Learning Coefficients, Fractals, and Trees in Parameter Space

Max Hennick*
Department of Mathematics and Statistics
University of New Brunswick & TrojAI
mhennick@unb.ca
&
Matthias Dellago
University of Innsbruck
dellago.matt@gmail.com

June 23, 2025

Abstract

It is well known that the local geometric structure of the loss surface about a particular value of the parameter space for a deep learning model (and other singular models) determines many of the behavioural properties of the model with that parameter value. In recent years the learning coefficient has emerged as a particularly important geometric invariant for predicting model properties. In this work we explore the interpretation of the learning coefficient as a fractal dimension of the loss surface, and show how it relates to more classical notions of fractal dimensions like the box counting dimension. Using this we show that there is a natural correspondence between the low loss parameters and an infinite depth, locally finite tree. We then use this to reframe the learning coefficient in terms of cylinder sets, which draws links between the geometry of parameter space, information theory, symbolic dynamics, and probability on trees.

1 Introduction

Due to the potential risks posed by powerful AI models it has become increasingly important to understand the "laws" that govern the behaviour of these models both during inference and training. One important avenue of study is trying to understand the emergence of structures within these models during the training process. To this end, recent work has suggested the use of a local variant of the *learning coefficient* from singular learning theory [Wat09] (which

*Primary author.

controls the Bayesian posterior over the model parameters) to study how the internal structures of neural networks change throughout training [LFW+24] [WHvW+24]. It has been suggested that the learning coefficient is effectively a fractal dimension that controls the scale of the low-loss pathways that models can diffuse through during the learning process [HB25] controlling large portions of the learning dynamics.

One would expect that these local geometric structures should have some information theoretic description, as modern deep learning models seemingly learn compressed representations of the data that allow for effective inference [TZ15]. However, the exact relationship between the geometric structures of the loss surface and the compression picture is not well understood.

In this work we attempt to remedy this using a fractal geometry inspired approach. In particular, we show that the learning coefficient λ of some subspace $W \subset \mathbb{R}^d$ describes a fractal dimension λ' like $\lambda = d - \lambda'$ such that λ' meets some general conditions for a fractal dimension. Then, we show that under natural conditions on the loss surface geometry there is a direct relationship between the learning coefficient and the classical box counting dimension. This allows one to construct locally finite, infinite depth trees that capture properties of low loss subspaces of parameter space. From this, one can establish a link between the "description length" of a parameter specification in the low loss subspace and the Bayesian posterior over the set of possible model parameters.

2 Singular Learning Theory

Here we give a very brief introduction to singular learning theory through one of the core results. Let parameter set $W \subset \mathbb{R}^d$ be a compact subset, and let L be the population loss (in particular, the KL-divergence) for some data distribution. Following theorem 7.1 of [Wat09] for a given error tolerance t and a prior $\rho(w)$ over W, the volume of solutions with error less than t is given by the integral:

$$V(t) = \int_{L(w)<t} \rho(w)dw \tag{1}$$

which is referred to as **the singular integral**. This takes on the value

$$V(t) = ct^{\lambda}(\log(\frac{1}{t}))^{m-1} + o(t^{\lambda}(\log(\frac{1}{t}))^{m-1}) \tag{2}$$

where λ is the **learning coefficient** which is given by

$$\lambda = \lim_{t \to 0} \frac{\log(V(at)/V(t))}{\log a} \tag{3}$$

Similarly, one can define the "local learning coefficient" $\lambda(w^*)$ about a parameter w^* by taking the compact set W to be an open ball about w^*

We note here that for simplicity we will assume that the prior is the normalized Lebesgue measure over W, however the majority of results hold when the prior is any (Baire) probability measure.

3 Fractal Dimensions

Broadly speaking, fractal dimensions capture how some quantity scales under some particular choice of measurement gauge. The study of fractal geometry itself is a rich field with many applications ranging from physics and engineering [MB89] all the way to number theory[LvF00]. With such diverse applications, it should be unsurprising that there are many different types of fractal dimensions. We make use of one of the simplest fractal dimensions, the box counting dimension:

Definition 3.1 (Box Counting Dimension)**.** Consider a covering of $X \subset \mathbb{R}^d$ by a grid of cubes with sidelength δ. Let $M(\delta)$ be the least number of boxes of sidelength δ needed to cover X. Then we have

$$D_m = \lim_{\delta \to 0} \frac{\log M(\delta)}{\log \delta} \tag{4}$$

and $M(\delta) \sim u\delta^{-D_m}$ for some constant u. If this limit does not exist we can define the lower box counting dimension as

$$\underline{D_m} = \liminf_{\delta \to 0} \frac{\log M(\delta)}{\log \delta} \tag{5}$$

and the upper box counting dimension

$$\overline{D_m} = \limsup_{\delta \to 0} \frac{\log M(\delta)}{\log \delta} \tag{6}$$

Given that there are many different types of fractal dimensions, one might expect that there is some set of conditions that must be met for something to be considered a fractal dimension. Surprisingly, this is not really the case. Determining whether or not something is a fractal dimension is mostly guided by intuition as some things that don't look like a fractal dimension can indeed behave like one. Despite this, there are some general properties fractal dimensions tend to have. While something might still be a fractal dimension without these properties, something that has them can reasonably be considered a fractal dimension. These are given below[Fal13]:

Definition 3.2 (Conditions for a Fractal Dimension)**.** A scaling factor can generally be considered a fractal dimension if the following hold:

- *Monotonicity*: If $E \subset F$, then $\dim E \leq \dim F$
- *Stability*: $\dim E \cup F = \max(\dim E, \dim F)$
- *Lipschitz Invariance*: If T is a bilipschitz transformation, $\dim T(E) = \dim E$
- *Countable Sets*: If E is countable, then $\dim E = 0$.

- *Open Sets*: If E is an open subset of $\mathbb{R}^d$ then dim $E = d$
- *Smooth Manifolds*: If E is a smooth m-manifold then dim $E = m$

Before continuing, we note that one can use a different type of limit to get an equivalent definition of the fractal dimension. Suppose you want to know how the volume scales in relation to some fixed constant a. One then has

$$\lim_{\delta \to 0} \frac{M(a\delta)}{M(\delta)} \sim a^{D_m} \tag{7}$$

Since fractal dimensions are about how something "scales" with respect to some measurement, we can now define a method of measurement for our purpose. Let $V(h)$ represent the volume of some liquid in a container at height h. We can then define the height scaling dimension as:

Definition 3.3 (Height Scaling Method)**.** Let $X \subset \mathbb{R}^d$ and consider a smooth height function $f : \mathbb{R}^d \to \mathbb{R}^+$. If X contains a zero of f, let $\Omega_z = \{x \in \mathbb{R}^d | f(x) = z\}$. Now let $A(z) = \int_X 1_{\Omega_z}(x)dx$. We now define

$$V(h) = \int_0^h A(z)dz \tag{8}$$

we then have the height scaling dimension defined as

$$D_h = \lim_{h \to 0} \frac{\log V(h)}{\log h} \tag{9}$$

if it exists. One can define the upper and lower dimensions similarly. Furthermore, we can express

$$D_h = d - D'_h \tag{10}$$

It is easy to see that this is very similar to the definition of the singular integral. In the next section we show that the singular integral can be written in such a form and show that it meets the conditions of a fractal dimension.

4 Results

4.1 Learning Coefficient as a Fractal Dimension

Recently it has been shown that one can interpret the (local) learning coefficient as being approximately half the Hölder exponent [Fur25]. Here we take a slightly different approach, and show that the learning coefficient meets the general criteria of a fractal dimension.

Consider the d-dimensional Lebesgue measure μ on $\mathbb{R}^d$. The Lebesgue measure can be written as the integral

$$\mu(A) = \int_{\mathbb{R}^d} 1_A(w)dw \tag{11}$$

where 1_A is the indicator function over A. Suppose that $A \subseteq W$ and that we normalize μ by simply dividing by $\mu(W) = c$, giving $\overline{\mu}(A) = \frac{\mu(A)}{c}$. Letting $\rho_A = \frac{1_A}{c}$, we get

$$\overline{\mu}(A) = \int_W \rho_A(w)dw \tag{12}$$

Now letting $W_t = \{w \in W | L(w) < t\}$ be a sublevel set, we can consider:

$$\overline{\mu}(W_t) = \int_W \rho_{W_t}(w)dw \tag{13}$$

however notice that outside of W_t we have $\rho_{W_t}(w) = 0$. Denoting the compliment $W \backslash W_t = W'$ we can rewrite

$$\overline{\mu}(W_t) = \int_{W_t} \rho_{W_t}(w)dw + \int_{W'} \rho_{W_t}(w)dw \tag{14}$$

but clearly $\int_{W'} \rho_{W_t}(w)dw = 0$ so we can rewrite

$$\overline{\mu}(W_t) = \int_{W_t} \rho_{W_t}(w)dw \tag{15}$$

Expanding this integral

$$\int_{W_t} \rho_{W_t}(w) = \frac{1}{c} \int_0^t \int_{\Omega_t} 1_{\Omega_t}(w)dw \tag{16}$$

Now let $\theta(a) = 1$ if $a \geq 0$ and 0 otherwise. From the above integral we can rewrite this as

$$\overline{\mu}(W_t) = \int_W \theta(t - L(w))\rho(w)dw \tag{17}$$

with ρ simply being the uniform distribution. According to Equation 4.9 in [Wat09] this is $V(t)$, and is a cross sectional integral as is given in definition 3.3.

Before showing that the learning coefficient determines a fractal dimension, first notice that using equation 7 we can rewrite mass dimensions like:

$$\lim_{\delta \to 0} \frac{\log(M(a\delta)/M(\delta))}{\log a} = D_m \tag{18}$$

Now, from definition 3.3 we know that we can write the coefficient

$$D_m = \lim_{t \to 0} \frac{\log V(t)}{\log t} \tag{19}$$

and from equation 18 we have

$$\lim_{t \to 0} \frac{\log(V(at)/V(t))}{\log a} = D_m \tag{20}$$

and this limit exists by equation 3 and $D_m = \lambda$. So, the learning coefficient meets definition 3.3. Additionally we also need to make use of Rademacher's theorem:

Theorem 4.1 (Rademacher's Theorem)**.** *If $U \subset \mathbb{R}^d$ is an open set and $f : U \to \mathbb{R}^m$ is Lipschitz continuous, then f is differentiable almost everywhere.*

It now remains to be seen that the height scaling of the learning coefficient corresponds to the classical idea of a fractal dimension.

Lemma 4.1. *If the loss function is locally Lipschitz about its zeros, then denoting the learning coefficient as $\lambda = d - \lambda'$, λ' is a fractal dimension.*

Proof. First, note that the fractal dimension is only defined when the set contains a local minima and that it suffices to prove that definition 3.3 provides a fractal dimension. Now, monotonicity is immediate from the definition as an integral over a non-negative function. The countable sets property follows similarly. Furthermore, if $W_t \subset E \subset \mathbb{R}^n$ is an open set such that the low loss parameters W_t concentrate on some open subset of $E \subset \mathbb{R}^n$ with $n \leq d$ as $t \to 0$ then we can see that the integral $A(z) = \int_E 1_{\Omega_z}(x)dx$ as $z \to 0$ is then exactly the Lebesgue measure on $\mathbb{R}^n$ so then $\lambda' = n$ and thus the volume decays only outside the low loss subset, meaning that the open sets property is met. The smooth manifold property follows similarly. Stability can be seen from the fact that if $V_E(h), V_F(h)$ are the volumes of E, F respectively and suppose that $\lambda'_E > \lambda'_F$. If E, F are not disjoint, then by monotonicity $\lambda'_{E\cap F} \leq \lambda'_F < \lambda'_E$ so the volume of E is the slowest decaying. We then can write

$$V_{E\cup F}(h) = V_E(h) + V_F(h) - V_{E\cap F}(h) \tag{21}$$

Then consider that as $h \to 0$ both $V_F(h), V_{E\cap F}(h)$ are dominated by V_E so then

$$\lambda = \lim_{h\to 0} \frac{\log V_{E\cup F}(h)}{\log h} \tag{22}$$

$$= \lim_{h\to 0} \frac{\log V_E(h)}{\log h} \tag{23}$$

We are now just left to prove that the scaling exponent is preserved under Bilipschitz transformations.

Let T be a transformation of W which is Bilipschitz with

$$\frac{1}{\alpha}\|w_1 - w_2\| \leq \|T(w_1) - T(w_2)\| \leq \alpha\|w_1 - w_2\| \tag{24}$$

This tells us that

$$\frac{1}{\alpha}\|w_1 + h - w_1\| \leq \|T(w_1 + h) - T(w_1)\| \leq \alpha\|w_1 + h - w_1\| \tag{25}$$

and by Rademacher's theorem T is differentiable almost everywhere, so by applying the limit definition of the derivative we get

$$\frac{1}{\alpha}\|w\| \leq \|\partial T(w)\| \leq \alpha\|w\| \tag{26}$$

with the derivative operator ∂. It follows then that

$$(\frac{1}{\alpha})^d \leq |\det \partial T(w)| \leq \alpha^d \tag{27}$$

Now for any Lebesgue measurable set we have

$$\mu(T(A)) = \int_A |\det \partial T(w)| dw \tag{28}$$

which tells us that

$$(\frac{1}{\alpha})^d \mu(A) \leq \mu(T(A)) \leq \alpha^d \mu(A) \tag{29}$$

Now if we assume that T respects the zeros of the loss so that if $L(w_1) = 0$ then $L(T(w_1)) = 0$ (so that the fractal dimension is still defined under the transformation) and that the loss itself is locally Lipschitz around each 0, then we can see from the above that the volume of the level sets changes by at most a constant factor of α. This means that the change in volume of $V(h)$ under T is distorted by at most α^d but as $h \to 0$ since T must preserve zeros then we must have that $\alpha \to 1$ so $V_{T(W)}(h) \approx V_W(h)$ so λ' is preserved under zero-preserving Bilipschitz transformations.

□

From the above we get a useful corollary:

Corollary 4.1. *The learning coefficient can be expressed as* $\lambda = \frac{d}{2} - \lambda'$

The above corollary is simply a consequence of the fact that $\lambda \leq \frac{d}{2}$ then just redefining λ' appropriately.

5 Relating Gauges

We can now show that the learning coefficient can be written in terms of the box counting dimension. In particular, it can be written in terms of how the box counting dimension changes as the maximum acceptable error decreases. We first start with some simple lemmas required for the forthcoming results.

Recall that a function L is k-Lipschitz on W if we have that for all $w_1, w_2 \in W$ we have that

$$|L(w_1) - L(w_2)| \leq kd(w_1, w_2) \tag{30}$$

Now suppose that there is some t such that if $L(w^*) < t$ then there is some small radius r about w^*, then L is at least k-Lipschitz on the open ball $B(w^*, r)$. Different values of w^* can have different values of r and k. However, we can actually pick fixed values for these that work for all possible points.

Lemma 5.1. *If for all* $w^* \in W$ *such that* $L(w^*) < t$ *we have that there is some small radius* r *about* w^* *such that* L *is at least* k*-Lipschitz on the open ball* $B(w^*, r)$*, then there is some fixed* k^**,* r^* *such that all such* w^* *are* k^**-Lipschitz on a ball of radius* r^**.*

Proof. First, note that if w^* is k-Lipschitz on $B(w^*, r)$, it must also be k-Lipschitz on $B(w^*, r^*)$ for all $0 < r^* \leq r$. So, we simply take the value of r^* to be the smallest radius required by any low loss parameter. Similarly, if w^* is k-Lipschitz on $B(w^*, r)$ it must also be k^*-Lipschitz for all $k^* \geq k$. □

While a simple result, it allows us to formalize what it means to say that the function does not oscillate too violently near the low loss subsets. In order to continue we will need to make use of the *implicit function theorem*, which we state below for convenience.

Theorem 5.1 (Implicit Function Theorem)**.** *Let $f : \mathbb{R}^{n+m} \to \mathbb{R}^m$ be a continuous function, and suppose that there exists $x_0 \in \mathbb{R}^m, y_0 \in \mathbb{R}^n$ such that $f(x_0, y_0) = 0$. If the Jacobian matrix $D_y f(x_0, y_0)$ is invertible then there exists neighbourhoods U, V of x_0, y_0 and a differentiable function ρ such that $f(x, \rho(x)) = 0$ for all $x \in U$.*

Define the level set of a compact set W with loss function L as $\overline{W}_t = \{w \in W | L(w) = t\}$. Using this, we give a result that gives conditions that ensure the level sets are sufficiently "nice":

Lemma 5.2. *Suppose L is at least d differentiable on $\mathbb{R}^d$. Then for any $t > 0$, the boundary of $W_t = \{w \in W | L(w) < t\}$ has Lebesgue measure 0 if $\nabla L(\overline{w}) \neq 0$ for all $\overline{w} \in \overline{W}_t$.*

Proof. Since $\nabla L(\overline{w}) \neq 0$ for all $\overline{w} \in \overline{W}_t$, there is some direction where the gradient is non-zero. Let this direction be denoted by w_i, so we can rewrite $L(w_{-i}, w_i)$. Furthermore we get $\frac{\partial L}{\partial w_i}(w_{-i}, w_i) \neq 0$ which is the same as the Jacobian $D_{w_i} L(w_{-i}, w_i)$. Next note that we can rewrite $L(w_{-i}, w_i) = t$ as $L(w_{-i}, w_i) - t = 0$ so we get by the implicit function theorem that for every point in $\overline{W}_t$ that there is a function $\rho : \mathbb{R}^{d-1} \to \mathbb{R}$ that defines the dth coordinate. Then if we have U as an open set about w_{-i} we can define a local diffeomorphism $\phi : U \to V \subset \overline{W}_t$ by $\phi(w_{-i}) = (w_{-i}, \rho(w_{-i}))$. This means that the boundary $\overline{W}_t$ is locally at most a $d-1$ dimensional submanifold, so one can then define the submanifold $\overline{W}_t$ by charts (w_{-i}, ϕ). It then follows immediately that we d-dimensional Lebesgue measure of the boundary must be 0.

□

While this might seem restrictive, it is simply a formal way of saying that so long as the level sets themselves are not too fractal, their individual contributions to the volume of the sublevel sets is 0. In fact, the above lemma is essentially a simplified version of Sard's theorem. Furthermore we note that the condition that $\nabla L(\overline{w}) \neq 0$ for all $\overline{w} \in \overline{W}_t$ is also necessary for the definition of the singular integral by theorem 4.3 of [Wat09].

5.1 The Learning Coefficient in Sidelength Gauges

5.1.1 Box Counting Dimension with the Euclidean Metric

Let W_t be the set of low loss parameters as before and consider a rectangular cover of W_t by a grid of cubes of side length δ. For each cube in this cover Q_i^δ, if

Q_i^δ does not contain any parameters in W_t, we remove Q_i^δ. The remaining cubes give us a cover $C'(\delta)$ of W_t. Let $|C'(\delta)|$ be the number of cubes in the cover, and let $C(\delta)$ be the minimal cover of W_t by δ-cubes. By definition we have that $|C(\delta)| \leq |C'(\delta)|$. Furthermore, each cube has Lebesgue measure $\mu(Q_i^\delta) = \delta^d$. This method of box counting is sometimes referred to as a dyadic covering, and defines a fractal dimension that is effectively equivalent to the box counting dimension [Fal13].

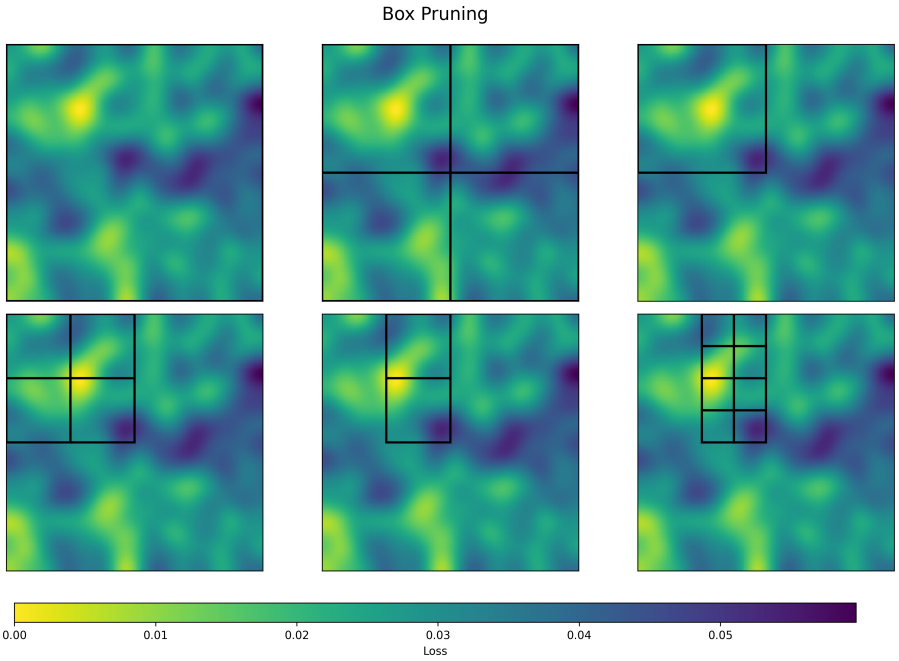

Figure 1: An example of the initial steps in the cube-covering method with $t = 0.001$.

Consider that for a cube of sidelength δ in $\mathbb{R}^d$, one can fit 2^d cubes of sidelength $\frac{\delta}{2}$ into said cube. Given a cover $C'(\delta)$ we can define a new cover by cubes of sidelength $\frac{\delta}{2}$ by splitting each cube Q_i^δ into cubes $Q_{i,j}^{\frac{\delta}{2}}$ and removing any cubes which don't contain true parameters. One can iterate on this process ad infinitum. With this process in mind, we are now equipped to show the following:

Theorem 5.2 (Learning Coefficient in the Sidelength Gauge)**.** *If $D_m(t)$ is the box counting dimension of W_t in the sidelength gauge δ. Assuming the conditions of lemmas 5.1 and 5.2 hold then for small δ we have that*

$$\lambda \log a + o(1) \approx b\delta^{D_m(t) - D_m(at)} + o(1) \tag{31}$$

with b being a constant.

Proof. First note that by lemma 5.1 that for all points $w \in W_t$ such that $L(w) < t - \varepsilon$ where $\varepsilon > 0$ is some arbitrary constant, then there is some sidelength scale δ such that if cube Q contains such if $w \in W_t$ then $Q \subset W_t$. Or, in other words, there is a scale such that one point in the cube with small error means all points in the cube have small error, unless the cube contains a point with loss arbitrarily close to t, meaning it contains some amount of points in exterior of W_t.

Now let Q_i^n denote a cube after n steps of the iteration process described previously such that $Q_i^n \subset W_t$ and let $\overline{Q}_i^n$ denote cubes that contain exterior points. We then have the following inequality

$$\sum_i \overline{\mu}(Q_i^n) \leq V(t) \leq \sum_i \overline{\mu}(Q_i^n) + \sum_j \overline{\mu}(\overline{Q}_j^n) \tag{32}$$

We can see as well that $\sum_i \overline{\mu}(Q_i^n) \leq \sum_i \overline{\mu}(Q_i^{n+1})$ because if a cube is completely contained in W_t it doesn't lose any volume during the pruning process, and at the $n+1$ step some set of cubes which contain some number of exterior points can be shrunk and pruned, leaving them containing only interior points. This also means that $\sum_j \overline{\mu}(\overline{Q}_j^n) \geq \sum_j \overline{\mu}(\overline{Q}_j^{n+1})$. In fact, one can see that this inequality must be strict near the boundary, so $\sum_j \overline{\mu}(\overline{Q}_j^n)$ is decreasing in n. This implies as well that $\sum_i \overline{\mu}(Q_i^n)$ is increasing proportional to the change in $\sum_j \overline{\mu}(\overline{Q}_j^n)$.

By monotone convergence we have that

$$\lim_{n\to\infty} \sum_i \overline{\mu}(Q_i^n) + \sum_j \overline{\mu}(\overline{Q}_j^n) = \overline{\mu}(W_t) + \overline{\mu}(\overline{W}_t) \tag{33}$$

but under the assumptions of lemma 5.2 we know that $\overline{\mu}(\overline{W}_t) = 0$, so

$$\lim_{n\to\infty} \sum_i \overline{\mu}(Q_i^n) = V(t) \tag{34}$$

Furthermore this tells us that for large n, $\sum_j \overline{\mu}(\overline{Q}_j^n)$ can be made arbitrarily small.

In the above construction we used the cube division process we know that $|C(\delta)| \leq |C'(\delta)|$ so we can replace the $n \to \infty$ limit with $\delta \to 0$ and rewrite

$$\lim_{\delta\to 0} \sum_i \overline{\mu}(Q_i^\delta) = V(t) \tag{35}$$

since the total volume of the minimal cover is at most equivalent to the volume of the cube cutting cover, and the volumes must agree in their respective limits.

Consider now the cover by cubes of sidelength δ for small δ such that

$$V(t) - \sum_i \overline{\mu}(Q_i^\delta) + o(\delta) \tag{36}$$

we know that $\overline{\mu}(Q_i^\delta) = \overline{\mu}(Q_j^\delta)$ for all i, j since each Q_i are cubes of identical size. We have then

$$V(t) = |C(\delta)|\overline{\mu}(Q_i^\delta) + o(\delta) = |C(\delta)|\frac{\delta^d}{c} + o(\delta) \tag{37}$$

since the Lebesgue measure of a d-cube is δ^d. By definition 3.1 we get

$$V(t) = u_t \delta^{-D_m(t)} \frac{\delta^d}{c} + o(\delta) \tag{38}$$

$$= \frac{u_t}{c} \delta^{d-D_m(t)} + o(\delta) \tag{39}$$

Consider then that

$$\frac{V(at)}{V(t)} = \frac{u_{at}\delta^{d-D_m(at)} + o(\delta)}{u_t \delta^{d-D_m(t)} + o(\delta)} \tag{40}$$

It then follows that

$$\frac{u_{at}\delta^{d-D_m(at)} + o(\delta)}{u_t \delta^{d-D_m(t)} + o(\delta)} = b\delta^{D_m(t)-D_m(at)} + o(1) \tag{41}$$

where b is some scalar $b = \frac{u_{at}}{u_t}$ and since $\frac{V(at)}{V(t)} = \lambda \log a + o(1)$ we get

$$b\delta^{D_m(t)-D_m(at)} + o(1) = \lambda \log a + o(1) \tag{42}$$

as desired. □

What this tells us is that the learning coefficient controls how the box counting dimension changes in the error gauge. We also note that if the conditions of lemma 5.2 do not hold, we still have that equations 32 and 33 hold. In this case however, the measure of the boundary $\overline{\mu}(\overline{W}_t)$ has some constant irreducible volume. One can view this as a by-product of the fact that we care about the approximation by "inner covers", which end up effectively tracking the boundary. Considering this case, we still have that in the limit the interior sets

$$\lim_{n\to\infty} \sum_i \overline{\mu}(Q_i^n) = V(t) \tag{43}$$

still holds, however we have that

$$b\delta^{d-D_m(t)} - \overline{\mu}(\overline{W}_t) \geq V(t) \tag{44}$$

instead of some vanishing term in δ.

One reason why casting the learning coefficient as a box counting dimension is useful is that it gives what one might consider a "generalization" since the learning coefficient depends on the existence of a local minima where the rate of change of the box counting dimension can be defined anywhere. More importantly however, it allows us to cast the learning coefficient as a property of a family of trees.

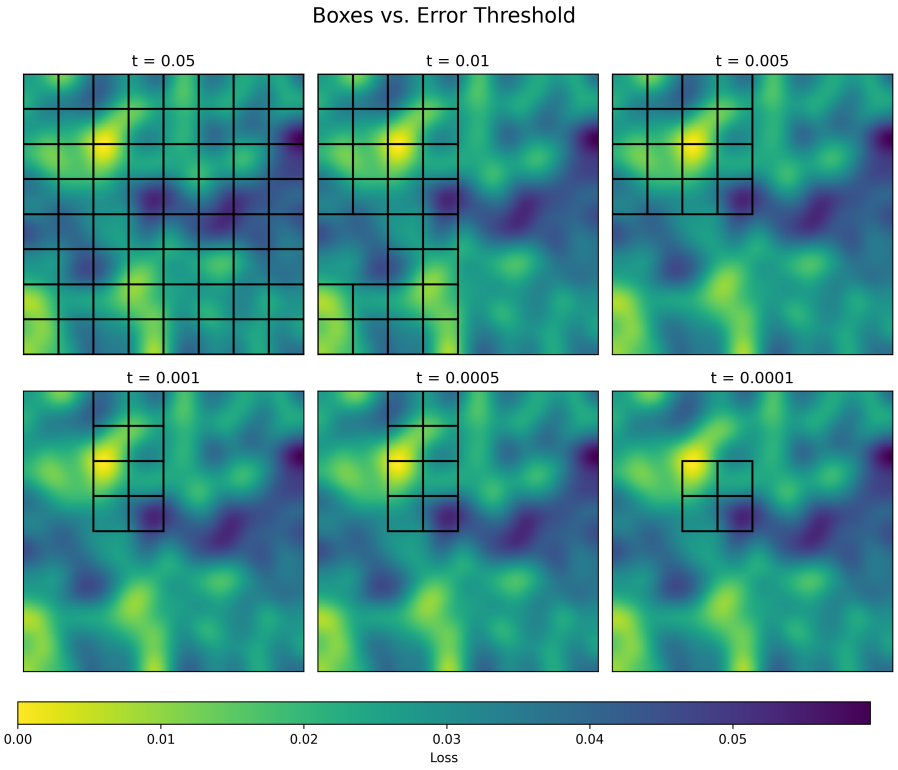

Figure 2: How the cube cover in the bottom right frame of figure 1 changes as t decreases.

5.2 Tree Structure and the Learning Coefficient

Why should one care about the box counting dimension? It turns out that the box counting dimension effectively tells us how to embed the set of low loss parameters within an infinite depth, locally finite tree.

Lemma 5.3. *$W_t \subseteq W$ be the set of parameters with loss t. Every such subset has an associated infinite depth, locally finite tree.*

Proof. Similar to the process outlined at the start of section 5.1.1, let $W_t \subset W \subseteq Q_0$ where Q_0 is a closed cube. For simplicity we assume that the sidelength of Q_0 is normalized to 1. Starting with Q_0, for each positive integer k we subdivide the cubes of sidelength 2^{-k} that intersect W_t into 2^n cubes of sidelength 2^{-k+1}. One then forms a tree by selecting the vertices at depth k to be all cubes of sidelength 2^{-k} that intersect W_t and defining edges between vertices v_i^k, v_j^{k+1} if v_j^{k+1} is a sub-cube defined by the initial division of the cube associated to the vertex v_i^k. We can see from this as well that for any vertex v, v must have a finite amount of children.

Denote the tree constructed by the above process as $\mathcal{T}_{W_t}$. Notice then that for any $w \in W_t$ one can associate a unique infinite sequence of cubes ... $\subset Q_1(w) \subset Q_0(w)$. This sequence of nested cubes defines an infinite ray/branch γ_w in the tree. Let $\partial\mathcal{T}_{W_t}$ represent the set of all such rays for the tree. We see then that the map $\pi : W_t \to \partial\mathcal{T}_{W_t}$ given by $\pi(w) = \gamma_w$ defines an embedding of W_t into the boundary of an infinite depth, locally finite tree. □

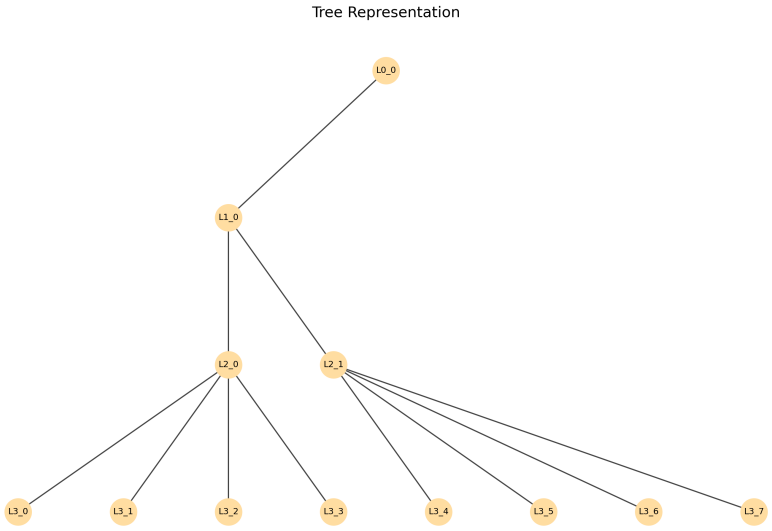

Figure 3: The tree (up to depth 4) corresponding to the cube counting steps in figure 1.

There is an issue of ambiguity where one point can have multiple rays if it lies on the boundary of a cube. However, one can resolve this by picking a side of the boundary to assign to the point. This is effectively the same as binary fractions having two binary expansions.

Importantly there is a natural metric on $\partial\mathcal{T}_{W_t}$.

Definition 5.1 (Ultrametric)**.** For $w_1, w_2 \in W_t$ let

$$r = \min_k \{Q_k(w_1) \neq W_2\} \tag{45}$$

then the the ultrametric on $\partial\mathcal{T}_{W_t}$ is

$$d_\tau(\gamma_{w_1}, \gamma_{w_1}) = 2^{-r} \tag{46}$$

Intuitively we can see that two points which are close in W_t are close in $\partial\mathcal{T}_{W_t}$. We can show this formally:

Lemma 5.4. *The value $d(w_1, w_2)$ gives a lower bound on the distance $d_\tau(\pi(w_1), \pi(w_2))$ and the distance $d_\tau(\pi(w_1), \pi(w_2))$ gives an upper bound on the Euclidean distance between w_1 and w_2.*

Proof. Notice that for any cube Q_k in the subdivision process that the diameter is 2^{-k} then if w_1 and w_2 are in Q_k then in Euclidean space $d(w_1, w_2) \leq 2^{-k}$ and since w_1, w_2 both belong to Q_k we know that they must agree up to k vertices so $d_\tau(\pi(w_1), \pi(w_2)) \leq 2^{-k}$. Similarly, if $d_\tau(\pi(w_1), \pi(w_2)) \leq 2^{-k}$ then $d(w_1, w_2) \leq 2^{-k}$ since they must both be contained in the same cube up to iteration at least k. □

Now for tree $\mathcal{T}_{W_t}$ let $\mathcal{T}_{W_t}(n)$ be the set of vertices at depth n. When discussing trees, there are ways to define the rate at which trees grow. We give such a definition below:

Definition 5.2 (Tree Growth Rate)**.** Let $\mathcal{T}$ be an infinite depth, locally finite tree with $|\mathcal{T}(n)|$ vertices at depth n. The upper growth rate is

$$\log \overline{\mathrm{gr}\mathcal{T}} = \limsup_{n\to\infty} \frac{\log |\mathcal{T}(n)|}{n} \tag{47}$$

with the lower growth rate

$$\log \underline{\mathrm{gr}\mathcal{T}} = \liminf_{n\to\infty} \frac{\log |\mathcal{T}(n)|}{n} \tag{48}$$

and the growth rate

$$\log \mathrm{gr}\mathcal{T} = \lim_{n\to\infty} \frac{\log |\mathcal{T}(n)|}{n} \tag{49}$$

if it exists.

We can now prove the following:

Lemma 5.5. *The box counting dimension D_m of W_t is equivalent to the growth rate of the tree associated to W_t (up to a factor of $\log \frac{1}{2}$).*

Proof. For simplicity, assume that both the upper and lower limits exist so we can work simply with lim case (as the proof is identical for both the upper and lower dimensions). Now consider the box counting dimension of W_t given by the cube covering method:

$$D_m(W_t) = \lim_{\delta\to 0} \frac{\log M(W_t, \delta)}{\log \delta} \tag{50}$$

Now we can rewrite $\delta = 2^{-n}$ which gives

$$D_m(W_t) = \lim_{n\to\infty} \frac{\log M(W_t, 2^{-n})}{-n \log 2} \tag{51}$$

By the construction in lemma 5.3 we know that $M(W_t, 2^{-n})$ is the same as $|\mathcal{T}_{W_t}(n)|$ and the factor of $-\log 2$ in the above is constant so we can rewrite

$$(-\log 2)D_m(W_t) = \lim_{n\to\infty} \frac{\log |\mathcal{T}_{W_t}(n)|}{n} \tag{52}$$

which is exactly the definition of the growth rate of the tree. □

Converting the box counting dimension to the growth rate of the tree is useful as it allows one to talk more information theoretically about quantities relating to low-loss parameters. In fact, it tells us that the learning coefficient describes the growth of "cylinder sets".

5.3 Learning Coefficient as the Growth Rate of Cylinder Sets

First, in the tree picture if our parameter space has dimension d then we know that the number of cubes of sidelength $\frac{\delta}{2}$ that can fit in the cube of sidelength δ is 2^d so each node in the tree associated with the parameter set W_t has at most 2^d children. What we can do is define an alphabet Σ consisting of 2^d symbols and pick some canonical symbol assignment scheme since every cube (and thus possible subcube inside of it) are identical. Each ray then in the tree corresponds to a unique infinite length string which specifies a model with loss less than t. Assuming the loss is bounded above, if we take t large enough we get $W_t = W$.

We can now give a functional definition of cylinder sets, which relate trees to information theory and symbolic dynamics.

Definition 5.3 (Cylinder Sets)**.** Let Σ be an alphabet of symbols, let Σ^* be the set of all finite words formed by symbols in Σ, and let Σ^∞ be all (left rooted) infinite length words. For each sequence $s \in \Sigma^*$ define the cylinder set $\Gamma(s)$ as the set of all $s^\infty \in \Sigma^\infty$ that begin with s.

Cylinder sets have many different, equivalent definitions but we simply pick this definition as it is well-suited for our purposes. Now if one picks any vertex v in an infinite depth tree, v corresponds to some finite length prefix and thus all points in the boundary of the tree which can be "traced back" to v form a subset of the cylinder set $\Gamma(v)$. Based on this, we give the following:

Lemma 5.6. *The learning coefficient captures the growth rate of the sequence of trees associated with the sets W_t as $t \to 0$.*

Proof. Suppose W is a cube (or take a cube which encloses W), we can see that for any W the first node in the tree effectively corresponds to the whole of W. We can call this node e, corresponding to the empty word. Now taking W_0 as the set of all parameters such that $L(w) = 0$. Associating the tree $\mathcal{T}_{W_0}$ gives a boundary $\partial\mathcal{T}_{W_0}$. This forms a subset of the cylinder set $\Gamma(e)$. We then have a sequence of nested trees/sets $\{\partial\mathcal{T}_{W_t}\}_{t\in[0,\infty)}$. Since L is bounded there is some finite t such that $\partial\mathcal{T}_{W_t} = \partial\mathcal{T}_W$ and we can see that $\partial\mathcal{T}_W = \Gamma(e)$.

Let $F : W \to \Sigma^\infty$ denote the embedding of W into an infinite depth tree. Notice that we get a measure on the space of infinite length strings by a push-forward μ^* of the normalized Lebesgue measure $\overline{\mu}$ on W by $\mu^*(A) = \mu(F^{-1}(A))$ for $A \subset \Gamma(e)$. By equation 17 this says:

$$\mu^*(\partial \mathcal{T}_{W_t}) = \int_W \theta(t - L(w))\rho(w)dw \tag{53}$$

$$= V(t) \tag{54}$$

It follows straightforwardly then that the learning coefficient captures how the growth rate of the sequence of trees associated with W_t changes as $t \to 0$ by theorem 5.2. □

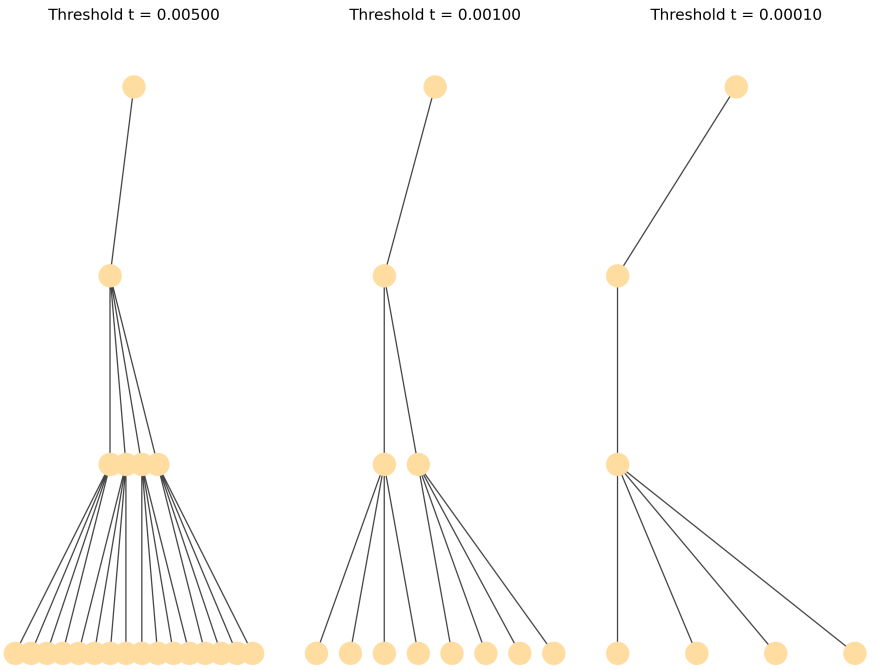

Figure 4: Visualizing how the tree in figure 4 changes as t decreases.

While it has been suggested before that there is a direct relationship between the size of a local minima and the number of bits needed to specify that minima up to some precision [LFW+24] [HS97]. Using the cylinder set/tree picture, we can give a somewhat more formal description of this phenomena.

Theorem 5.3 (Local Description Length)**.** *Let w_1, w_2 be two points in W such that $L(w_1), L(w_2) = 0$ and suppose L is at least k-Lipschitz about its zeros, and*

let t be some error rate such that $t > k$. Then there is a natural association of w_1, w_2 to cylinder sets $\Gamma(w_1), \Gamma(w_2)$ described by prefixes s_1, s_2 such that $\mu^*(\Gamma(w_1)) \geq \mu^*(\Gamma(w_2))$ if and only if Len$(s_1) \leq$ Len(s_2).

Proof. First notice that by the Lipschitz assumption there must exist cubes Q_1, Q_2 described by the iteration process described in section 5.1.1 for each w_1, w_2 such that for all $w \in Q_i$, we have $L(w) < t$ where each Q_i is associated to some prefix s_i, and $\text{Len}(s_i) = n$ where n is the number of steps taken to get box Q_i within the iteration process. Let Q_i be the largest such box for each w_i so each subtree rooted at s_i must have full growth rate, and and where the rays of this subtree correspond to the cylinder set $\Gamma(s_i)$. We can see that $\overline{\mu}(Q_i) \propto \frac{1}{2^{\text{Len}(s_i)}}$ so boxes whose associated prefix is shorter have a larger volume.

In the other direction, suppose without loss of generality, suppose Q_1 is larger than Q_2. Then we must have that Q_1 corresponds to a shorter prefix than Q_2 by construction, and the same measure argument as applied previously gives the result. □

This tells us that the learning coefficient describes how hard it is to specify a model as our error tolerance decreases.

6 Discussion and Conclusion

6.1 Related Work

The relationship between model description length and performance (namely the generalization behaviour) has been well-known for over 50 years [Sol64]. However, particularly in the case of neural networks, it is known that the generalization of the model is related to flat minima [HS97] [MM20], but the exact relationship between flat minima and description length has not had an adequate formal explanation.

There has also been recent work in relating singular learning theory to information theory via program synthesis. In particular, [CMW21] explores programs as singularities of analytic varieties which correspond to phases in a Bayesian posterior. This direction of investigation was expanded upon in [MT25] which effectively relates internal structure of Turing machines to the local geometry of a parameter space that encodes them. While there is seemingly a correspondence to the geometric structure of program synthesis and the work done here, the formal relationship is unclear.

One of the primary goals of framing the learning coefficient as a more information theoretic quantity is to help interpret the internal phases of a neural network during training. The local learning coefficient has been used to study phase transitions in neural networks [LFW$^+$24]. Using a refined verion of the LLC, [WHvW$^+$24] was able to study how attention heads take on different functional roles during training. While the learning coefficient is a Bayesian quantity, its relationship with phase transitions in SGD was explored more directly

in [CLM+23], providing insight into the influence of the Bayesian psoterior on the dyanmics of stochastic gradient descent.

6.2 Future Work

Here we outline some potential directions for future work. While the work here provides a more information theoretic perspective on singular learning theory, opening up new routes of analysis. In particular, the use of infinite depth, locally finite trees provides a direct link to symbolic dynamics through the cylinder set topology and suggests that one might be able reframe singular learning theory in terms of symbolic dynamics in some cases, though this has not yet been explored. In a similar way, one can use the notion of an infinite length string to relate models to the Kolmogorov complexity through the lower incompressibility ratio, however the formal relationship of the lower incompressibility ratio with quantities like the LLC is not clear. Finally, one should be able to use the theory of probability on trees to study different properties of the Bayesian posterior by looking at "flows" through the tree. This is probably the most straightforward way to extend some of the results here as theorem 5.3 can be reframed in terms of probability flows through trees in a straightforward way.

6.3 Conclusion

Here we have shown that the structure of low loss parameter sets is fractal, then relating the learning coefficient to other more classical notions of fractal dimensions. Using this, we showed how one can reconceptualize the space of low loss parameters as an infinite depth tree and that this tree captures many meaningful properties about the low loss parameters in a natural way and that the learning coefficient corresponds to effectively to how this tree changes as the maximum allowable loss decreases. By doing this, we have built a bridge between singular learning theory, computable analysis, and trees, and information theory. We hope that such theoretical insights can be used further the understanding of the internal structure of AI systems and how they evolve during training.

References

[CLM+23] Zhongtian Chen, Edmund Lau, Jake Mendel, Susan Wei, and Daniel Murfet. Dynamical versus bayesian phase transitions in a toy model of superposition, 2023.

[CMW21] James Clift, Daniel Murfet, and James Wallbridge. Geometry of program synthesis, 2021.

[Fal13] K. Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. Wiley, 2013.

[Fur25] Zach Furman. The local learning coefficient as a fractal dimension. https://zachfurman.com/notes/llc-fractal-dimension.pdf, 2025. Accessed: 2025-06-03.

[HB25] Max Hennick and Stijn De Baerdemacker. Almost bayesian: The fractal dynamics of stochastic gradient descent, 2025.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 01 1997.

[LFW+24] Edmund Lau, Zach Furman, George Wang, Daniel Murfet, and Susan Wei. The local learning coefficient: A singularity-aware complexity measure, 2024.

[LvF00] Michel Lapidus and Machiel van Frankenhuijsen. *Fractal Geometry and Number Theory*. Springer, 01 2000.

[MB89] B. B. Mandelbrot and A. Blumen. Fractal geometry: What is it, and what does it do? [and discussion]. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 423(1864):3–16, 1989.

[MM20] Rotem Mulayoff and Tomer Michaeli. Unique properties of flat minima in deep networks, 2020.

[MT25] Daniel Murfet and Will Troiani. Programs as singularities, 2025.

[Sol64] R.J. Solomonoff. A formal theory of inductive inference. part i. *Information and Control*, 7(1):1–22, 1964.

[TZ15] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle, 2015.

[Wat09] Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. 2009.

[WHvW+24] George Wang, Jesse Hoogland, Stan van Wingerden, Zach Furman, and Daniel Murfet. Differentiation and specialization of attention heads via the refined local learning coefficient, 2024.