# Singular learning theory V : symmetry and RLCT



The central statement of singular learning theory is that the Bayes generalisation error of a singular model is determined by the RLCT, and in contradistinction to the case of regular models, the RLCT is not determined by the number of parameters in the model. With a fixed class of models, the RLCT varies as the true distribution varies. In connection with this, Watanabe [W, §7.6] makes the remarkable, and somewhat cryptic, statement that

$$\begin{aligned} &\text{simple function} \iff \text{complicated singularities} \\ &\text{complicated function} \iff \text{simple singularities.} \end{aligned} \tag{1.1}$$

Here "function" refers to the true distribution, and the complexity of a singularity is measured by the RLCT (complex singularities have smaller RLCTs). Since Kolmogorov complexity, and other complexity measures, lie at the heart of information theory, the above statement is potentially one of the deepest insights of statistical learning theory and information science, but it is not widely known.

In order to make Watanabe's discovery more accessible, we exhibit in this note the connection between complexity of functions and complexity of singularities in what we hope is a simple and transparent way, by emphasising the role of symmetry.

A highly symmetric function is simple, because it can be described with less information: a rotationally invariant function $f(x,y)$ may be described as a function $g(r)$. Let us explain how highly symmetric true distributions $q(y|x)$ lead to small RLCTs.

<u>Remark</u> Recall the RLCT is a measure of "effective number of parameters" in a model close to the most complex singularity of the set of true parameters, in the sense that in local coordinates $u_1, \ldots, u_d$ the set of true parameters is $u_1 = \cdots = u_{2\lambda} = 0$ where $\lambda$ is the RLCT (this is only strictly true in the "mildly singular" case, e.g. reduced rank regression). Since varying the remaining $u_{2\lambda+1}, \ldots, u_d$ doesn't change the "fit" of the model, we do not count them as parameters.

We take the same setup as the "Fisher for feedforward" notes (fforw), but where $f(x, w)$ is not necessarily a ReLU network. Then

$$K(w) = \int q(y|x)\, q(x) \log \frac{q(y|x)}{p(y|x,w)}\, dx\, dy. \tag{1.1}$$

We <u>do not</u> assume the true distribution is realisable, set $W_\alpha = \{ w \mid K(w) = \alpha \}$. As discussed in (sd14) beginning on p. (16), for $C \subseteq W$ compact (ignoring priors) if $W_0 \cap C \neq \phi$ the local RLCT $2\lambda_C$ is a measure of the effective codimension of $W_0 \cap C$ in $C$ (very roughly). In particular, it seems reasonable to assume that the more directions at $P \in W_0 \cap C$ <u>tangent</u> to $W_0 \cap C$, the smaller the RLCT (this is strictly true in the "mildly singular" case from (sd14) p. (17)).

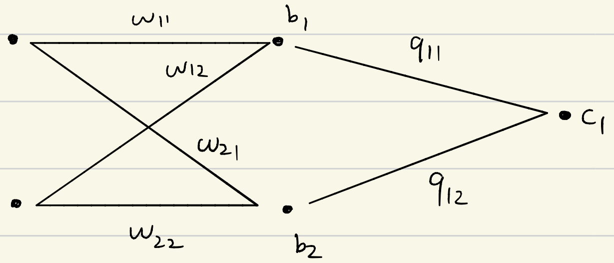
Suppose given a group $G$, and (not nec. continuous) actions $G \times W \longrightarrow W$, written $(g, w) \longmapsto g \cdot w$, and $G \times \mathbb{R}^N \longrightarrow \mathbb{R}^N$, written $(g, x) \longmapsto gx$.

<u>Def$^n$</u> We say the pairing $f: \mathbb{R}^N \times W \longrightarrow \mathbb{R}^M$ is <u>$G$-invariant</u> if

$$f(gx, gw) = f(x, w) \tag{1.2}$$

for all $x \in \mathbb{R}^N, w \in W, g \in G$ (equiv. $f(gx, w) = f(x, g^{-1}w)$ for all $x, w, g$).

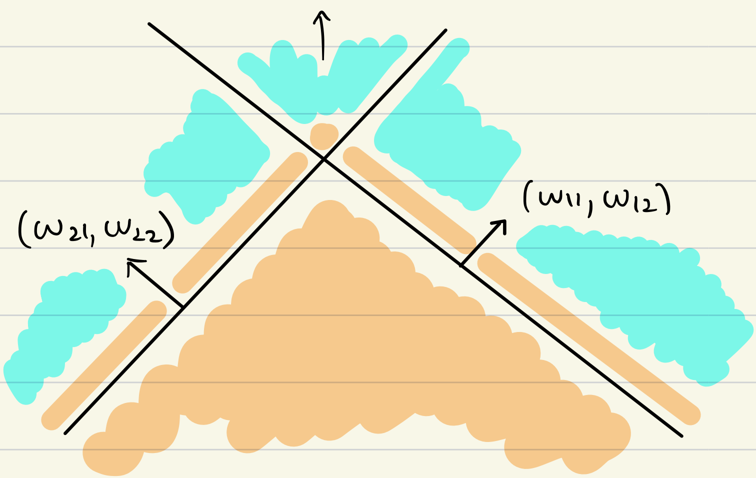
**Example 1** Consider a two layer feedforward ReLU net




(2.1)

with $W = \overline{B}_K(0) \subseteq \mathbb{R}^9$ for some $K$, $N=2$ and $M=1$, and $w = (w_{11}, w_{12}, w_{21}, w_{22}, b_1, b_2, q_{11}, q_{12}, c_1)$ determining for $x = (x_1, x_2)$

$$f(x,w) = q_{11}\,\mathrm{ReLU}(w_{11}x_1 + w_{12}x_2 + b_1) + q_{12}\,\mathrm{ReLU}(w_{21}x_1 + w_{22}x_2 + b_2) + c_1. \qquad (2.2)$$




$c_1 < 0$
$q_{11} > 0$
$q_{12} > 0$

Let $G = O(2)$ act on $\mathbb{R}^N = \mathbb{R}^2$ as usual, and on $W$ by

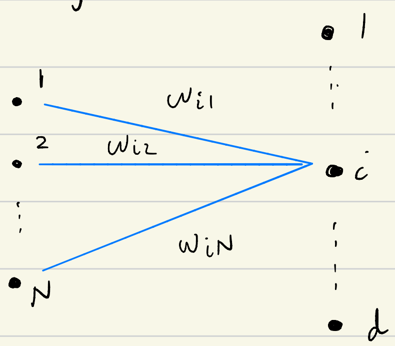
$$g \cdot w = (g\, w_{1\bullet},\ g\, w_{2\bullet},\ b_1, b_2, q_{11}, q_{12}, c_1). \qquad (2.3)$$

where $w_{1\bullet} = (w_{11}, w_{12})^T$, $w_{2\bullet} = (w_{21}, w_{22})^T$. We claim that $f : \mathbb{R}^N \times W \longrightarrow \mathbb{R}$ is $G$-invariant, with respect to these actions. Writing $\langle , \rangle$ for the dot product this follows from

$$
\begin{aligned}
f(gx, gw) &= q_{11}\,\mathrm{ReLU}(\langle g w_{1\bullet}, gx\rangle + b_1) \\
&\quad + q_{12}\,\mathrm{ReLU}(\langle g w_{2\bullet}, gx\rangle + b_2) + c_1 . \\
&= q_{11}\,\mathrm{ReLU}(\langle w_{1\bullet}, x\rangle + b_1) \\
&\quad + q_{12}\,\mathrm{ReLU}(\langle w_{2\bullet}, x\rangle + b_2) + c_1 \\
&= f(x, w).
\end{aligned}
\tag{3.1}
$$

More generally:

<u>Lemma</u> Let $f : \mathbb{R}^N \times W \longrightarrow \mathbb{R}^M$ be a ReLU network of arbitrary depth, and represent $W$ as $W_1 \times W'$ where $W_1 = (\mathbb{R}^N)^d$ are the weights in the first layer:




$$w = (w_{1\bullet}, w_{2\bullet}, \dots, w_{d\bullet}, w') \in W$$
$$w_{i\bullet} = (w_{i1}, \dots, w_{iN}) \in \mathbb{R}^N$$

We let $G = O(N)$ act on $\mathbb{R}^N$ in the standard way, and on $W$ by

$$
\begin{aligned}
G \times W_1 \times W' &\longrightarrow W_1 \times W' \\
(g, (w_{1\bullet}, \dots, w_{d\bullet}), w') &\longmapsto (g\, w_{1\bullet}, \dots, g\, w_{d\bullet}, w').
\end{aligned}
$$

Then $f$ is $G$-invariant.

<u>Proof</u> In the first layer, $f$ computes pre-activations as

$$x \longmapsto (\langle w_{1\bullet}, x\rangle, \dots, \langle w_{d\bullet}, x\rangle). \quad \square$$

<u>Lemma</u> If $f$ is $G$-invariant then $p(y|gx, gw) = p(y|x,w)$

<u>Proof</u> Clear since $p(y|x,w) := \frac{1}{(2\pi)^{M/2}} \exp\left(-\frac{1}{2}\| y - f(x,w)\|^2\right)$. $\square$

<u>Proposition</u> Let $G$ be a group acting on $\mathbb{R}^N$ and $W$ such that $f$ is $G$-invariant and

(i) $q(y|gx) = q(y|x)$ for all $x \in \mathbb{R}^N$, $y \in \mathbb{R}^M$, $g \in G$
(ii) $q(x) = q(gx)$ for all $x \in \mathbb{R}^N$, $g \in G$
(iii) For $g \in G$ the action $g\cdot(-) : \mathbb{R}^N \longrightarrow \mathbb{R}^N$ is smooth and $|\det(Dg)(x)| = 1$ for all $x \in \mathbb{R}^N$.

Then $K(w) = K(gw)$ for all $w \in W$, $g \in G$.

<u>Proof</u>

$$K(gw) = \int q(y|x)\, q(x) \log \frac{q(y|x)}{p(y|x, gw)}\, dx\, dy$$

$$= \int q(y|x)\, q(x) \log \frac{q(y|x)}{p(y|g^{-1}x, w)}\, dx\, dy$$

$$= \int q(y|g^{-1}x)\, q(g^{-1}x) \log \frac{q(y|g^{-1}x)}{p(y|g^{-1}x, w)}\, dx\, dy$$

Since $g\cdot(-) : \mathbb{R}^N \longrightarrow \mathbb{R}^N$ is smooth and bijective,

$$= \int q(y|x)\, q(x) \log \frac{q(y|x)}{p(y|x,w)} |\det(Jg)(x)|\, dx\, dy$$

$$= K(w). \square$$

<u>Corollary</u> The level sets $W_\alpha \subseteq W$ are $G$-invariant.

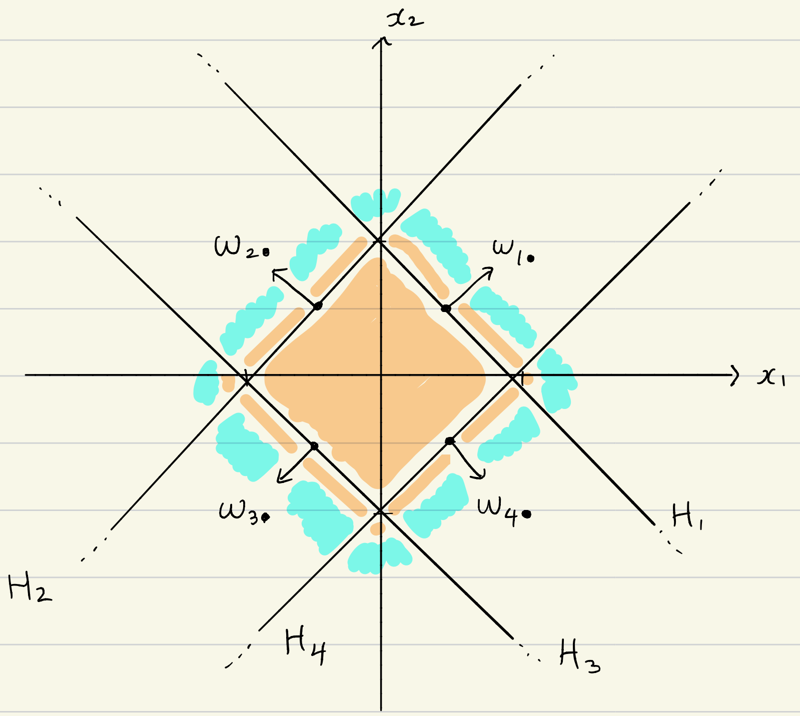
Example 2 In the situation of Example 1, we assume $q(x)$ is an $O(2)$-invariant distribution on $\mathbb{R}^2$, e.g. a normal distribution centered at $\underline{0}$, and an example of $q(y|x)$ satisfying hypothesis (i) of the proposition is

$$q(y|x) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\tfrac{1}{2}\| y - h(r) \|^2\right)$$

where $r = \|x\|$, for any continuous function $h: [0,\infty) \longrightarrow \mathbb{R}$. Then the proposition applies and all the level sets $W_\alpha$ are $O(2)$-invariant.

The problem here is that an $O(2)$-invariant true distribution which is nonconstant seems not to be realisable by a finite-depth ReLU network. However we can consider realisable true distributions which are invariant under finite subgroups $G \subseteq O(2)$.

Example 3 Consider a depth two ReLU network $f(-,w): \mathbb{R}^2 \longrightarrow \mathbb{R}$ where the hyperplanes $\langle w_{i\bullet}, x\rangle + b_i = 0$ (for $1 \le i \le d$ indexing a node in the first hidden layer) are mapped to one another by a finite subgroup $G \subseteq O(2)$ generated by rotations by $\frac{2\pi}{m}$, e.g.

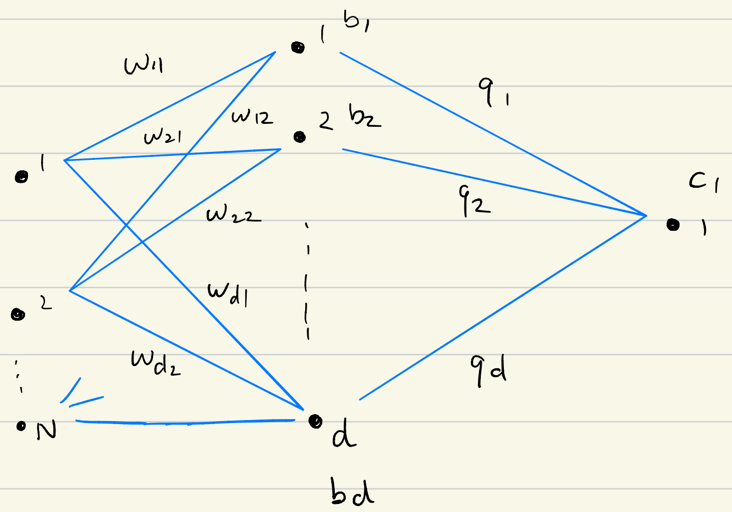



(5.1)

$m = 4$

This is a constraint on $w_0 \in W$ which ensures that $q(y|x) := p(y|x, w_0)$ satisfies the invariance condition of the Proposition on p. 4 for $g \in G$. So this true distribution is both $G$-invariant <u>and</u> realisable (by construction).

More generally, let $G \leq O(N)$ be a finite subgroup, $G = \langle g \rangle$ and suppose the network is given by weights $w_0 = ((w_{i\bullet})_{i=1}^d, b_\bullet, q_\bullet, c_1)$ as in




(6.1)

such that there is a permutation $\delta \in S_d$ with

(i) as functions $\mathbb{R}^N \longrightarrow \mathbb{R}$ for $1 \leq i \leq d$

$$\langle w_{i\bullet}, g x \rangle + b_i = \langle w_{\delta(i)\bullet}, x \rangle + b_{\delta(i)} \qquad (6.2)$$

(ii) $q_{\delta(i)} = q_i$ for $1 \leq i \leq d$.

Then

$$f(gx, w_0) = c_1 + \sum_{i=1}^{d} q_i \,\mathrm{ReLU}(\langle w_{i\bullet}, gx \rangle + b_i) \qquad (6.3)$$

$$= c_1 + \sum_i q_{\delta(i)} \mathrm{ReLU}(\langle w_{\delta(i)\bullet}, x \rangle + b_{\delta(i)})$$

$$= f(x, w_0).$$

Note that condition (ii) can be realised by taking all $q_i$ equal. To reason about (i) let us assume none of the $w_{i\bullet}$ are zero vectors, so $\langle w_{i\bullet}, - \rangle : \mathbb{R}^N \longrightarrow \mathbb{R}$ is surjective and $\mathbb{R}^N/K \cong \mathbb{R}$ where $K$ is the kernel. Let $t_i \in \mathbb{R}^N$ be such that $\langle w_{i\bullet}, -t_i \rangle = b_i$. Then

$$\langle w_{i\bullet}, x \rangle + b_i = 0 \iff \langle w_{i\bullet}, x - t_i \rangle = 0$$

and (6.2) is equivalent to (writing $t_i = g g^{-1} t_i$)

$$\langle g^{-1} w_{i\bullet}, x - g^{-1} t_i \rangle = \langle w_{\sigma(i)\bullet}, x - t_{\sigma(i)} \rangle$$

which we can arrange by e.g. $g^{-1} w_{i\bullet} = w_{\sigma(i)\bullet}$ and $g^{-1} t_i = t_{\sigma(i)}$.

<u>Lemma</u> Let $G \leq O(N)$ be a finite group generated by $g$, and suppose $(w_{1\bullet}, \ldots, w_{d\bullet}) \in (\mathbb{R}^N)^d$ and $(t_1, \ldots, t_d) \in (\mathbb{R}^N)^d$ and $\sigma \in S_d$ are such that

(a) $g^{-1} w_{i\bullet} = w_{\sigma(i)\bullet}$ for all $1 \leq i \leq d$
(b) $g^{-1} t_i = t_{\sigma(i)}$ for all $1 \leq i \leq d$.

Then let $w_0 = \left( (w_{i\bullet})_{i=1}^d, (-\langle w_{i\bullet}, t_i \rangle)_{i=1}^d, (q)_{i=1}^d, c \right)$ be the parameters for a two-layer feedforward ReLU network as in (6.1), from $\mathbb{R}^N \longrightarrow \mathbb{R}$, with biases $b_i = -\langle w_{i\bullet}, t_i \rangle$, and $q, c \in \mathbb{R}$ arbitrary. Then

$$f(gx, w_0) = f(x, w_0) \quad \forall x \in \mathbb{R}^N.$$

<u>Proof</u> Same calculation as (6.3), i.e.

$$f(gx, w_0) = c + q\sum_{i=1}^{d} \mathrm{ReLU}(\langle w_{i\bullet}, gx\rangle + b_i)$$

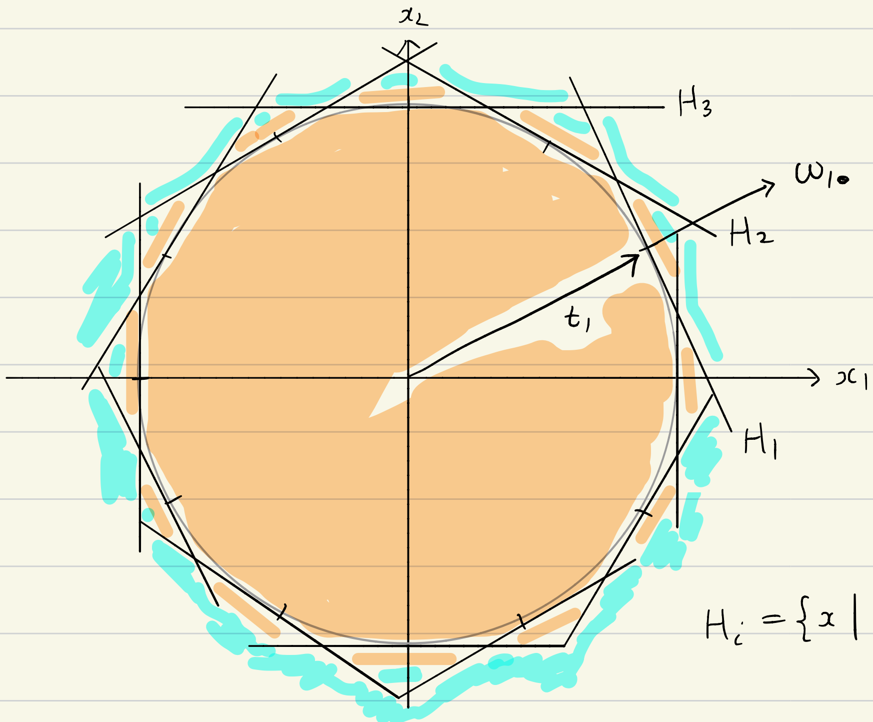
$$= c + q\sum_{i=1}^{d} \mathrm{ReLU}(\langle w_{i\bullet}, gx - t_i\rangle) \qquad (8.1)$$

$$= c + q\sum_{i=1}^{d} \mathrm{ReLU}(\langle g^{-1}(w_{i\bullet}), x - g^{-1}(t_i)\rangle)$$

$$= c + q\sum_{i=1}^{d} \mathrm{ReLU}(\langle w_{\sigma(i)\bullet}, x - t_{\sigma(i)}\rangle)$$

$$= f(x, w_0). \ \square$$

Example 4 To revisit Example 3 more concretely, let $g \in O(2)$ be rotation by $\frac{2\pi}{m}$ anticlockwise for some $m \geq 3$ and let $G = \langle g \rangle \cong \mathbb{Z}/m\mathbb{Z}$. Let $t_1 := g^{1/2}(1,0)^T$ where $g^{1/2}$ is rotation by $\frac{2\pi}{2m}$; define also $w_{i\bullet} = t_i := g^i t_1$, for $1 \leq i \leq m$.




$m = 6$

$q = 1$

$c < 0$

(8.2)

$H_i = \{x \mid \langle w_{i\bullet}, x\rangle + b_i = 0\}$

Then $g^{-1}w_{i\bullet} = g^{i-1}t_1 = w_{(i-1)\bullet}$ and $g^{-1}t_i = t_{i-1}$ where indices are read modulo $m$, so with $\sigma$ the cyclic permutation the hypotheses hold.

Taking $q(y|x) := p(y|x, w_0)$ for such $w_0$ constructs a $\mathbb{Z}_m$-invariant realisable true distribution for any $m \geqslant 3$. Notice that this true distribution is realisable for the ReLU architecture with $d$ nodes in the hidden layer for any $d \geqslant m$ (taking some $q_i$'s to be zero).

Consider a two-layer ReLU network architecture $f : \mathbb{R}^2 \times W \longrightarrow \mathbb{R}$ with $d$ nodes in the hidden layer, where $W$ is compact and $O(2)$-invariant. For $m < d$ let $q_m(y|x)$ be the $\mathbb{Z}_m$-invariant realisable true distribution constructed above and $\lambda_m$ its RLCT relative to some fixed prior.

Conjecture $\lambda_m$ is a decreasing function of $m$.

Less formally

A more symmetric true distribution $\Longrightarrow$ smaller RLCT

"simpler function" "more complicated singularity"

In their published work on RLCTs, Watanabe and collaborators tend to focus on rather simple true distributions, because it is already difficult to theoretically analyse the RLCT in these cases. However the deep idea on p. 1 is best illustrated with more interesting true distributions, and experimental approximation to the RLCT.

Remark We justify the application of singular learning theory to ReLU networks by the "soft ReLU trick".

Remark The idea of studying figures in the plane in connection with properties of neural networks is inspired by Minsky & Papert's book "Perceptrons: an introduction to computational geometry".