Statistical Foundations of Learning

Debarghya Ghoshdastidar

School of Computation, Information and Technology

Technical University of Munich

Regression

Outlook

- Problem: $\mathcal{D}$ distribution on $\mathcal{X} \times \mathbb{R}$ . . . will mostly assume $\mathcal{X} \subset \mathbb{R}^p$
  Given training sample $S = \{(x_i, y_i)\}_{i=1}^m \sim \mathcal{D}^m$, find a predictor $h : \mathcal{X} \to \mathbb{R}$

- Training/learning by (regularised) squared regression:

$$\underset{h \in \mathcal{H}}{\text{minimise}} \ \frac{1}{m} \sum_{i=1}^{m} (h(x_i) - y_i)^2 + \lambda \cdot \text{complexity}(h)$$

- Two perspectives for guarantees:
  - Approximation: Assume $y = f(x)$. Which functions f can be learned by our model?

$$\sup_{x \in \mathcal{X}} |f(x) - h(x)| \leq ?$$

  - Generalisation: How well does learned h predict on new data?

$$\mathbb{E}_{(x,y) \sim \mathcal{D}} \left[(y - h(x))^2 \right] \leq ?$$

Outline

- Neural network regression: Universal approximation theorem
- Kernel regression: Universal kernels, Stability / Generalisation

How many neurons needed to learn a Lipschitz function?

- Let $\mathcal{X} = [0,1)$ and $f : \mathcal{X} \to \mathbb{R}$ be a ρ-Lipschitz function
$$|f(x) - f(x')| \leq \rho \cdot |x - x'| \qquad \text{for all} x, x' \in \mathcal{X}$$

- Construct $\widetilde{h}(x)$ with values
  - Let $t_i = \frac{i-1}{N}$, $i = 1, \ldots, N$. Define $h(x) = f(t_i)$ for $x \in [t_i, t_{i+1})$
  - How well does $\widetilde{h}$ approximate f?
$$\sup_{x\in[0,1)} |f(x) - h(x)| \leq \max_i \sup_{x\in[t_i,t_{i+1})} |f(x) - f(t_i)| \leq \frac{\rho}{N}$$

- Suppose we use step activation $\mathbf{1}\{z \geq 0\}$. So $\widetilde{h}(x) = \sum_{i=1}^{M} a_i \cdot \mathbf{1}\{x + b_i \geq 0\}$
  - How many M needed to model $\widetilde{h}(x)$? How many needed to ensure $\sup_x |f(x) - \widetilde{h}(x)| \leq \epsilon$?

How many ReLU units needed to learn a Lipschitz function?

- With step activation, a 2-layer NN

$$\widetilde{h}(x) = f(0) \cdot \mathbf{1}\{x \geq 0\} + \sum_{i=2}^{N} \big(f(t_i) - f(t_{i-1})\big) \cdot \mathbf{1}\{x - t_i \geq 0\} \qquad \text{with } N \geq \frac{\rho}{\epsilon}$$

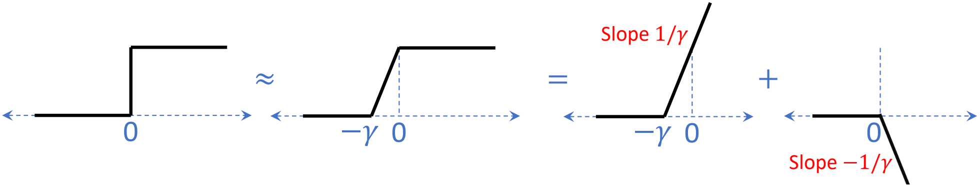
guarantees $\sup_x |f(x) - \widetilde{h}(x)| \leq \epsilon$

- Problem: What is N, if the activations are $\text{ReLU}(z) = \max\{z, 0\}$?

$$h(x) = \sum_{i=1}^{M} a_i \cdot \text{ReLU}(w_i x + b_i)$$

- How can we approximately construct $\mathbf{1}\{z \geq 0\}$ using $\text{ReLU}(\cdot)$?

How many ReLU units needed to learn a Lipschitz function?



- $\mathbf{1}\{z \geq 0\} \approx \text{ ramp}_\gamma(z) = \begin{cases} 0 & z < -\gamma \\ 1 & z \geq 0 \\ \frac{z+\gamma}{\gamma} & z \in [-\gamma, 0) \end{cases} \quad ; \qquad \gamma \in (0,1)$
  - Observe $\sup_z \left| \mathbf{1}\{z \geq 0\} - \text{ ramp}_\gamma(z) \right| = 1$
- $\text{ramp}_\gamma(z) = \frac{1}{\gamma}\text{ReLU}(z+\gamma) - \frac{1}{\gamma}\text{ReLU}(z)$

Approximating Lipschitz functions by ReLU network

Theorem Reg.1 (Approximating Lipschitz functions by ReLU network)

Let $f : [0,1) \to \mathbb{R}$ be a ρ-Lipschitz continuous function. There is a 1-hidden layer neural network with $\left\lceil \frac{4\rho}{\epsilon} \right\rceil$ ReLU units whose output $h(x)$ satisfies $\sup\limits_{x\in[0,1)} |f(x) - h(x)| \leq \epsilon$

Extensions of construction/proof idea:

- $f : [0,1)^p \to \mathbb{R}$ is ρ-Lipschitz: $|f(x) - f(x')| \leq \rho \cdot \|x - x'\|_2 \leq \rho\sqrt{p} \cdot \max_i \left|x^{(i)} - x'^{(i)}\right|$
  - We can ϵ-approximate f by a ReLU net with $\sim \frac{\rho\sqrt{p}}{\epsilon^p}$ ReLU units
- Uniformly continuous $g : [0,1)^p \to \mathbb{R}$
  - For any $\epsilon > 0$, there is $\delta_\epsilon > 0$, such that $\|x - x'\|_2 \leq \delta \implies |f(x) - f(x')| \leq \epsilon$
  - Discretise into hypercubes of length $\sim \delta_\epsilon$ instead of $\sim \frac{\epsilon^p}{\rho}$

Proof: The ReLU network

- Let $N \geq \frac{2\rho}{\epsilon}$ and $t_i = \frac{i-1}{N}$, $i = 1, \ldots, N$

$$\widetilde{h}(x) = f(0) \cdot \mathbf{1}\{x \geq 0\} + \sum_{i=2}^{N} \big(f(t_i) - f(t_{i-1})\big) \cdot \mathbf{1}\{x - t_i \geq 0\}$$

guarantees $\sup\limits_{x \in [0,1)} |f(x) - \widetilde{h}(x)| \leq \epsilon/2$

- Choose $\gamma \leq \frac{1}{N}$, and define

$$\begin{aligned} h(x) &= f(0) \cdot \mathrm{ramp}_\gamma(x) + \sum_{i=2}^{N} \big(f(t_i) - f(t_{i-1})\big) \cdot \cdot \mathrm{ramp}_\gamma(x - t_i) \\ &= \frac{f(0)}{\gamma} \cdot \big(\mathrm{ReLU}\,(x + \gamma) - \mathrm{ReLU}\,(x)\,\big) \\ &\quad + \sum_{i=2}^{N} \frac{\big(f(t_i) - f(t_{i-1})\big)}{\gamma} \cdot \big(\mathrm{ReLU}\,(x - t_i + \gamma) - \mathrm{ReLU}\,(x - t_i)\,\big) \end{aligned}$$

... $2N$ ReLU units

Proof: Bounding $\sup_x |\widetilde{h}(x) - h(x)|$

- Recall $\mathbf{1}\{z \geq 0\}$ and $\text{ramp}_\gamma(z)$ differs only on $x \in (-\gamma, 0)$

$$\widetilde{h}(x) - h(x) = f(0) \cdot \left(1 - \frac{x+\gamma}{\gamma}\right) \cdot \underbrace{\mathbf{1}\{x \in (-\gamma, 0)\}}_{x \notin [0,1)}$$
$$+ \sum_{i=2}^{N} \underbrace{(f(t_i) - f(t_{i-1}))}_{\leq \rho \cdot |t_i - t_{i-1}| \leq \rho/N} \cdot \left(1 - \frac{x - t_i + \gamma}{\gamma}\right) \cdot \mathbf{1}\{x \in (t_i - \gamma, t_i)\}$$

- For $\gamma \leq \frac{1}{N}$, intervals are disjoint. Hence,

$$\sup_{x \in [0,1)} |\widetilde{h}(x) - h(x)| \leq \frac{\rho}{N} \leq \frac{\epsilon}{2}$$

Universal approximation with 1-hidden layer nets

- Earliest results by Cybenko (1989); Hornik et al. (1989)
  - Various versions exist now for wide or deep nets. See Wikipedia
  - We will see version by Allan Pinkus (Acta Numerica, 1999)
- Setup:
  - Let $C(\mathbb{R})$ = space of all continuous functions $f : \mathbb{R} \to \mathbb{R}$
  - $\sigma \in C(\mathbb{R})$ is a continuous activation function
  - Space of functions obtained from 1 hidden layer NN

$$\begin{aligned}\mathcal{H}_\sigma &= \left\{ \sum_{i=1}^{N} a_i \cdot \sigma(w_i x + b_i) \ : \ N = 1, 2, \ldots, w_i, b_i, a_i \in \mathbb{R} \right\} \\ &= \operatorname{span}\{\sigma(wx + b) \ : \ w, b \in \mathbb{R}\}\end{aligned}$$

Universal approximation with 1-hidden layer nets

Theorem Reg.2 (Universal approximation theorem (Pinkus, 1999))

Let $\sigma \in C(\mathbb{R})$. If σ is ***not a polynomial****, then $\mathcal{H}_\sigma$ is* ***dense*** *in $C(\mathbb{R})$ in the following sense:*

- *for every compact set $\mathcal{X} \subset \mathbb{R}$,*
- *for $f \in C(\mathcal{X})$ and $\epsilon > 0$,*

there is a function $h \in \mathcal{H}_\sigma$ such that $\sup_{x \in \mathcal{X}} |f(x) - h(x)| \leq \epsilon$.

- Proof skipped. Idea is to approximate any $f \in C(\mathbb{R})$ by an arbitrarily wide NN
- If σ is a polynomial, then $\mathcal{H}_\sigma$ is **not dense** in $C(\mathbb{R})$. Why?
  - If σ is a polynomial of degree d, then $h \in \mathcal{H}_\sigma$ cannot approximate weell a polynomial of degree $> d$

Can we approximate any function by bounded width NN?

- Let $\mathcal{F}$ = some class of function $f : [0,1]^p \to [0,1]^q$
  - Example: Continuous OR Convex OR ρ-Lipschitz OR L_p (where $\int |f(x)|^p dx < \infty$)
- Consider the deep ReLU NN of the form $h : \mathbb{R}^p \to \mathbb{R}^q$
  $h(x) = A_k \cdot \mathrm{ReLU}\left(A_{k-1} \cdot \mathrm{ReLU}\left(\ldots \mathrm{ReLU}\left(A_2 \cdot \mathrm{ReLU}\left(A_1 x + b_1\right) + b_2\right) \ldots\right) + b_{k-1}\right) + b_k$
  - Alternates between affine transforms, $Ax + b$, and coordinate-wise ReLU
  - $A_i \in \mathbb{R}^{p_i \times p_{i-1}}, b \in \mathbb{R}^{p_i}$, $p_0 = p$ and $p_k = q$
  - Depth of network $= k$, and width of network $w = \max\{p_0, p_1, \ldots, p_k\}$

Can we approximate any function by bounded width NN?

Theorem Reg.3 (Minimum width of ReLU NN for universal approximation)

Let $w_{\min}(p,q;\mathcal{F})$ = *minimum* w *such that ReLU NNs of width* $\leq w$ *(and arbitrary depth) can approximate any function* $f \in \mathcal{F}$

- *Hanin, Sellke (arXiv:1710.11278):*
  $\mathcal{F} = \{\textit{continuous functions}\} \implies p+1 \leq w_{\min}(p,q;\mathcal{F}) \leq p+q$
- *Park et al. (ICLR 2021):*
  $\mathcal{F} = \{L_p \textit{ functions}\} \implies w_{\min}(p,q;\mathcal{F}) = \max\{p+1,q\}$
- *Next slides:*
  $\mathcal{F} = \{\rho\textit{-Lipschitz functions}\} \implies w_{\min}(1,1;\mathcal{F}) \leq 2$

Will prove only last statement. Use steps provided in next slides (exercises marked in red)

Proof: Width 2 ReLU NN for Lipschitz functions (not in exam)

- The following is a possible construction based on Hanin, Sellke (arXiv:1710.11278).

- Let $f : [0,1) \to \mathbb{R}$ be ρ-Lipschitz

  - Discretise $[0,1)$ by points $t_i = \frac{i-1}{N}$, for $i = 1, \ldots, N$

  - Max-min string: We call a function $g : [0,1) \to \mathbb{R}$ of length k if there are k affine functions $r_1, \ldots, r_k$, $(r(x) = ax + b)$ such that

$$h(x) = \sigma_k\Big\{r_k(x), \sigma_{k-1}\Big\{r_{k-1}(x), \sigma_{k-2}\Big\{\ldots, \sigma_2\{r_3(x), \sigma_1\{r_2(x), r_1(x)\}\ldots\}\Big\}\Big\}$$

where σ_i is either $\max$ or $\min$

    - We will construct a max-min string $g(x)$ of length $2N$ that matches $f(x)$ on $\{t_1, \ldots, t_N\}$

  - Above max-min string $h(x)$ of length $2N$ can be modelled by a ReLU NN of width 2 and depth $2N$

  - Bound $|h(x) - f(x)|$ for $x \notin \{t_1, \ldots, t_N\}$ using ρ-Lipschitz (not sure how bad is bound)

Proof: Max-min string on $S = \{t_1, \ldots, t_N\}$

- Choose $b > \max\{|f(t_i)| \,:\, i = 1, \ldots, N\}$

- We construct h recursively.

    - Define $g_1(x) = f(t_1)$ (constant function)

    - For each $j = 1, 2, \ldots,$, let $\ell_j(x) = N \cdot b \cdot (t_{j+1} - x)$

    - Define $g_{j+1}(x) = \max\left\{f(t_{j+1}) - \ell_j(x), \min\{g_j(x), f(t_{j+1}) + \ell_j(x)\}\right\}$

(1.1) Show that $\ell_j(x) = 0$ for $x = t_j$ and $\ell_j(x) \geq b$ for $x = t_1, \ldots, t_{j-1}$.
Hence, by induction, show that $g_j(x) = f(x)$ for $x \in \{t_1, \ldots, t_j\}$.

- $h(x) = g_N(x)$ is a max-min string of length $2N$ that matches $f(x)$ on $\{t_1, \ldots, t_N\}$

(1.2) Derive a bound on $\sup\limits_{x \in [0,1)} |f(x) - h(x)|$ using ρ-Lipschitzness. Hence, choose N

Proof: Modelling $\max, \min$ by ReLU NN

- Let α, β be two scalar such that $|\beta| < b$

(1.3) Show that $\max\{\alpha, \beta\}$ can be modelled by a 1-hidden layer NN with 2 ReLU units as

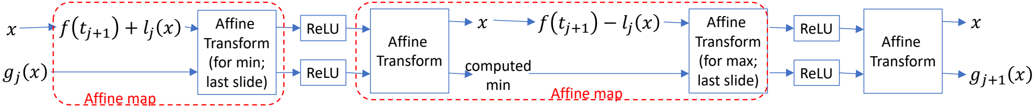
$$\max\{\alpha,\beta\} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}^\top \cdot \mathrm{ReLU}\left(\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} 0 \\ b \end{pmatrix} \right) - b$$

- Above construction also works when α, β are functions of x (but then we need to also propagate x through the NN)

(1.4) What is the corresponding ReLU NN for computing $\min\{\alpha, \beta\}$?

Proof: Modelling $h(x)$ by ReLU NN

- Idea: Model the map $\begin{pmatrix} x \\ g_j(x) \end{pmatrix} \mapsto \begin{pmatrix} x \\ g_{j+1}(x) \end{pmatrix}$ with a 2-hidden layer ReLU NN



- One can combine consecutive affine maps into a single affine map, $\mathbb{R}^2 \to \mathbb{R}^2$, resulting in a NN with 2 ReLU layers

(1.5) Compute the resulting affine maps

Outline

- Neural network regression: Universal approximation theorem
- Kernel regression: Universal kernels, Stability / Generalisation

Positive Semidefinite Kernels

- Kernel: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is any symmetric function
  - Informally, $k(x, x')$ measures similarity between $x, x' \in \mathcal{X}$
  - Examples: Gaussian kernel $k(x, x') = e^{-\|x-x'\|^2/\gamma}$, Quadratic kernel $k(x, x') = (\langle x, x' \rangle)^2$

Theorem Reg.4 (Positive semidefinite definite (psd) kernel)

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel. Then the following statements are equivalent:

1. *For all $n = 1, 2, \ldots$ and all $x_1, \ldots, x_n \in \mathcal{X}$, the $n \times n$ matrix K with entries $K_{ij} = k(x_i, x_j)$ is positive semidefinite ($u^\top K u \geq 0$ for all $u \in \mathbb{R}^n$)*
2. *There exists an inner product space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ for all $x, x' \in \mathcal{X}$*

A kernel k satisfying above (equivalent) conditions is a psd kernel

$\mathcal{H}$ is called the reproducing kernel Hilbert space (rkhs) for k

Reproducing kernel Hilbert space (summary)

- What is real inner product space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$?
  - $\mathcal{H}$ is a set of elements
  - $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is a valid inner (dot) product defined on $\mathcal{H}$
- When is $\mathcal{H}$ a Hilbert space?
  - From $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, we can define a norm $\|\phi\|_{\mathcal{H}} = \sqrt{\langle \phi, \phi \rangle_{\mathcal{H}}}$ and a metric $d(\phi, \phi') = \|\phi - \phi'\|_{\mathcal{H}}$
  - $\mathcal{H}$ is a Hilbert space if "it contains limiting points"
    Any sequence $\{\phi_n\}_{n=1}^{\infty} \in \mathcal{H}$ such that $d(\phi_m, \phi_n)$ becomes arbitrarily small as $m, n \to \infty$ (Cauchy sequence) has a limit $\phi_n \to \phi \in \mathcal{H}$
- How do we construct rkhs for kernel k?
  - Many possible Hilbert spaces and feature maps for k, but they are isomorphic

Reproducing kernel Hilbert space (summary)

- Assume $\displaystyle\int\int k^2(x,x')\,\mathrm{d}x\,\mathrm{d}x' < \infty$ (k has finite trace)

- Constructing ϕ and $\mathcal{H}$
  - Given kernel k, for every $x \in \mathcal{X}$, define the map $\phi_x : \mathcal{X} \to \mathbb{R}$, $\phi_x(\cdot) = k(x,\cdot)$
  - Define set $\mathcal{H}_1 = \mathrm{span}\{\phi_x \mid x \in \mathcal{X}\} = \left\{ \displaystyle\sum_{i=1}^{m} c_i \phi_{x_i}(\cdot) \mid m \in \mathbb{N}, c_i \in \mathbb{R}, x_i \in \mathcal{X} \right\}$
  - $\mathcal{H}_1$ may not contain limits of sequences, so add them. $\mathcal{H} =$ closure of $\mathcal{H}_1$

- Constructing inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, and hence, rkhs $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$
  - For every $\phi_x, \phi_{x'}$, define $\langle \phi_x, \phi_{x'} \rangle_{\mathcal{H}} = k(x,x')$ . . . why? end of next slide

Reproducing kernel Hilbert space (summary)

- Any $f, g \in \mathcal{H}_1$ is of the form $f = \sum\limits_{i=1}^{m} c_i \phi_{x_i}, g = \sum\limits_{j=1}^{m'} c'_j \phi_{x'_j}$

$$\langle f, g \rangle_{\mathcal{H}} = \left\langle \sum_{i=1}^{m} c_i \phi_{x_i}, \sum_{j=1}^{m'} c'_j \phi_{x'_j} \right\rangle_{\mathcal{H}} = \sum_{i,j} c_i c'_j \langle \phi_{x_i}, \phi_{x'_j} \rangle_{\mathcal{H}} = \sum_{i,j} c_i c'_j k(x_i, x'_j)$$

  - Any $f \in \mathcal{H} \backslash \mathcal{H}_1$ would be of form $\sum_{i=1}^{\infty} c_i \phi_{x_i}$ with $\sum\limits_{i=1}^{\infty} c_i^2 < \infty$. Define $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ as above

- Why do we define $\langle \phi_x, \phi_{x'} \rangle_{\mathcal{H}} = k(x, x')$?
  - Define an evaluation functional $\delta_x : \mathcal{H} \to \mathbb{R}$ such that $\delta_x(f) = f(x)$
  - Riesz representation theorem: There is unique $\phi_x \in \mathcal{H}$ such that $\delta_x(f) = \langle f, \phi_x \rangle_{\mathcal{H}}$

In present case, $k(x, x') = \phi_x(x') = \delta_{x'}(\phi_x) = \langle \phi_x, \phi_{x'} \rangle$

Universal kernel

- Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is **universal** if
  - for every $\epsilon > 0$, every continuous function $f : \mathcal{X} \to \mathbb{R}$ and all compact subsets $C \subset \mathcal{X}$,
  - there exists $h \in \mathrm{span}\{\phi_x \; : \; x \in \mathcal{X}\}$ such that $\sup\limits_{x \in C} |f(x) - h(x)| \leq \epsilon$
- Taylor criterion for universality (proof skipped):
  - Let $\mathcal{X} = \{x \in \mathbb{R}^p \mid \|x\|_2 \leq r\}$ and kernel $k(x, x') = g(\langle x, x' \rangle)$
  - If g can be expressed as a power series $g(z) = \sum_{i=0}^{\infty} a_i z^i$ that converges for all $|z| < r^2$

    then k is universal
- Example: Exponential $k(x, x') = e^{\gamma \langle x, x' \rangle}$, $\gamma > 0$ is universal
  - Here, $g(z) = e^{\gamma z} = \sum_{i=0}^{\infty} \frac{\gamma^i}{i!} z^i$ is convergent for all radius r

Representer theorem: Do we need to know $\mathcal{H}, \phi$ for regression?

Theorem Reg.5 (Representer theorem)

- *Let* $\mathcal{H}$ *be rkhs for a psd kernel* k
  Given $S = \{(x_i, y_i)\}_{i=1}^m \subset \mathcal{X} \times \mathbb{R}$, *consider regularised loss minimisation (RLM)*
  $$\underset{h\in\mathcal{H}}{\text{minimise}}\ L_S(h) + r\left(\|h\|_{\mathcal{H}}^2\right)$$
  - $L_S : \mathcal{H} \to \mathbb{R}$ *arbitrary loss function, computed on* S;
    $r : \mathbb{R} \to \mathbb{R}$ *non-decreasing regularisation function*
- *Then optimal solution can be expressed as* $\widehat{h}(\cdot) = \sum_{i=1}^{m} \alpha_i k(x_i, \cdot)$ *for some* $\alpha_1, \ldots, \alpha_m$

Proof: Let $\mathcal{G} = \text{span}\{\phi_{x_1}, \ldots, \phi_{x_m}\}$ and $\mathcal{G}^\perp$ its complement.
Can write any $h = h_s + h_\perp$, where $h_s \in \mathcal{G}, h_\perp \in \mathcal{G}^\perp$.
$L_S(h) = L_S(h_s)$ but $r\left(\|h\|_{\mathcal{H}}^2\right) \geq r\left(\|h_s\|_{\mathcal{H}}^2\right)$. So for any $h \in \mathcal{H}$, h_s has smaller objective

Kernel Ridge Regression

- Given $S = \{(x_i, y_i)\}_{i=1}^m \subset \mathcal{X} \times \mathbb{R}$

$$\underset{h \in \mathcal{H}}{\text{minimise}} \; \frac{1}{m} \sum_{j=1}^{m} (h(x_j) - y_j)^2 + \lambda \|h\|_{\mathcal{H}}^2$$

- **Exercise:** Use representer theorem—optimal $\widehat{h}(\cdot) = \sum_{i=1}^{m} \alpha_i k(x_i, \cdot)$—to show that above problem is equivalent to

$$\underset{\boldsymbol{\alpha} \in \mathbb{R}^m}{\text{minimise}} \; \frac{1}{m} \|K\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2 + \lambda \cdot \boldsymbol{\alpha}^\top K \boldsymbol{\alpha}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m), \boldsymbol{y} = (y_1, \ldots, y_m), K = [k(x_i, x_j)]_{i,j=1,\ldots,m}$

  - Assuming K is full rank, $\boldsymbol{\alpha} = (K + \lambda m I)^{-1} \boldsymbol{y}$
  - Above is a Tikhonov RLM. Can we derive stability guarantees?

Recap: Stability of Tikhonov RLM solution (rephrased)

- Recall Riesz representation, $h(x) = \langle h, \phi_x \rangle_{\mathcal{H}}$. Hence RLM is

$$\underset{h \in \mathcal{H}}{\text{minimise}} \ \frac{1}{m} \sum_{j=1}^{m} \underbrace{(\langle h, \phi_{x_j} \rangle - y_j)^2}_{\ell_{x_j, y_j}(h)} + \lambda \|h\|_{\mathcal{H}}^2$$

Theorem Reg.6 (Tikhonov RLM is a stable learner)

- *If* $\ell =$ *convex,* ρ*-Lipschitz loss with respect to* $h \in \mathcal{H}$

  then Tikhonov RLM based on loss ℓ *is on-average-replace-one stable with rate* $\dfrac{2\rho^2}{\lambda m}$

- *Expected generalisation error of* $\widehat{h}$ *satisfies*

$$\mathbb{E}_{S \sim \mathcal{D}^m}\left[L_{\mathcal{D}}(\widehat{h})\right] \leq \mathbb{E}_{S \sim \mathcal{D}^m}\left[L_S(\widehat{h})\right] + \frac{2\rho^2}{\lambda m}$$

Is squared loss Lipschitz? What is ρ?

- Observe for $\ell_{x,y}(h) = (\langle h, \phi_x \rangle - y)^2$

$$\begin{aligned}
|\ell_{x,y}(h) - \ell_{x,y}(h')| &= \left| \langle h - h', \phi_x \rangle_{\mathcal{H}} \left(\langle h, \phi_x \rangle_{\mathcal{H}} + \langle h', \phi_x \rangle_{\mathcal{H}} - 2y \right) \right| \\
&\leq \left| \langle h - h', \phi_x \rangle_{\mathcal{H}} \right| \cdot \left| \langle h, \phi_x \rangle_{\mathcal{H}} + \langle h', \phi_x \rangle_{\mathcal{H}} - 2y \right| \\
&\leq \|h - h'\|_{\mathcal{H}} \cdot \underbrace{\|\phi_x\|_{\mathcal{H}}}_{=\sqrt{k(x,x)}} \cdot \Big| \|h\|_{\mathcal{H}} \cdot \|\phi_x\|_{\mathcal{H}} + \|h'\|_{\mathcal{H}} \cdot \|\phi_x\|_{\mathcal{H}} - 2y \Big|
\end{aligned}$$

(we use Cauchy-Schwarz)

- Assume $y \in [-c, c]$ and $k(x,x) \leq r$ for all x
  Then $\ell_{x,y}(h) = (\langle h, \phi_x \rangle - y)^2$ is $2r(rB + c)$-Lipschitz over $\{h \in \mathcal{H} \ : \ \|h\|_{\mathcal{H}} \leq B\}$

- **Exercise:** Let $L_{\mathcal{D}}^{sq}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[(h(x) - y)^2 \right]$
  Show that $L_{\mathcal{D}}^{sq}(\widehat{h}) \leq \min\limits_{\|h\|_{\mathcal{H}} \leq B} L_{\mathcal{D}}^{sq}(h) + \sqrt{\dfrac{8\rho^2 B^2}{m}}$ where $\rho = 2r(rB + c)$

Recap: Rademacher complexity

- Rademacher complexity (can be defined for any loss):

  - Consider finite set $Z = \{z_1, \ldots, z_m\}$, and $\mathcal{F}$ be class of real-valued functions defined on Z

  - Rademacher complexity of $\mathcal{F}$ with respect to set Z

$$R(\mathcal{F} \circ Z) = \mathbb{E}_{\sigma_1,\ldots,\sigma_m \sim_{iid} \text{Unif}\{\pm 1\}} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \, f(z_i) \right]$$

- Generalisation error bound using Rademacher complexity:

  - Loss satisfies $|\ell(h(x), y)| \leq M$ for all $h \in \mathcal{H}, (x, y) \in \mathcal{X} \times \mathcal{Y}$. Let $\mathcal{F} = \{\ell(h(x), y) \ : \ h \in \mathcal{H}\}$

  - For any $\delta \in (0, 1)$, with probability $1 - \delta$ over training samples $S \sim \mathcal{D}^m$,

$$\sup_{h \in \mathcal{H}} \big(L_{\mathcal{D}}(h) - L_S(h) \big) \ \leq \ 2R(\mathcal{F} \circ S) + 4M \sqrt{\frac{2 \ln(\frac{4}{\delta})}{m}}$$

Rademacher complexity for kernel models

Theorem Reg.7 (Rademacher complexity for kernel models)

Let $X = \{x_1, \ldots, x_m\}$ *and* $K = [k(x_i, x_j)]_{i,j=1,\ldots,m}$ *be the kernel matrix defined on* X.

Let $\mathcal{H}$ *is the rkhs for kernel* k, *and* $\mathcal{H}_B = \{h \in \mathcal{H} \ : \ \|h\|_{\mathcal{H}} \leq B\}$, *then the Rademacher complexity is given by*

$$R(\mathcal{H}_B \circ X) = \mathbb{E}_{\sigma_1,\ldots,\sigma_m \sim_{iid} Unif\{\pm 1\}} \left[\sup_{h \in \mathcal{H}_B} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left\langle h, \phi_{x_i} \right\rangle_{\mathcal{H}} \right]$$

and is bounded as $R(\mathcal{H}_B \circ X) \leq \dfrac{B\sqrt{\mathrm{trace}(K)}}{m} \leq \dfrac{B\sqrt{r}}{\sqrt{m}}$ *where* $k(x,x) \leq r$ *for all* x

Proof: Exercise

Rademacher complexity based bounds for kernel regression

- For generalisation bounds, we need Rademacher complexity of loss class $\mathcal{F} \circ S$, where

$$S = \{(x_i, y_i)\}_{i=1,\ldots,m} \qquad \text{and} \qquad \mathcal{F} = \{f_h(x,y) = \ell(h(x), y) \;:\; h \in \mathcal{H}\}$$

Theorem Reg.8 (Talagrand's lemma)

Consider the sets $X = \{x_1, \ldots, x_m\}$, $S = \{(x_i, y_i)\}_{i=1,\ldots,m}$ *and a function class* $\mathcal{H}$.

If the loss $\ell = \ell_{x,y}(h)$ *is* ρ*-Lipschitz with respect to* $h \in \mathcal{H}$, *then the Rademacher complexity of the loss class* $\mathcal{F} = \{f_h(x,y) = \ell(h(x), y) \;:\; h \in \mathcal{H}\}$ *is bounded as*

$$R(\mathcal{F} \circ S) \leq \rho \cdot R(\mathcal{H} \circ X)$$

- If $y \in [-c, c]$, then loss is bounded by $M = (rB + c)$ and $\rho = 2rM$-Lipschitz

- For any $\delta \in (0,1)$, with probability $1 - \delta$ over training samples $S \sim \mathcal{D}^m$,

$$\sup_{h \in \mathcal{H}_B} \left(L_{\mathcal{D}}^{sq}(h) - L_S^{sq}(h) \right) \leq \frac{2\rho B \sqrt{\text{trace}(K)}}{m} + 4M \sqrt{\frac{2\ln(\frac{4}{\delta})}{m}}$$

Consistency of kernel ridge(less) regression

- Kernel ridge regression: $\underset{h\in\mathcal{H}}{\text{minimise}} \ \frac{1}{m}\sum_{j=1}^{m}(h(x_j)-y_j)^2 + \lambda_m\|h\|_{\mathcal{H}}^2$
  - Ridge-"less" case $(\lambda = 0)$: $\widehat{h}(\cdot) = \sum_{i=1}^{m}\alpha_i k(x_i,\cdot)$ is still a possible solution

Theorem Reg.9 (Consistency and inconsistency of kernel (least squares) regression)

- *Weak consistency of ridge regression (Christmann, Steinwart, Bernoulli, 2007):*
  If k *is a universal kernel, and distribution* $\mathcal{D}$ *satisfies* $\mathbb{E}_{(x,y)\sim\mathcal{D}}[|y|^2] < \infty$, *then if* $\lambda_m \to 0$ *and* $\lambda_m^4 m \to \infty$ *as* $m\to\infty$, *then the ridge solution satisfies* $L_{\mathcal{D}}^{sq}(\widehat{h}) \to L_{\mathcal{D}}^{*}$
- *Inconsistency of ridgeless regression (Rakhlin, Zhai, COLT, 2019; Malinar et al. arXiv:2207.06569):*
  Let $k(x,x') = e^{-\gamma\|x-x\|^2}$ *(Gaussian kernel) or* $e^{-\gamma\|x-x'\|}$ *(Laplace kernel) on* $\mathbb{R}^p$.
  There is a distribution $\mathcal{D}$ *such that* $L_{\mathcal{D}}^{sq}(\widehat{h}) - L_{\mathcal{D}}^{*} = \Omega(1)$