Lecture 9 : Point matching and reconstruction on trees

Lecturer: Yuval Peres Scribe: *Nathan Levy* <levy@stat.berkeley.edu>

9.1 One color point matching

We proved the following proposition:

Proposition 9.1. *Let R_0 be the distance between the origin and its matched point for a point process conditioned to have a point at the origine.*

- *The stable (i.e. iterated nearest neighbour) matching for a Poisson point process in $\mathbb{R}^d$ satisfies $\mathbb{P}(R_0 > r) \leq \frac{C_d}{r^d}$.*
- *For $d = 1$, any invariant determinist matching for a Poissn point process satisfies $\mathbb{E}R_0 = \infty$.*

For $d > 1$, it is an open problem to know whether there exists a matching with all moments of R_0 finite. The first point of the proposition assures that $\mathbb{E}R_0^{d-1} < \infty$ but we do not know if $\mathbb{E}R_0^d < \infty$ is possible.

Proposition 9.2. *The stable matching in $\mathbb{R}^d$ yields $\mathbb{E}R_0^d = \infty$.*

Proof. For a center x (i.e. a point of the point process), let R_x denote the distance between x and its matched point. Draw a ball $B(x, R_x)$ around every center x.

Step 1 : Let us show that those balls "cover" $\mathbb{R}^d$, i.e. that the uncovered set has volume 0 a.s.

Otherwise, using the fact that you can cover $\mathbb{R}^d$ with countably many copies of $[0, 1]^d$:

$$\mathbb{P}\big(\mathcal{L}eb(\text{uncovered set in } [0, 1]^d) > 0\big) > 0.$$

Consider now a different point process equivalent (i.e. mutually absolutely continuous) to the Poisson point process: take a Poisson point process and, with probability $1/2$, add a center uniformly in $[0, 1]^d$. With positive probability, the added center falls in the uncover set. In this case, that center is matched to no one because every other center is matched to a center in its ball hence closer. We have an unmatched center with positive probability for the new point process, that and the equivalence of the two processes is in contradiction with the fact that every center is matched a.s. for the Poisson point process.

Step 2 : Let us show that almost all $\mathbb{R}^d$ is covered by infinitely many balls.

Let

$$A_L = \{y \in \mathbb{R}^d : \text{all centers } x \text{ s.t. } y \in B(x, R_x) \text{ satisfy } \overline{B(x, R_x)} \subset [-L, L]^d\}.$$

Then

$$\bigcup_{L=1}^{\infty} A_L = \{\text{y} : \text{y covered by finitely many } B(x, R_x)\}.$$

Let us prove that $\forall L \ \mathbb{E}\mathcal{L}eb(A_L) = 0$. Otherwise, one can now consider an equivalent point process where one erases all the centers in $[-L, L]^d$ and their matched centers with probability 1/2. That process is equivalent to the Poisson point process and the remaining centers are matched in the same way. Nevertheless, with positive probability A_L is an uncovered area of positive Lebesgue measure, which is absurd by step 1.

Step 3 : Mass transport principle.

Consider the following mass transport function : for $y, z \in \mathbb{Z}^d$

$$m(y,z) = \mathbb{E} \sum_{\text{centers } x \in y + [-1/2;1/2]^d} \mathcal{L}eb\big(B(x, R_x) \cap z + [-1/2; 1/2]^d\big).$$

Then

$$\sum_y m(y,z) = \mathbb{E} \sum_{x \text{ center}} \mathcal{L}eb\big(B(x, R_x) \cap z + [-1/2; 1/2]^d\big) = \infty \text{ by step 1}$$

hence,

$$\infty = \sum_z m(y,z) = \mathbb{E} \sum_{\text{centers } x \in y + [-1/2;1/2]^d} C_d R_x^d = \mathbb{E} R_0^d.$$

For that last equality, we used the fact that $\mathbb{E}\big(\#\text{centers } x \in y + [-1/2; 1/2]^d\big) = 1$. □

9.2 Reconstruction on trees

Here is an overview of the simplest model. Consider a b-ary tree with root o and choose a parameter $\epsilon \leq 1/2$. We assign to every vertex v a spin $\sigma_v = \pm 1$ by the following random method. Let $\sigma_o = \pm 1$ with probability 1/2 each. Then, if w is the father of v let $\sigma_v = \sigma_w$ with probability $1 - \epsilon$ and $\sigma_v = -\sigma_w$ with probability ϵ. The idea is that the spin is broadcasted from the root downward and is changed with an error probability ϵ on every edges.

Question : Consider the best predictor of σ_o knowing $\{\sigma_v, |v| = n\}$ and set:

$$\Delta_n = \mathbb{P}(\text{predictor right}) - \mathbb{P}(\text{predictor wrong}).$$

Note that the best predictor is the one that maximizes Δ_n. We ask for which ϵ we have $\lim_n \Delta_n > 0$, in that case, we say that the reconstruction is possible. It means that by looking at the vertices of the n^{th} generation, you have asymptotic information on the root. That is equivalent to the fact that the sequence $(\{\sigma_v, |v| = n\})_{n=1}^{\infty}$ has a non trivial tail σ-field.

Applications : This problem appeared simultaneously in different areas.

- Statistical physics : This model is a different description of a known problem : Ising model on tree.
- Phylogenetic : Is it possible to guess some characteristics of an ancestor looking at the characteristics of many descendents at the n^{th} generation ? The problems considered in phylogenetic are much more complicate but the methods developed for the reconstruction on trees can still be applied.
- Computer science : Our model is a good description of the broadcasting of information along a noisy channel.

We next stat and prove an easy proposition that gives a necessary condition for reconstruction.

Proposition 9.3. *If $\theta = 1 - 2\epsilon \leq \frac{1}{b}$ then reconstruction is impossible (i.e. $\Delta_n \xrightarrow[n\to\infty]{} 0$).*

Proof. We construct a coupling between the probability distribution on the configurations conditoned to have +1 at the root (call σ the configuration) and the distribution on the configuration conditioned to have -1 at the root (call τ the configuration). We set $\sigma_o = +1$ and $\tau_o = -1$. Then, if w is the father of v, we have two possibilities :

First case : $\sigma_w = \tau_w$ then with probability $1 - \epsilon$, set $\sigma_v = \tau_v = \sigma_w$ and with probability ϵ set $\sigma_v = \tau_v = -\sigma_w$, do that independently of what has been done so far.

Second case : $\sigma_w = +1$ and $\tau_w = -1$, pick a uniform random variable U in $[0, 1]$ independently of what has been done so far.

$$\begin{cases} \text{if } U < \epsilon & \text{then set } \sigma_v = \tau_v = +1 \\ \text{if } \epsilon \leq U < 1-\epsilon & \text{then set } \sigma_v = +1 \text{ and } \tau_v = -1 \\ \text{if } 1-\epsilon \leq U & \text{then set } \sigma_v = \tau_v = -1. \end{cases}$$

Basically: if the configurations agree on the father then they agree on all the offspring forever and if the configurations disagree on the father then they agree on the son with probability 2ϵ and they keep disagreeing with probability $1 - 2\epsilon$. That way we see inductively that the case $\sigma_w = -1$ and $\tau_w = +1$ is impossible.

Now look at the vertices on which the configurations disagree ; they form a Galton Watson tree with offspring expectation $b(1 - 2\epsilon)$. Therefore, if $b(1 - 2\epsilon) \leq 1$ that tree is a.s. finite and there is a level from which the configurations always agree. If you look at the vertices below that level, you cannot tell the two configurations apart, hence reconstruction is impossible. □

That proposition is not sharp, $1 - 2\epsilon > \frac{1}{b}$ is not a sufficient condition for reconstruction.

Proposition 9.4.

$$b(1-2\epsilon)^2 \leq 1 \Leftrightarrow \Delta_n \xrightarrow[n\to\infty]{} 0.$$

Ramae (1997) gave the first argument showing that reconstruction is possible if $b(1 - 2\epsilon)^2 > 1$. He considered $S_n = \sum_{|v|=n} \sigma_v$ and the predictor $\widehat{\sigma}_0 = \text{sign}(S_n)$, which gives reconstruction.

We now give another argument showing that for $1 - 2\epsilon \leq 1/b$ reconstruction is impossible.

Here is another way of describing the process : if w is the father of v, set $\sigma_v = \sigma_w$ with probability $1 - 2\epsilon$ and $\sigma_v = \mathcal{R}$ with probability 2ϵ where $\mathcal{R}$ is a random variable taking the values ± 1 with probability 1/2 each. Basically, the information is broadcasted without mistake with probability $1 - 2\epsilon$ and with probability 2ϵ a mutation appears and the spin of the son is then random.

Once a $\mathcal{R}$ appears, below it, the configuration is independent of the spin of the root. Consider now the subtree where no $\mathcal{R}$ has ever appeared i.e. where the information from the root has been well broadcasted. It is Galton Watson Tree with offspring expectation $(1 - 2\epsilon)b \leq 1$ hence it is a.s. finite and there is a generation from which the spins are independent of the root and so no information is broadcasted anymore.

Question : Why is that proof not sharp ?

In the last description, if one has $\frac{1}{b} < (1 - 2\epsilon) < \frac{1}{\sqrt{b}}$, the Galton Watson Tree can be infinite but by looking at the n^{th} generation, one can not tell apart the +1 that descend from a $\mathcal{R}$ and therefore are independent of the root from the those who descend directly from the root (i.e. there is no $\mathcal{R}$

on the path from the vertex to root). Heuristically, the last ones are swallowed by the noise created by the $\mathcal{R}$'s. The order of that noise is given by the CLT.