Mathematical Aspects of Deep Learning

Mathematics of Deep Learning: Lecture 6 – Simple hierarchical models

Transcribed by Govind Ramnarayan (edited by Mathew Brennan, Asad Lodhia and Elchanan Mossel)

Some Very Simple (Gaussian) Hierarchical Models

We start by describing some very simple hierarchical models on the *d-ary tree* . Recall that in the d-ary tree, each node in the tree has exactly d children (As opposed to the d-regular tree, in which the root has d children, and every other node has $d - 1$ children). Both models will feature an unknown character at the root, chosen according to some distribution. Each node propagates its character to its children with some Gaussian noise. The goal will be to recover the character at the root given the characters at some level ℓ.

1. Version 1 (Brownian Motion)

We will denote the character at the vertex v by X_v, with the root being denoted as 0.

X_0 will be distributed according to an unknown distribution μ, with

$$X_v = X_0 + \sum_{w \in \text{path}(0,v) \setminus \{0\}} \sigma \mathcal{N}_w$$

where $\mathcal{N}_w$ are i.i.d. $N(0, 1)$ random variables, and σ is a fixed parameter.

The goal is to estimate X_0 given $\{X_v : v \in L_\ell\}$, the set of labels at level ℓ. However, this model will not be the one we focus on today.

2. Version 2 (Ornstein-Uhlenbeck)

This is similar to the above model, but instead $X_0 \sim N(0,1)$ and if v is a parent of w then

$$X_w = \eta X_v + \sqrt{1-\eta^2}\mathcal{N}_w$$

where $\mathcal{N}_w$ are i.i.d $N(0,1)$.

An important observation about this model is that $X_w \sim N(0,1)$ for *all nodes w* .

Just like in the previous model, we want to estimate X_0 given $\{X_v : v \in L_\ell\}$. The question is, how should we do it? A natural attempt is to take the average of the children, which we now analyze.

Note that $(X_0, X_v : v \in L_\ell)$ is jointly Gaussian. This means that the estimate that minimizes the ℓ_2 error

$$\mathbb{E}\left[X_0 \middle| (X_v)_{v\in L_\ell}\right]$$

can be written as $\sum_{v\in L_\ell} a_v X_v$, by the definition of being jointly Gaussian. Also, by symmetry of the model, all the a_v's are the same, so we can write this as $a\sum_{v\in L_\ell} X_v$.

So to find the correct estimate for X_0, it suffices to find the value of a that minimizes

$$\mathbb{E}\left[\left(X_0 - a\sum X_v\right)^2\right]$$

which we now proceed to do. We begin by computing the mean squared error

$$\begin{aligned}\mathbb{E}\left[\left(X_0 - a\sum X_v\right)^2\right] &= \mathbb{E}[X_0^2] + a^2\sum \mathbb{E}[X_v^2] - 2ad^\ell \operatorname{Cov}(X_0, X_v) + a^2 \sum_{v \neq w \in L_\ell} \operatorname{Cov}(X_v, X_w) \\ &= 1 + d^\ell a^2 - 2ad^\ell \eta^\ell + a^2 d^\ell \sum_{j=1}^{\ell} \eta^{2j}(d-1)d^{j-1} \\ &= 1 + d^\ell \left(-2a\eta^\ell + \frac{a^2(d-1)}{d} \cdot \frac{(d\eta^2)^{\ell+1} - 1}{(d\eta^2) - 1}\right)\end{aligned}$$

The second line follows from expanding out the sum and the fact that there are d^ℓ nodes at level ℓ. The third line follows from the fact that the covariance of X_v and X_w is dependent on the distance between them in the underlying tree: namely, if the distance between them is D, then $\mathbb{E}[X_v X_w] = \eta^D$. Since all the vertices in the final sum are at the level ℓ, they are at an even distance on the tree (the distance is $2j$, where j denotes the distance between the vertices and their lowest common ancestor on the tree). The fourth line follows from the formula for computing the sum of a geometric series.

We can compute the value of a that minimizes

$$1 + d^\ell \left(-2a\eta^\ell + \frac{a^2(d-1)}{d} \cdot \frac{(d\eta^2)^{\ell+1} - 1}{(d\eta^2) - 1}\right) \tag{1}$$

by taking the derivative with respect to a (this can also be accomplished approximately by setting the two terms in the sum to be equal and solving for a). Doing so, we get that

$$a = \eta^\ell \left(\frac{d-1}{d} \cdot \frac{(d\eta^2)^\ell}{d\eta^2 - 1}\right)^{-1} \tag{2}$$

Now we realize that the optimal squared error $\mathbb{E}\left[\left(X_0 - a\sum X_v\right)^2\right]$ at this value of a, as given by the computation above, behaves very differently when $d\eta^2 < 1$ versus when $d\eta^2 > 1$. When $d\eta^2 < 1$, we notice that $\frac{(d\eta^2)^{\ell+1} - 1}{(d\eta^2) - 1}$ is approximately a constant. Furthermore, the optimal value of a is roughly η^ℓ, and so our expression for the optimal squared error becomes

$$\mathbb{E}\left[\left(X_0 - a\sum X_v\right)^2\right] \approx 1 + d^\ell a(-2\eta^\ell + a\Theta(1)) \approx 1 + d^\ell \eta^{2\ell} \approx 1$$

Note also that a squared error of 1 would be accomplished by simply ignoring the leaves and estimating X_0 by 0.

By contrast, when $d\eta^2 > 1$, we argue that the optimal squared error is less than 1. Plugging in the minimizing value of a gives that

$$\begin{aligned}
\mathbb{E}\left[\left(X_0 - a\sum X_v\right)^2\right] &= 1 + d^\ell\left(-2a\eta^\ell + \frac{a^2(d-1)}{d}\cdot\frac{(d\eta^2)^{\ell+1} - 1}{(d\eta^2) - 1}\right)\\
&= 1 + d^\ell(-2a\eta^\ell + a\eta^\ell)\\
&= 1 - d^\ell\eta^\ell\left(\eta^\ell\frac{d}{d-1}\frac{d\eta^2 - 1}{(d\eta^2)^{\ell+1} - 1}\right)\\
&\leq 1 - \frac{(d\eta^2)^\ell}{(d\eta^2)^{\ell+1}}\cdot\frac{d}{d-1}\cdot(d\eta^2 - 1)\\
&< 1
\end{aligned}$$

In summary, we have a threshold at $d\eta^2 = 1$ for this simple model. If $d\eta^2 > 1$, we have shown that we can estimate better than random with the optimal linear estimator. When $d\eta^2 < 1$, we know that we cannot do better than random. But this model was a linear Gaussian model, which is not at all similar to the type of structure posed by deep learning (specifically, we know that deep learning has *non-linearities* that are applied at each layer, whereas the entire model here was linear). We'll introduce some non-linearities in the next model to try to make it more interesting, and analyze the resulting behaviour.

Model 3: "Threshold" Version of Model 2

In this model we deal with the same d-ary tree as before, but the initial distribution of X_0 and the noise model we use will change.

Specifically, X_0 will be distributed uniformly in $[q] = \{1, \ldots, q\}$ and if w is the parent of v, then

$$X_v = \begin{cases} X_w & \text{w.p. } \eta \\ \text{Unif}[q] & \text{w.p. } 1 - \eta \end{cases}$$

Remark. An alternative view of the model is that each edge in the d-ary tree is erased with probability $1 - \eta$. Then, the root of each component get a random symbol from $[q]$, and within each subtree the character is copied. So the labels within a component are the same, and the labels across components are independent.

Our goal is as follows.

Goal: Estimate X_0 from $(X_v : v \in L_\ell)$

Before continuing, we note that the error of any estimator is always at least $\Theta(1)$ (where we are thinking of d and η as constants), since there is a chance that there is no copying at the first level! But this was also a problem for Model 2.

A First Attempt (Plurality): One naive estimator is to take the value $a \in [q]$ that is most common among the $(X_v : v \in L_\ell)$ (i.e., taking the plurality vote amongst the leaves at level ℓ). This is one of the most natural attempts, but is this the optimal estimator? We now analyse (a variant of) this estimator.

We'll analyse a variant of the plurality estimator. Let

$$y_v = \begin{cases} 1 - 1/q & \text{if } X_v = 1 \\ -1/q & \text{if } X_v \neq 1 \end{cases}$$

The estimator we will use is the following:

1. If $S := \sum_{v \in L_\ell} y_v > 0$, then output 1.
2. Otherwise, output a random character from $\{2, \dots, q\}$.

Claim. *The estimator described above outperforms a random estimator if and only if plurality does.*

We won't formally prove this claim now – it will follow trivially from the results in the rest of lecture. However, intuitively S synchronizes with the plurality estimator when the root symbol is 1, and is random otherwise. If plurality does well, it should do well when the root symbol is 1 by symmetry. It is clear that in this case, S succeeds whenever plurality succeeds, so it outperforms the random estimator. If S works correctly, it means that 1 appeared more than $\frac{1}{q}$ times at the leaves very frequently. The only way plurality could work poorly in this case is if the plurality element is not 1 on many of these occasions. However, by symmetry, every non-root element has lower probability of showing up at the leaves than the root element, so this will only happen if, for many of those times, 1 showed up in just a little more than $\frac{1}{q}$ fraction of the leaves, but was beaten by a non-root symbol due to random deviation. However, intuitively, in this case 1 would not have showed up more than $\frac{1}{q}$ times so often, because random deviation would often make it go below $\frac{1}{q}$. We will indeed see that this is the case in the remainder of lecture.

First we note a trivial claim about when we cannot hope to recover X_0 better than random asymptotically.

Claim. *[Trivial Lower Bound] If $d\eta < 1$ no estimator can do better than random asymptotically.*

Proof. If $d\eta < 1$, then the fraction of leaves that are connected to the root tends to zero. Therefore, no correlations between the leaves and the root will survive. ∎

How do we analyse this estimator? We'll use the following three facts:

$$\mathbb{E}\left[S|X_0 = 1\right] = \mathbb{E}\left[\text{v connected to root}\right] \cdot \left(1 - \frac{1}{q}\right) = (d\eta)^{\ell}\left(1 - \frac{1}{q}\right)$$

$$\mathbb{E}\left[S|X_0 \sim U\right] = 0$$

which follows from the fact that each y_v is an unbiased random variable. Here U denotes the uniform distribution on $[q]$. Also,

$$\text{Var}(S) \approx \Theta(d^{\ell})$$

which follows from a direct computation very similar to the calculation in the previous model, where we minimized the ℓ_2-error.

From the above three observations, we can conclude that there is some $\epsilon > 0$ such that

$$d\eta^2 > 1 \Rightarrow \frac{\mathbb{E}\left[S|X_0 = 1\right] - \mathbb{E}\left[S|X_0 \sim U\right]}{\sqrt{\text{Var}(S)}} > \epsilon$$

and furthermore that

$$d\eta^2 < 1 \Rightarrow \frac{\mathbb{E}\left[S|X_0 = 1\right] - \mathbb{E}\left[S|X_0 \sim U\right]}{\sqrt{\text{Var}(S)}} \to 0$$

As we will see, in the case where $d\eta^2 > 1$, we will be able to recover the root better than random by using the estimator S.

Claim. *[Second Moment Argument]*
If $\frac{\mathbb{E}[S|X_0=1]-\mathbb{E}[S|X_0\sim U]}{\sqrt{\text{Var}(S)}} > \epsilon$, then we can recover X_0 using S better than random.

Proof. Let μ denote the measure of S given a random character at the root, and let μ^+ denote the measure of S given that the root is 1. Furthermore, define f such that

$d\mu^+ = f d\mu$, and of course by definition $\mu = 1 d\mu$. Then

$$\begin{aligned} \left(\mathbb{E}_{\mu^+}[S] - \mathbb{E}_{\mu}[S]\right)^2 &= \left(\int S(d\mu^+ - d\mu)\right)^2 \\ &= \mathbb{E}_{\mu}[S(f-1)]^2 \\ &\le \mathbb{E}_{\mu}\left[S^2\right] \cdot \mathbb{E}_{\mu}\left[(f-1)^2\right] \end{aligned}$$

by Cauchy-Schwarz. Furthermore, we can upper bound the expression $\mathbb{E}_{\mu}\left[(f-1)^2\right]$ by noting that $\mathbb{E}_{\mu}\left[(f-1)^2\right] \le d_{\mathrm{TV}}(\mu, \mu^+)$, which follows since $(\mu(v) - \mu^+(v))^2 \le |\mu(v) - \mu^+(v)|$ for all v. Finally, we can use this to lower bound the total variation distance between μ and μ^+,

$$d_{\mathrm{TV}}(\mu, \mu^+) \ge \frac{\left(\int S(d\mu - d\mu^+)\right)^2}{\mathbb{E}_{\mu}\left[S^2\right]} \ge \epsilon^2$$

where in the final step we used our assumption. ■

In conclusion, this estimator performs very similarly to the scenario for the "boring" case of Model 2! But wait:

1. We didn't prove that the estimator fails if $d\eta^2 < 1$.
2. Maybe there's a better estimator. Maybe even one that works for $d\eta > 1$?

These questions are addressed by the following two theorems.

Theorem.

1. *If $q = 2$, then $d\eta^2 = 1$ is the threshold (and $q = 3, 4$ behave similarly).*
2. *If $q \ge 5$, then $d\eta^2$ is not the threshold (but we have no simple formula for the threshold!)*

In the theorem below, we define *count reconstruction* estimators to be estimators that just depend on the number of each color q among the symbols at the leaves of the tree.

Theorem. *Count reconstruction estimators will always fail for $d\eta^2 < 1$.*

In particular, $d\eta^2 = 1$ is the threshold for all sane count reconstruction estimators, like the one we have already described.

In the interest of time, we won't actually prove both of these theorems. We will sketch some easy versions of a "proof" of the first of these two theorems.

Proof. [“Proof” that $d\eta^2$ is not threshold for $q \geq 5$]

Baby version ($q = \infty$): We first handle this easier special case. When $q = \infty$, the labels in the model can be constructed as follows:

1. Color root randomly (with some arbitrary color).
2. Copy the color with probability η.
3. Otherwise, choose a completely new color that has *not been seen previously* and use it.

Now consider the estimator defined as follows: If two nodes $u, v \in L_\ell$ have the root as their lowest common ancestor and $x_u = x_v$, then return x_v, and otherwise return an arbitrary color.

1. **Correctness:** Any repeated color must have been given by a common ancestor in this model. For u and v as chosen, this ancestor must have been the root!
2. **Probability of correctness:** The probability that we get such leaves is at least the probability that two children of the root survive, and that the branching process survives for each of them.

   $$\mathbb{P}[\text{correct}] \geq \eta^2 \cdot \mathbb{P}[\text{branching process survives}]^2 \geq \epsilon$$

   where we used the fact that the branching process will survive asymptotically if $d\eta > 1$.

How do we extend this to large, finite q? First, we make a definition that will be relevant in our extension. ■

Definition. An *r-diluted d-ary tree* is a tree where every node at level ℓ must have at least d descendents at level $\ell + r$ (if that level exists).

The extension to large, finite q will result from the following two claims:

Claim. *If $d\eta > 1$, there exist $r, \epsilon > 0$ such that with probability at least ϵ, the root cluster contains an infinite r-diluted binary tree.*

Claim. *Given d, r, ϵ, there exists a value $\eta' < 1$ such that, for the branching process with copy probability η', the probability that the root cluster contains an infinite r-diluted $(d^r - 1)$-ary tree is at least $1 - \frac{\epsilon}{10q}$.*

We will delay proving these two claims for now. Instead, we proceed to give a good estimator for the case where q is large and finite, given these claims.

The Estimator: Look for a monochromatic, r-diluted binary tree of color i. If it exists, return color i.

By the first claim above, if the root is colored with i, then the probability that such a tree exists with color i is $\geq \epsilon$. Now suppose that the root is not colored with i. Then we argue that there is no r-diluted binary tree of color i with very high probability. Let v be a child of a node w in the tree. First we note that

$$\begin{aligned}\Pr[x_v \neq i | x_w \neq i] &= \eta + (1 - \eta)\left(1 - \frac{1}{q}\right) \\ &= 1 - \frac{1}{q} + \frac{\eta}{q} \xrightarrow[q\to\infty]{} 1\end{aligned}$$

By taking q large enough such that $1 - \frac{1}{q} + \frac{\eta}{q} \geq \eta'$ where η' is from the second claim above, we see that we cannot fit an r-diluted binary tree of color i in the remainder of the tree (since there is no space for the necessary 2 leaves at any level!).

Now we'll prove the second claim. Before we proceed, we note that this proof was given in detail at the beginning of the proceeding lecture, so the proof here reflects the proof done in that lecture.

Proof. [Proof of Second Claim]
We proceed by induction on k, the number of iterations. Let p_k denote the probability that the Claim holds for a tree of depth $k \cdot r$ (that is, that the root cluster contains an r -diluted $(d^r - 1)$-ary tree of depth $k \cdot r$ after percolation has occured for $k \cdot r$ levels). We will use induction to show that, if η' is large enough, then there is some $p^* \geq 1 - \epsilon$ such that $p_k \geq p^*$ for all k.

Define the function f as follows

$$f(p) = \Pr[\mathrm{Bin}(d^r, p) \geq d^r - 1]$$

We can see that $f(1) = 1$, and furthermore that f is monotone in p. Furthermore, we can compute $f(p)$ by simply counting.

$$f(p) = p^{d^r} + d^r(1-p)p^{d^r-1}$$

Now we compute the derivative of f at $p = 1$:

$$f'(1) = d^r + d^r(1 - p)p^{d^r - 1} = 0$$

From this, we can conclude that for all sufficiently large $p^* < 1$, we have that

$$f(p^*) \geq p^*$$

We choose a $p^* > 1 - \epsilon$ and $\eta' < 1$ such that

$$q = \Pr[\text{connected to } d^r \text{ descendents at distance } r] \geq \frac{p^*}{f(p^*)}$$

Finally, we can proceed to show that $p_k \geq p^*$ for all k by induction. We note that $p_0 = 1$, as the 0-depth tree is just the root. Furthermore

$$\begin{aligned} p_{r+1} &\geq q \cdot \Pr[\text{Bin}(d^\ell, p_r) \geq d^\ell - 1] \\ &\geq \frac{p^*}{f(p^*)} \cdot f(p_r) \\ &\geq \frac{p^*}{f(p^*)} f(p^*) = p^* \end{aligned}$$

where the final inequality proceeds from the induction hypothesis that $p_r \geq p^*$ and the monotonicity of f in p.

■

For the rest of the lecture, we will sketch the proof of the theorem that count reconstruction estimators will always fail for $d\eta^2 < 1$ (the word sketch should be re-emphasized here). Informally, this says that if $d\eta^2 < 1$, then we cannot reconstruct the character at the root better than random simply by looking at the count statistics on the leaves.

Theorem. *[Kesten-Stigum Theorem] If $d\eta^2 < 1$, then the count vector $(c_1^\ell, \ldots, c_q^\ell)$ satisfies a CLT that does not depend on the root, where c_i^ℓ denotes the number of leaves at level ℓ with color i.*

As a concrete example, we can use this CLT for the counts to give a CLT for our earlier count estimator S. Let $\psi(i) = e_1(i) - \frac{(1,\ldots,1)}{q}$. Then we can see that $y_v = \psi(x_v)$, and we can apply the Kesten-Stigum Theorem for $S_\ell = \sum_{v \in L_\ell} y_v$.

We now give a high level sketch of why the Kesten-Stigum theorem works for this sum. The idea is to use the right martingale. We consider exposing one node of the tree at a time, and subtracting the "expected value" of the node given its parent to keep the exposed variables zero mean. More formally, define

$$S'_\ell = \sum_{v:|v|\leq\ell} (d\eta)^{-|v|}(y_v - \eta y_{\text{parent}(v)})$$

where $|v|$ denotes the level of the vertex v in the tree (and $y_{\text{parent}}(\text{root}) = 0$). Note that $\mathbb{E}\left[y_v | y_{\text{parent}(v)}\right] = \eta \cdot y_{\text{parent}(v)}$, so whenever we add a vertex to the sum, the expected value of the sum remains the same. By induction, it follows that when we add an entire level to the sum, it remains the same. Hence, we can conclude that

1. S'_ℓ is a martingale.
2. $S'_\ell = \sum_{v:|v|=\ell} d^{-\ell}\eta^{-\ell} y_v$

where the second item follows by induction and noting that the sum telescopes on adjacent levels. We know that $S_0 = \sum_{v:|v|=0} (d\eta)^0 y_v$, so the claim holds for $\ell = 0$. We now assume that $S'_\ell = \sum_{v:|v|=\ell} d^{-\ell}\eta^{-\ell} y_v$ and prove the statement for $\ell + 1$.

$$\begin{aligned} S'_{\ell+1} &= \sum_{v:|v|\leq\ell+1} (d\eta)^{-|v|}(y_v - \eta y_{\text{parent}(v)}) \\ &= S_\ell + \sum_{v:|v|=\ell+1} (d\eta)^{-\ell-1} y_v - \sum_{v:|v|=\ell+1} (d\eta)^{-\ell-1}\eta y_{\text{parent}(v)} \\ &= S_\ell + \sum_{v:|v|=\ell+1} (d\eta)^{-\ell-1} y_v - \sum_{v:|v|=\ell} (d\eta)^{\ell} y_v \\ &= \sum_{v:|v|=\ell+1} (d\eta)^{-\ell-1} y_v \end{aligned}$$

Now, we would like to appeal to a Martingale Central Limit Theorem to establish that S'_ℓ satisfies a CLT that is independent from the root. Note that Azuma's inequality would tell us that $\lim_{\ell\to\infty} S'_\ell$ is highly concentrated around S_0 if we knew that $\sum_{k=0}^{\infty} c_k^2 < \infty$, where the c_k denote the Martingale differences. In this case, the leaves would *not* satisfy a CLT independent of the root.

$$\begin{aligned} \sum_{\ell' \leq \ell} (S_{\ell'} - S_{\ell'-1})^2 &\geq \sum_{\ell' \leq \ell} d^{\ell'} (d\eta)^{-2\ell'} \Theta(1) \\ &= \Theta(1) \sum_{\ell' \leq \ell} \left(\frac{1}{d\eta^2}\right)^{\ell'} \end{aligned}$$

If $d\eta^2 < 1$, this series diverges, which means that c satisfies a CLT independent of the root. However, this is not quite enough; a CLT tells us a statement of the form $\mathbb{P}[\sum X_i \in [a - c\sqrt{n}, a + c\sqrt{n}]]$, but just knowing that the counts are concentrated in an interval does not rule out the possibility of an estimator that can reconstruct the root from the parity of c_1^ℓ (or similar functions that are not constrained by being in an interval). We note that it does not really make sense for any sane estimator to use the parity when trying to reconstruct the root, but we still have to rule it out! So we will additionally need a Local Central Limit Theorem, of the form

Theorem. *[Local CLT] Let X_i be iid integer valued and not supported on an arithmetic progression with stride ≥ 2. Then we have that*

$$\Pr\left[\sum_{i=1}^{n} X_i = m\right] = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\left(\frac{m - n\mu}{2\sigma\sqrt{n}}\right)^2\right) + o(1/\sqrt{n})$$

where σ^2 is the variance of X_i and μ is the expectation.

The local CLT helps us rule out “parity-like” estimators that still only depend on the counts at the leaves.

elmos / May 29, 2017

Mathematical Aspects of Deep Learning / Proudly powered by WordPress