The Total Area Statistic for Dyck Paths

James Harbour

April 7, 2024

Contents

1 Preliminaries

$$c_n = \frac{1}{n+1}\binom{2n}{n}, \quad C = C(x) = \sum_{n\geq 0} c_n x^n = 1 + xC^2 = \frac{1-\sqrt{1-4x}}{2x}.$$

$$D_n := \{\text{Dyck paths from } 0 \text{ to } (n,n)\}, \quad D_0 := \{\{(0,0)\}\}.$$

Given $\gamma \in D_n$, let $\operatorname{area}(\gamma)$ denote the area between γ and the line $y = x$.

2 Counting Dyck Paths

3 Computing the Total Area

We follow the approaches in [CEF07] and [MSV96]

Theorem 3.0.1. *Let A_n be the total area of all of the c_n Dyck paths of length n. Then*

$$A_n = \frac{1}{2}\left(4^n - \binom{2n+1}{n}\right)$$

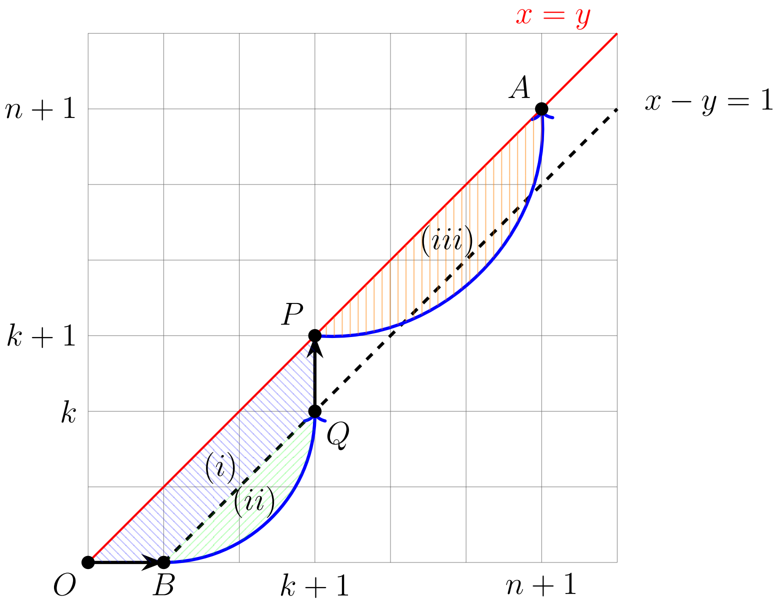

Figure 1: Recursive decomposition of γ

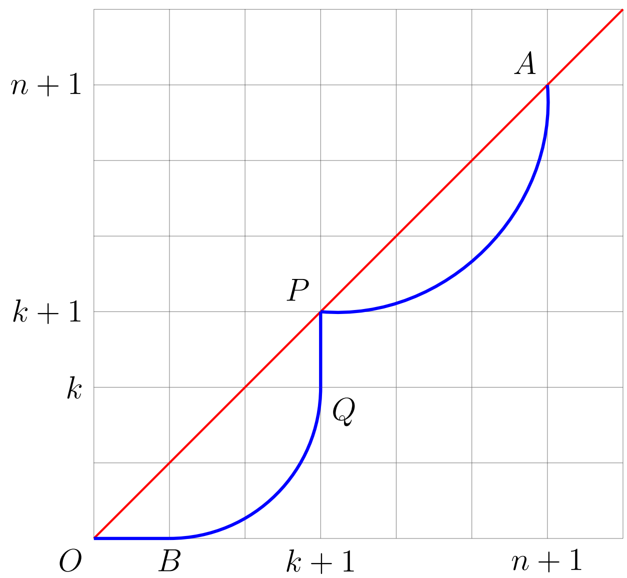

Figure 2: The whole path $\gamma \in D_{n+1}$

Proof. First we find a recursive formula for A_n. Note that $A_0 = 0$. If $P \in \{(1,1), \ldots, (n+1, n+1)\}$, let

$$D_{n+1}^P := \{\gamma \in D_{n+1} : \text{the first point on } y = x \text{ in } \gamma \text{ after } (0,0) \text{ is } P\}.$$

Then we have a decomposition

$$D_{n+1} = \bigsqcup_{k=0}^{n} D_{n+1}^{(k+1,k+1)} \implies A_{n+1} = \sum_{\gamma \in D_{n+1}} \operatorname{area}(\gamma) = \sum_{k=0}^{n} \sum_{\gamma \in D_{n+1}^{(k+1,k+1)}} \operatorname{area}(\gamma)$$

Fix $k \in \{0, \ldots, n\}$, $P := (k+1, k+1)$, and let $\gamma \in D_{n+1}^P$.

Define points $O = (0,0)$, $B = (1,0)$, and $A = (n+1, n+1)$. Since $n+1 > 0$, the first step of γ is OB. Now note that γ passes through $Q = (k+1, k)$ at or after B by definition. Hence we have a depiction of γ in Figure 2. We break up the area between γ and the line $y = x$ according to Figure 1 as follows.

1. The trapezoid $OBQP$ labelled (i) has area $k + \frac{1}{2}$.
2. The green area labelled (ii) has area $\operatorname{area}(\gamma|_{BQ} - B)$, where we note that $\gamma|_{BQ} - B \in D_k$.
3. The orange area labelled (iii) has area $\operatorname{area}(\gamma|_{PA} - P)$, where we note that $\gamma|_{PA} - P \in D_{n-k}$.

Hence, the area of the individual γ decomposes as

$$\operatorname{area}(\gamma) = k + \frac{1}{2} + \operatorname{area}(\gamma|_{BQ} - B) + \operatorname{area}(\gamma|_{PA} - P),$$

whence we obtain the a decomposition for A_{n+1} by

$$A_{n+1} = \overbrace{\sum_{k=0}^{n} \sum_{\gamma \in D_{n+1}^{(k+1,k+1)}} k + \frac{1}{2}}^{\text{sum from } (i)} + \overbrace{\sum_{k=0}^{n} \sum_{\gamma \in D_{n+1}^{(k+1,k+1)}} \operatorname{area}(\gamma|_{BQ} - (1,0))}^{\text{sum from } (ii)} + \overbrace{\sum_{k=0}^{n} \sum_{\gamma \in D_{n+1}^{(k+1,k+1)}} \operatorname{area}(\gamma|_{PA} - P)}^{\text{sum from (iii)}}.$$

Observe that, for fixed $P = (k+1, k+1)$, under the decomposition in Figure 1 the curves $\gamma|_{BQ}$ and $\gamma|_{PA}$ after translating to the origin area freely ranging over D_k and D_{n-k} respectively. Noting that any $\gamma \in D_{n+1}^P$ may be written as

$$\gamma = \overline{OB} \dotplus \gamma|_{BQ} \dotplus \overline{QP} \dotplus \gamma|_{PA}$$

where $\dotplus$ denotes path concatenation, we have a set decomposition

$$D_{n+1}^P = \{\overline{OB} \dotplus (\alpha + B) \dotplus \overline{QP} \dotplus (\beta + P) : \alpha \in D_k, \beta \in D_{n-k}\}.$$

Moreover, the map $\gamma \mapsto (\gamma|_{BQ} - B, \gamma|_{PA} - P)$ furnishes a bijection between D_{n+1}^P and $D_k \times D_{n-k}$.

Treating the sum from (i), we directly compute

$$\begin{aligned}
\sum_{k=0}^{n} \sum_{\gamma \in D_{n+1}^{(k+1,k+1)}} k + \frac{1}{2} &= \sum_{k=0}^{n} \left(k + \frac{1}{2}\right) |D_{n+1}^{(k+1,\,k+1)}| \\
&= \sum_{k=0}^{n} \left(k + \frac{1}{2}\right) |D_k \times D_{n-k}| = \sum_{k=0}^{n} \left(k + \frac{1}{2}\right) c_k c_{n-k}
\end{aligned}$$

Treating the sum from (ii), we find

$$\begin{aligned}
\sum_{k=0}^{n} \sum_{\gamma \in D_{n+1}^{(k+1,k+1)}} \operatorname{area}(\gamma|_{BQ} - B) &= \sum_{k=0}^{n} \sum_{(\alpha,\beta) \in D_k \times D_{n-k}} \operatorname{area}(\alpha) = \sum_{k=0}^{n} \sum_{\alpha \in D_k} \sum_{\beta \in D_{n-k}} \operatorname{area}(\alpha) \\
&= \sum_{k=0}^{n} \sum_{\alpha \in D_k} c_{n-k} \operatorname{area}(\alpha) = \sum_{k=0}^{n} c_{n-k} A_k
\end{aligned}$$

Treating the sum from (iii), we similarly compute

$$\begin{aligned}
\sum_{k=0}^{n} \sum_{\gamma \in D_{n+1}^{(k+1,k+1)}} \operatorname{area}(\gamma|_{PA} - P) &= \sum_{k=0}^{n} \sum_{(\alpha,\beta) \in D_k \times D_{n-k}} \operatorname{area}(\beta) = \sum_{k=0}^{n} \sum_{\beta \in D_{n-k}} \sum_{\alpha \in D_k} \operatorname{area}(\beta) \\
&= \sum_{k=0}^{n} \sum_{\beta \in D_{n-k}} c_k \operatorname{area}(\beta) = \sum_{k=0}^{n} c_k A_{n-k}
\end{aligned}$$

Combining these three results, we obtain a recursive formula

$$\begin{aligned}
A_{n+1} &= \sum_{k=0}^{n} \left(k + \frac{1}{2}\right) c_k c_{n-k} + \sum_{k=0}^{n} c_{n-k} A_k + \sum_{k=0}^{n} c_k A_{n-k} \\
&= \frac{1}{2} \sum_{k=0} c_k c_{n-k} + \sum_{k=0}^{n} k c_k c_{n-k} + 2 \sum_{k=0}^{n} c_{n-k} A_k \qquad (1)
\end{aligned}$$

We will utilize generating function manipulations to find an explicit formula for A_n. To set notation, given a formal power series $S = S(x)$, we denote the coefficient of x^n in S by $[x^n]\{S\}$. Consider the generating functions

$$C = C(x) = \sum_{n=0}^{\infty} c_n x^n, \quad A = A(x) = \sum_{n=0}^{\infty} A_n x^n.$$

In terms of these generating functions, equation (1) becomes a relation between n^{th} coefficients by

$$\begin{aligned}
[x^n]\left\{\frac{A(x)}{x}\right\} = A_{n+1} &= \frac{1}{2}\sum_{k=0}^{n} c_k c_{n-k} + \sum_{k=0}^{n} k c_k c_{n-k} + 2\sum_{k=0}^{n} c_{n-k} A_k \\
&= \frac{1}{2}[x^n]\{C(x)^2\} + [x^n]\{xC(x)'C(x)\} + 2[x^n]\{C(x)A(x)\} \\
&= [x^n]\left\{\frac{1}{2}C(x)^2 + xC(x)'C(x) + 2C(x)A(x)\right\}.
\end{aligned}$$

Hence, we have an equality of generating functions

$$\frac{A}{x} = \frac{1}{2}C^2 + xC'C + 2CA. \tag{2}$$

We intend to solve this equation for A. First, we note some useful equalities:

$$C = 1 + xC^2 = \frac{1-\sqrt{1-4x}}{2x}, \quad C' = \frac{C^2}{\sqrt{1-4x}}.$$

$$\begin{aligned}
A = \frac{1}{2}xC^2 + x^2C'C + 2xCA &\implies A(1-2xC) = \frac{1}{2}xC^2 + x^2C'C \\
&\implies A = \frac{1}{1-2xC}\left(\frac{1}{2}xC^2 + x^2C'C\right) = \frac{1}{\sqrt{1-4x}}\left(\frac{1}{2}xC^2 + x^2C'C\right)
\end{aligned} \tag{3}$$

Expanding the inner expression, we compute

$$\frac{1}{2}xC^2 + x^2C'C = \frac{1}{2}(C-1) + \frac{x}{\sqrt{1-4x}}C(C-1) \tag{4}$$

As an intermediate computation, we note

$$C - 1 = \frac{1-\sqrt{1-4x}-2x}{2x} \implies C(C-1) = \frac{1-3x-\sqrt{1-4x}+x\sqrt{1-4x}}{2x^2}.$$

Returning to equation (4), we find

$$\begin{aligned}
\frac{1}{2}(C-1) + \frac{x}{\sqrt{1-4x}}C(C-1) &= \frac{1}{2}\cdot\frac{1-\sqrt{1-4x}-2x}{2x} + \frac{x}{\sqrt{1-4x}}\frac{1-3x-\sqrt{1-4x}+x\sqrt{1-4x}}{2x^2} \\
&= \frac{1-\sqrt{1-4x}-2x}{4x} + \frac{1-3x-\sqrt{1-4x}+x\sqrt{1-4x}}{2x\sqrt{1-4x}}
\end{aligned}$$

Finally, substituting back into equation (3), we write

$$\begin{aligned}
A = \frac{1}{\sqrt{1-4x}}\left(\frac{1}{2}xC^2 + x^2C'C\right) &= \frac{1}{\sqrt{1-4x}}\left(\frac{1-\sqrt{1-4x}-2x}{4x} + \frac{1-3x-\sqrt{1-4x}+x\sqrt{1-4x}}{2x\sqrt{1-4x}}\right) \\
&= \frac{1-\sqrt{1-4x}-2x}{4x\sqrt{1-4x}} + \frac{1-3x-\sqrt{1-4x}+x\sqrt{1-4x}}{2x(1-4x)} \\
&= \frac{\sqrt{1-4x}-(1-4x)-2x\sqrt{1-4x}}{4x(1-4x)} + \frac{2-6x-2\sqrt{1-4x}+2x\sqrt{1-4x}}{4x(1-4x)} \\
&= \frac{1-2x-\sqrt{1-4x}}{4x(1-4x)}
\end{aligned}$$

Lastly, with this expression we compute the n^{th} coefficient of A as

$$\begin{aligned}
[x^n]\{A\} &= [x^n]\left\{\frac{1}{4x(1-4x)}\right\} + [x^n]\left\{\frac{-2x}{4x(1-4x)}\right\} + [x^n]\left\{\frac{-\sqrt{1-4x}}{4x(1-4x)}\right\} \\
&= \frac{1}{4}[x^{n+1}]\left\{\frac{1}{1-4x}\right\} - \frac{1}{2}[x^n]\left\{\frac{1}{1-4x}\right\} - \frac{1}{4}[x^{n+1}]\left\{\frac{1}{\sqrt{1-4x}}\right\} \\
&= \frac{1}{4}[x^{n+1}]\left\{\sum_{n=0}^{\infty} 4^n x^n\right\} - \frac{1}{2}[x^n]\left\{\sum_{n=0}^{\infty} 4^n x^n\right\} - \frac{1}{4}[x^{n+1}]\left\{\sum_{n=0}^{\infty} \binom{2n}{n} x^n\right\} \\
&= \frac{1}{4}4^{n+1} - \frac{1}{2}4^n - \frac{1}{4}\binom{2(n+1)}{n+1} = \frac{4^n}{2} - \frac{1}{2}\binom{2n+1}{n}
\end{aligned}$$

$\square$

4 Area Averages and Asymptotic info

So we know the total area

$$A_n = \frac{1}{2}\left(4^n - \binom{2n+1}{n}\right)$$

Let $\mathbb{P}$ denote the uniform probability measure on D_n. Let $X_n : D_n \to [0, n^2]$ be given by $X_n(\gamma) = \text{area}(\gamma)$. Using Stirling's approximation, we find

$$\begin{aligned}
\mathbb{E}[X_n] = \frac{1}{c_n}A_n &= \frac{1}{2} \cdot \frac{n+1}{\binom{2n}{n}}\left(4^n - \binom{2n+1}{n}\right) \\
&= \frac{1}{2}\left(\frac{(n+1)4^n}{\binom{2n}{n}} - (2n+1)\right) \\
&\sim \frac{1}{2}\left(\frac{(n+1)4^n}{\frac{2^{2n}}{\sqrt{\pi n}}} - (2n+1)\right) \\
&= \frac{1}{2}\left(\sqrt{\pi n}(n+1) - (2n+1)\right) \sim \frac{\sqrt{\pi}}{2}n^{3/2} \asymp n^{3/2}
\end{aligned}$$

By Chebyshev's inequality

$$\mathbb{P}\left(X_n \geq cn^{\frac{3}{2}+\varepsilon}\right) \leq \frac{\mathbb{E}[X_n]}{cn^{\frac{3}{2}+\varepsilon}} \lesssim \frac{n^{\frac{3}{2}}}{cn^{\frac{3}{2}+\varepsilon}} = \frac{1}{cn^{\varepsilon}} \xrightarrow{n\to\infty} 0$$

A more detailed look on this asymptotic using Stirling's approximation gives

$$\begin{aligned}
\#\{\gamma \in D_n : \text{area}(\gamma) \geq cn^{\frac{3}{2}+\varepsilon}\} = c_n\,\mathbb{P}\left(X_n \geq cn^{\frac{3}{2}+\varepsilon}\right) &\lesssim \frac{1}{cn^{\frac{3}{2}+\varepsilon}}\left(4^n - \binom{2n+1}{n}\right) \\
&= \frac{1}{cn^{\frac{3}{2}+\varepsilon}}\left(4^n - \frac{1}{2}\binom{2(n+1)}{n+1}\right) \lesssim \frac{1}{cn^{\frac{3}{2}+\varepsilon}}\left(4^n - K\frac{4^{n+1}}{\sqrt{\pi(n+1)}}\right) \\
&\lesssim \frac{4^n}{n^{\frac{3}{2}+\varepsilon}}\left(1 - \frac{4K}{\sqrt{\pi(n+1)}}\right) = O\left(\frac{4^n}{n^{\frac{3}{2}+\varepsilon}}\right).
\end{aligned}$$

4.1 Large Deviations

In this section, we show that in fact the tail count of paths with area at least $cn^{\frac{3}{2}+\varepsilon}$ is significantly smaller than the above approximation makes it appear. We peruse into large deviations theory for this application. To ease our computations, we instead use the x-axis based model of Dyck paths. To translate from our previous model requires a -45 degree rotation and then scaling by $\sqrt{2}$, whence all of our area-based results are scaled by 2.

Observe first that if P is a catalan path of length $2n$ whose maximum height is less than or equal to $cn^{\frac{1}{2}+\varepsilon}$, then $\text{area}(P) \leq cn^{\frac{3}{2}}$ by a symmetry argument. Hence, by contraposition

$$\begin{aligned}\{\text{Catalan paths } P : \text{area}(P) > cn^{\frac{3}{2}+\varepsilon}\} &\subseteq \{\text{Catalan paths } P : \max_{k\leq 2n} P_k > cn^{\frac{1}{2}+\varepsilon}\} \\ &\subseteq \{\text{all length } 2n \text{ simple random walks } W : \max_{k\leq 2n} W_k > cn^{\frac{1}{2}+\varepsilon}\}\end{aligned}$$

Let $\{S_k\}_{k=0}^{\infty}$ be simple random walk in $\mathbb{Z}$ starting at 0. Then by the above set inclusions,

$$\mathbb{P}(X_n > cn^{\frac{3}{2}+\varepsilon}) \leq \mathbb{P}(\max_{0\leq k\leq 2n} S_k > cn^{\frac{1}{2}+\varepsilon})$$

Let $\{Y_i\}_{i=1}^{\infty}$ be i.i.d. ± 1-valued coinflips, so $S_n = \sum_{i=1}^{n} Y_i$ for $n \geq 1$.

$$M(t) = \mathbb{E}[e^{t\cdot Y_i}] = \frac{1}{2}e^{-t} + \frac{1}{2}e^{t}$$

$$\mathbb{E}[e^{tS_n}] = \mathbb{E}[\prod_{i=1}^{n} e^{tY_i}] = \prod_{i=1}^{n} \mathbb{E}[e^{tY_i}] = \mathbb{E}[e^{tY_1}]^n$$

$$\mathbb{P}(S_n \geq c) = \mathbb{P}(e^{tS_n} \geq e^{ta}) \leq \inf_{t>0} \mathbb{E}[e^{tS_n}]e^{-ta} = \inf_{t>0} M(t)^n e^{-ta}$$

$W_n := \max_{0\leq k\leq n} S_k$

$$\mathbb{P}(W_n \geq r, S_n = b) = \begin{cases} \mathbb{P}(S_n = b) & \text{if } b \geq r \\ \mathbb{P}(S_n = 2r - b) & \text{otherwise} \end{cases}$$

By the reflection principle,

$$\mathbb{P}\left(\max_{0\leq k\leq 2n} S_k > cn^{\frac{1}{2}+\varepsilon}\right) = 2\,\mathbb{P}(S_{2n} \geq cn^{\frac{1}{2}+\varepsilon} + 1) + \mathbb{P}(S_{2n} = cn^{\frac{1}{2}+\varepsilon})$$

Now, applying Chernoff's bound and explicitly computing minima, we find

$$\mathbb{P}(S_{2n} \geq cn^{\frac{1}{2}+\varepsilon}) \leq \inf_{t>0} M(t)^{2n} e^{-tcn^{\frac{1}{2}+\varepsilon}} = \inf_{t>0} \frac{1}{4^n}\left(e^{-t} + e^{t}\right)^{2n} e^{-tcn^{\frac{1}{2}+\varepsilon}} \leq \inf_{t>0} e^{t^2 n} e^{-tcn^{\frac{1}{2}+\varepsilon}}$$

$$\begin{aligned}\mathbb{P}(S_{2n} \geq cn^{\frac{1}{2}+\varepsilon} + 1) &\leq \inf_{t>0} M(t)^{2n} e^{-t(cn^{\frac{1}{2}+\varepsilon}+1)} = \inf_{t>0} \frac{1}{4^n}\left(e^{-t} + e^{t}\right)^{2n} e^{-t(cn^{\frac{1}{2}+\varepsilon}+1)} \\ &\leq \inf_{t>0} e^{t^2 n} e^{-t(cn^{\frac{1}{2}+\varepsilon}+1)} = e^{-\frac{c^2}{4}n^{2\varepsilon} - \frac{c}{2}n^{-\frac{1}{2}+\varepsilon} - \frac{1}{4n}}\end{aligned}$$

$$\begin{aligned}\mathbb{P}(S_{2n} = cn^{\frac{1}{2}+\varepsilon}) &\leq \mathbb{P}(S_{2n} \leq cn^{\frac{1}{2}+\varepsilon} + 1) - \mathbb{P}(S_{2n} \leq cn^{\frac{1}{2}+\varepsilon}) \\ &\leq \inf_{t<0} e^{t^2 n} e^{-tcn^{\frac{1}{2}+\varepsilon}} + \inf_{t<0} e^{t^2 n} e^{-t(cn^{\frac{1}{2}+\varepsilon}+1)} \leq 2\end{aligned}$$

Whence, we finally compute

$$\mathbb{P}\left(\max_{0\leq k\leq 2n} S_k > cn^{\frac{1}{2}+\varepsilon}\right) \leq 2e^{-\frac{c^2}{4}n^{2\varepsilon} - \frac{c}{2}n^{-\frac{1}{2}+\varepsilon} - \frac{1}{4n}} + 2 = O(\exp(-\widetilde{c}n^{2\varepsilon}))$$

References

[CEF07] Szu-En Cheng, Sen-Peng Eu, and Tung-Shan Fu. "Area of Catalan paths on a checkerboard". In: *European Journal of Combinatorics* 28.4 (2007), pp. 1331–1344.

[MSV96] Donatella Merlini, Renzo Sprugnoli, and M. Cecilia Verri. "The area determined by underdiagonal lattice paths". In: *Trees in Algebra and Programming — CAAP '96*. Ed. by Hélène Kirchner. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 59–71.