x_1, x_2, ..., x_n$ the data points, when $N > n$ and as N increases, the norm of the minimizer tends to the norm of the minimum norm RKHS interpolant.*

Proof. Let $f(x)$ be the minimum norm RKHS interpolant function for the datapoints.

$$f(x) = \sum_i \alpha_i k(x_i, x) \approx \sum_i \alpha_i z(x_i)^{\mathrm{T}} z(x) = \beta^{\mathrm{T}} z(x) = \hat{f}(x)$$

(the first equality holds due to Representer Theorem) Where $\beta = \sum_i \alpha_i z(x_i)$. The norm of the function from the random fourier features approximation is:

$$\|\beta\| = \beta^{\mathrm{T}}\bar{\beta} = (\sum_i \alpha_i z^{\mathrm{T}}(x_i))(\sum_i \bar{\alpha}_i \bar{z}(x_i)) = \sum_i \sum_j \alpha_i \bar{\alpha}_j z^{\mathrm{T}}(x_i)\bar{z}(x_j) \approx \sum_i \sum_j \alpha_i \bar{\alpha}_j k(x_i, x_j) = \|f\|$$

$\square$

References

[1] *Support Vector Machines*. Springer New York, 2008.

[2] Mikhail Belkin. Approximation beats concentration? an approximation view on inference with smooth radial kernels. *arxiv:1801.03437*, 2018.

[3] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. *arxiv:1802.01396*, 2018.

[4] Balázs Kégl, Tamás Linder, and Gábor Lugosi. Data-dependent margin-based generalization bounds for classification. In *Lecture Notes in Computer Science*, pages 368–384. Springer Berlin Heidelberg, 2001.

[5] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, pages 1177–1184. Curran Associates, Inc., 2008.

[6] Walter Rudin. *Fourier Analysis on Groups*. John Wiley & Sons, Inc., jan 1990.

[7] Bernhard SchÃ¶lkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *Lecture Notes in Computer Science*, pages 416–426. Springer Berlin Heidelberg, 2001.

[8] Holger Wendland. *Scattered Data Approximation*. Cambridge Monographs on APplied and Computational Mathematics. Cambridge University Press, 2004.

Appendix