Distributed Density Estimation

1 Catalan Objects

1.1 Dyck Paths

Dyck paths are one interpretation of the Catalan numbers. Here, we will instead consider a more general form of Dyck Paths, which correspond to numbers in the *Catalan Trapezoid*.

A Dyck path can be constructed as a $1D$ random walk with $2n$ steps, where the ends of the walk are pinned to zero (Figure **??**). This path also has the additional restriction that the walk can never reach negative values. The number of possible Dyck paths is the n^{th} Catalan number –

$$C_n = \frac{1}{n+1} \cdot \binom{2n}{n}$$

We will attempt to support queries to a uniformly random instance of a Dyck path. Specifically, we will want to query the position of the i^{th} index.

1.2 Catalan Trapezoid

First, we define Catalan trapezoids as presented in [Reu14]. Let $C_k(n, m)$ be the $(n, m)^{th}$ entry of the Catalan trapezoid of order k, where $C_1(n, m)$ corresponds to the Catalan triangle.

The interpretation is as follows. Consider a sequence of n $(+1)$s and m (-1), such that the sum of any initial sub-string is not less than $1 - k$. This means that we start our Dyck path at a height of $k - 1$, and we are never allowed to cross below zero. The total number of such paths is exactly $C_k(n, m)$.

Now, we state a result from [Reu14] without proof

$$C_k(n,m) = \begin{cases} \binom{n+m}{m} & 0 \le m < k \\ \binom{n+m}{m} - \binom{n+m}{m-k} & k \le m < n+k-1 \\ 0 & m > n+k-1 \end{cases}$$

1.3 Generating Dyck Paths

Our general recursive step is as follows. We consider a sequence of length $2S$ comprising of $2U$ up moves $(+1)$ and $2D$ down moves (-1). Additionally, the sum of any initial sequence prefix? canon be less than $k-1$. Without loss of generality, let's assume that $2D \leq S$. If this were not the case, we could simply flip the sequence and negate the elements. This essentially means that the overall Dyck path is non-decreasing.

Lemma 1. $S - 2D = \mathcal{O}(\log n\sqrt{S}) \implies U - D = \mathcal{O}(\log n\sqrt{S})$

We want to sample the height of this path after S steps. This is the same as sampling the number of $(+1)$s that get assigned to the first half of the elements in the sequence. We define p_d as the probability that exactly $D-d$ (-1)s get assigned to the first half. This means that exactly $U+d$ $(+1)$s get assigned to the first half. Consequently, the second half will contain exactly $D+d$ (-1)s and $U-d$ $(+1)$s.

Let us first compute this probability.

$$p_d = \frac{D_{left} \cdot D_{right}}{D_{tot}}$$

Where D_{left} denotes the number of valid starting sequences (first half) and D_{right} denotes the number of valid ending sequences. Here, *valid* means that each half sequence gets the appropriate number of ups and downs and the initial sums never drop below $1-k$. For, D_{right}, we will start the Dyck path from the end of the $2S$ sequence. In this case the invalidation threshold will be a different k'. This k' is the final height of the $2S$ sequence. So, $k' = k + 2U - 2D = k + 4S - 2D$. We will use this fact extensively moving forward.

Also, D_{tot} is the total number of possible sequences of length $2S$, given the initial conditions. This value is considered by constructing paths in the original direction i.e. the value of k is the same.

1.4 The Simple Case

The problem of sampling reduces to the binomial sampling case when $k > \mathcal{O}(\log n)\sqrt{S}$ for some constant c. In this case, the we can simply approximate the probability as

$$\frac{\binom{S}{D-d} \cdot \binom{S}{D+d}}{\binom{2S}{2D}}$$

This is because the random paths will not have initial sums less than $1-k$ with high probability. Note that this uses the assumption that we have an increasing path.

1.5 Path Segments Close to Zero

The problem arises when we $k < \mathcal{O}(\log n)\sqrt{S}$. In this case we need to compute the actual probability, Using the formula from [Reu14], we find that.

$$D_{left} = \binom{S}{D-d} - \binom{S}{D-d-k} \qquad D_{right} = \binom{S}{U-d} - \binom{S}{U-d-k'} \tag{1}$$

Here, $k' = k + 2U - 2D$, and so $k' = \mathcal{O}(\log n)\sqrt{S}$.

prove using Lemma 1

Finally, we compute the total number of Dyck paths as

$$D_{tot} = \binom{2S}{2D} - \binom{2S}{2D-k}$$

Now, we are going to use the following Lemma from [GGN10].

Lemma 2. *Let $\{p_i\}$ and $\{q_i\}$ be distributions satisfying the following conditions*

1. *There is a poly-time algorithm to approximate p_i and q_i up to $\pm n^{-2}$*
2. *Generating an index i according to q_i is closely implementable.*
3. *There exists a poly(logn)-time recognizable set S such that*
   - $1 - \sum_{i \in S} p_i$ *is negligible*
   - *There exists a constant c such that for every i, it holds that $p_i \leq \log^{\mathcal{O}(1)} n \cdot q_i$*

Then, generating an index i according to the distribution $\{p_i\}$ is closely-implementable.

In this process, we will first disregard all values of d where $|d| > \Theta(\sqrt{S})$. The probability mass associated with these values can be shown to be negligible .

bound variance of path

Next, we will construct an appropriate $\{q_i\}$ and show that $p_d < \log^{\mathcal{O}(1)} n \cdot q_d$ for all $|d| < \Theta(\sqrt{S})$ and some constant c. We will use the following distribution

$$q_d = \frac{\binom{S}{D-d} \cdot \binom{S}{D+d}}{\binom{2S}{2D}} = \frac{\binom{S}{D-d} \cdot \binom{S}{U-d}}{\binom{2S}{2D}}$$

It is shown in [GGN10] that this distribution is closely implementable.

Lemma 3. *First we show that $D_{left} \leq \frac{c_1 \cdot k}{\sqrt{S}} \cdot \binom{S}{D-d}$ for some constant c_1.*

Proof. This involves some simple manipulations.

$$D_{left} = \binom{S}{D-d} - \binom{S}{D-d-k} \tag{2}$$

$$= \binom{S}{D-d} \cdot \left[1 - \frac{(D-d)(D-d-1)\cdots(D-d-k+1)}{(S-D-d+k)(S-D-d+k-1)\cdots(S-D-d+1)}\right] \tag{3}$$

$$\leq \binom{S}{D-d} \cdot \left[1 - \left(\frac{D-d-k+1}{S-D+d+k}\right)^k\right] \tag{4}$$

$$\leq \binom{S}{D-d} \cdot \left[1 - \left(\frac{U+d+k-(U-D+d+k-1)}{U+d+k}\right)^k\right] \tag{5}$$

$$\leq \binom{S}{D-d} \cdot \left[1 - \left(\frac{U+d+k-\mathcal{O}(\sqrt{U})}{U+d+k}\right)^k\right] \tag{6}$$

$$\leq \frac{k}{\Theta(\sqrt{S})} \cdot \binom{S}{D-d} \tag{7}$$

□

Lemma 4. *Similarly, we show that* $D_{right} < \frac{c_2 \cdot k'}{\sqrt{S}} \cdot \binom{S}{U-d}$ *for some constant* c_2*.*

Proof.

$$D_{right} = \binom{S}{U-d} - \binom{S}{U-d-k'} \tag{8}$$

$$= \binom{S}{U-d} \cdot \left[1 - \frac{(U-d)(U-d-1)\cdots(U-d-k'+1)}{(S-U-d+k')(S-U-d+k'-1)\cdots(S-U-d+1)}\right] \tag{9}$$

$$\leq \binom{S}{U-d} \cdot \left[1 - \left(\frac{U-d-k'+1}{S-U+d+k'}\right)^{k'}\right] \tag{10}$$

$$\leq \binom{S}{U-d} \cdot \left[1 - \left(\frac{2D-U-d-k+1}{2U-D+k+d}\right)^{k'}\right] \tag{11}$$

$$\leq \binom{S}{U-d} \cdot \left[1 - \left(\frac{U+k+d-(2U-2D+2d+2k-1)}{U+k+d}\right)^{k'}\right] \tag{12}$$

$$\leq \binom{S}{U-d} \cdot \left[1 - \left(\frac{U+k+d-\mathcal{O}(\sqrt{U})}{U+k+d}\right)^{k'}\right] \tag{13}$$

$$\leq \frac{k'}{\Theta(\sqrt{S})} \cdot \binom{S}{U-d} \tag{14}$$

□

Finally, we need to lower bound the value of D_{tot}.

Lemma 5. *We claim that* $D_{tot} < \frac{c_3 \cdot k \cdot k'}{S} \cdot \binom{2S}{2D}$ *for some constant* c_3.

Proof.

$$D_{tot} = \binom{2S}{2D} - \binom{2S}{2D-k} \tag{15}$$

$$= \binom{2S}{2D} \cdot \left[1 - \frac{(2D)(2D-1)\cdots(2D-k+1)}{(2S-2D+k)(2S-2D+k-1)\cdots(2S-2D+1)}\right] \tag{16}$$

$$\geq \binom{2S}{2D} \cdot \left[1 - \left(\frac{2D-k+1}{2S-2D+1}\right)^k\right] \tag{17}$$

$$\geq \binom{2S}{2D} \cdot \left[1 - \left(\frac{2U-(2U-2D+k-1)}{2U+1}\right)^k\right] \tag{18}$$

$$\geq \binom{2S}{2D} \cdot \left[1 - \left(\frac{(2U+1)-k'}{2U+1}\right)^k\right] \tag{19}$$

$$\geq \frac{k \cdot k'}{\Theta(S)} \cdot \binom{2S}{2D} \tag{20}$$

□

Theorem 1. *We can now put these lemmas together to show that* $p_d/q_d \leq c = c_1 \cdot c_2/c_3 = \Theta(1)$. *This satisfies all the conditions of Lemma 2 from [GGN10]. We simply need to set the accept probability less than* $p_d/(c \cdot q_d)$.

References

[GGN10] Oded Goldreich, Shafi Goldwasser, and Asaf Nussboim. On the implementation of huge random objects. *SIAM Journal on Computing*, 39(7):2761–2822, 2010.

[Reu14] Shlomi Reuveni. Catalan's trapezoids. *Probability in the Engineering and Informational Sciences*, 28(03):353–361, 2014.