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 [2]:

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 [2]:

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.

References

[1] Shai Shalev-Shwartz and Shai Ben David. Understanding machine learning: From theory to algorithms. `http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning`.

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

Appendix