MA4199 Project – Bias Variance Tradeoff

Ng Wei Le

1 Approximation Theorem

Definition 1. The fill distance for a set of points $X = \{x_1, ..., x_N\} \subseteq \Omega$ for a bounded domain Ω is defined to be

$$h_{X,\Omega} \coloneqq \sup_{x \in \Omega} \min_{1 \leq j \leq N} \|x - x_j\|_2$$

.

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 1. *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 [4]:

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:

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

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:

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

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

Another form we can have is using proposition 14.1 in [4]:

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:*

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

.

With the definition of quasi-uniformness being:

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

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 $|h(x) - h^*(x)|$.

2 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 3. 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 2. *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 4. 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, [3] 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 [1] 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. Classical bounds for kernel methods ([2]) 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 2.

References

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

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

[3] 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.

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

Appendix