Partial Sampling of Huge Random Objects

Amartya Shankha Biswas
CSAIL, MIT
asbiswas@mit.edu

Ronitt Rubinfeld
CSAIL, MIT
ronitt@csail.mit.edu

Anak Yodpinyanee
CSAIL, MIT
anak@csail.mit.edu

Abstract

Consider an algorithm performing a computation on a huge random object (for example a random graph or a "long" random walk). Is it necessary to generate the entire object prior to the computation, or is it possible to provide query access to the object and sample it incrementally "on-the-fly" (as requested by the algorithm)? Such an *implementation* would emulate the random object by answering appropriate queries in a consistent manner. Specifically, all responses to queries must be consistent with an instance of the random object sampled from the true distribution (or close to it). This paradigm would be useful when the algorithm is sub-linear and therefore it is inefficient to sample the entire object up front.

Our results focus on undirected graphs with independent edge probabilities, that is, each edge is chosen as an independent Bernoulli random variable. We provide a general implementation for generators in this model. Then, we use this construction to obtain the first efficient local implementations for the Erdös-Rényi $G(n, p)$ model, and the Stochastic Block model. As in previous local-access implementations for random graphs, we support `Vertex-Pair`, `Next-Neighbor` queries, and `All-Neighbors` queries. In addition, we introduce a new `Random-Neighbor` query. We also give the first local-access generation procedure for `All-Neighbors` queries in the (sparse and directed) Kleinberg's Small-World model. Note that, in the sparse case, an `All-Neighbors` query can be used to simulate the other types of queries efficiently. All of our generators require no pre-processing time, and answer each query using $\mathcal{O}(\text{poly}(\log n))$ time, random bits, and additional space.

We also show how to implement random Catalan objects, specifically focusing on Dyck paths (balanced random walks that are always positive). Here, we support `Height` queries to find the location of the walk and `First-Return` queries to find the time when the walk returns to a specified location. As an application, we show that this can be used to implement `Next-Neighbor` queries on random rooted and binary trees, and `Matching-Bracket` queries on random well bracketed expressions (the Dyck language).

Finally, we study random q-colorings of graphs with maximum degree Δ. This is a new setting where the random object also has a "huge" description (the underlying graph) that can be accessed through adjacency list queries. This setting is similar to Local Computation Algorithms [RTVX11, ARVX12] with the added restriction that the output must follow a specific distribution in addition to being legal. We show how to sample the color of a single node in sub-linear time when $q > \alpha\Delta$ where α is a small constant.

2012 ACM Subject Classification Author: Please fill in 1 or more \ccsdesc macro

Keywords and phrases Dummy keyword

Funding *Amartya Shankha Biswas*: funding
Ronitt Rubinfeld: funding

Anak Yodpinyanee: funding

Contents

Thresholding.

Mention that this is imperfect sampling (close impl.)

we have?

Connect with previous part on SBM

1 Sampling Catalan Objects

Earlier, we were interested in querying the following random object. In a random permutation of n white marbles and n black marbles, how many white marbles are present in the first k positions. As we have seen before, [GGN10] gives us a method of sampling from this (hypergeometric) distribution. In constructing a generator for the Stochastic Block model, we generalized this by adding more colors (multivariate hypergeometric distribution). We also took this to the extreme where all marbles are distinguishable (i.e. a random permutation), and saw that this could also be implemented efficiently. Now we focus on a more challenging variant of this question with more complicated conditional dependences among the placement of the marbles.

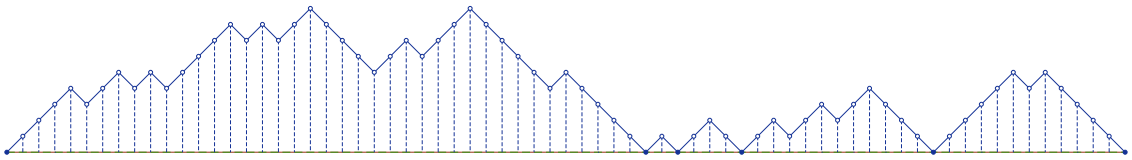

Figure 1 Simple Dyck path with $n = 35$.

Important extension of interval summable queries.

We consider a sequence of n white and n black marbles such that every prefix of the sequence has at least as many white marbles as black ones. This can be interpreted as a Dyck path; a $2n$ step *balanced* one-dimensional walk with exactly n up and down steps. In Figure 1, each step moves one unit along the positive x-axis (time) and one unit up or down the positive y-axis (position). The prefix constraint implies that the y-coordinate of any point on the walk is ≥ 0 i.e. the walk never crosses the x-axis. The number of possible Dyck paths (see Theorem 25) is the n^{th} Catalan number $C_n = \frac{1}{n+1} \cdot \binom{2n}{n}$. Many important combinatorial objects occur in Catalan families of which these are an example.

Our goal will be to support queries to a uniformly random instance of a Dyck path, which will in turn allow us to sample other random Catalan objects such as rooted trees, and bracketed expressions. Specifically, we will want to answer the following queries:

- $\textsc{Height}(i)$: Returns the position of the path after i steps
- $\textsc{First-Return}(i)$: Returns an index $j > i$ such that $\textsc{Height}(j) = \textsc{Height}(i)$ and for any k between i and j, $\textsc{Height}(k)$ is strictly greater than $\textsc{Height}(i)$.

1.1 Bijections to other Catalan objects

The $\textsc{Height}$ query is natural for Dyck paths, but the $\textsc{First-Return}$ query is important in exploring other Catalan objects. For instance, consider a random well bracketed expression; equivalently an uniform distribution over the Dyck language. One can construct a trivial bijection between Dyck paths and words in this language by replacing up and down steps with opening and closing brackets respectively. The $\textsc{Height}$ query corresponds to asking for the nesting depth at a certain position in the word, and $\textsc{First-Return}(i)$ returns the position of the matching bracket for position i.

There is also a natural bijection between Dyck paths and rooted ordered trees by letting the path be a transcript of the DFS traversal of a tree. Starting with the root, for each "up-step" we move to a new child of the current node, and for each "down-step", we backtrack towards the root. Here, the $\textsc{Height}$ query returns the depth of a node and the $\textsc{First-Return}$ query can be used to find the *next child* of a node.

Figure? Degree queries by repeated application.

Moving forwards, we will focus on Dyck paths for the sake of simplicity.

1.2 Catalan Trapezoids and Generalized Dyck Paths

In order to sample Dyck paths locally, we will need to analyze more general Catalan objects. Specifically, we consider a sequence of U up-steps and D down-steps, 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 (Figure 2).

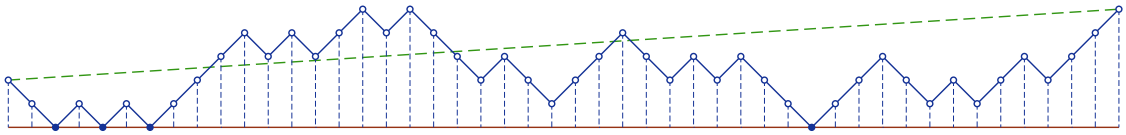

Figure 2 Complex Dyck path with $U = 25$, $D = 22$ and $k = 3$. Notice that the boundary is shifted.

We will denote the set of such *generalized Dyck paths* as $\mathbb{C}_k(U, D)$ and the number of paths as $C_k(U, D) = |\mathbb{C}_k(U, D)|$, which is an entry in the *Catalan Trapezoid* of order k (presented in [Reu14]). We also use $\mathsf{C}_k(U, D)$ to denote the uniform distribution over $\mathbb{C}_k(n, m)$. Now, we state a result from [Reu14] without proof

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

For $k = 1$ and $n = m$, these represent the vanilla Catalan numbers i.e. $C_n = C_1(n, n)$ (number of simple Dyck paths). Our goal is to sample from the distribution $\mathsf{C}_1(n, n)$.

Consider the situation after sampling the height of the Dyck path at various locations $\langle x_1, x_2, \cdots, x_m\rangle$. The revealed locations partition the path into disjoint *intervals* $[x_i, x_{i+1}]$ where the heights of the endpoints have been sampled. We define $y_i = \textsc{Height}(x_i)$ and notice that these intervals are independent of each other. Specifically, the path in the interval $[x_i, x_{i+1}]$ will be sampled from $\mathsf{C}_k(U, D)$, where $k-1 = y_i$, $U+D = x_{i+1}-x_i$, and $U-D = y_{i+1}-y_i$, and this happens independent of any samples outside the interval. Next, we will show how one can sample heights within such an interval, and in Section 1.4 we will move on to the more complicated $\textsc{First-Return}$ queries.

Is this necessary?

We also maintain a threshold $\mathcal{T} = \Theta(\log^4 n)$. If a query lands in an interval that has length less than $\mathcal{T}$, then we brute force sample the entire interval one step at a time. Assuming that the probabilities of these events can be approximated efficiently (Lemma 30, this take poly$(\log n)$ time.

1.3 Sampling the Height

We will implement $\textsc{Height}(t)$ by showing how to (efficiently) sample the position of the path in the midpoint of an existing interval. We can then extend this to arbitrary positions by running a binary search on the appropriate interval using the midpoint samples. If the interval in question has odd length, we sample one step on the boundary and proceed with a shortened interval.

Our general recursive step is as follows. We consider an interval of length $2B$ comprising of $2U$ up-steps and $2D$ down-steps where the sum of any prefix cannot be less than $k-1$ i.e. this interval should be sampled from $\mathsf{C}_k(2U, 2D)$ (the factor of two makes the analysis cleaner). Without loss of generality, we assume that $2D \le B$; if this were not the case, we could simply flip the sequence and

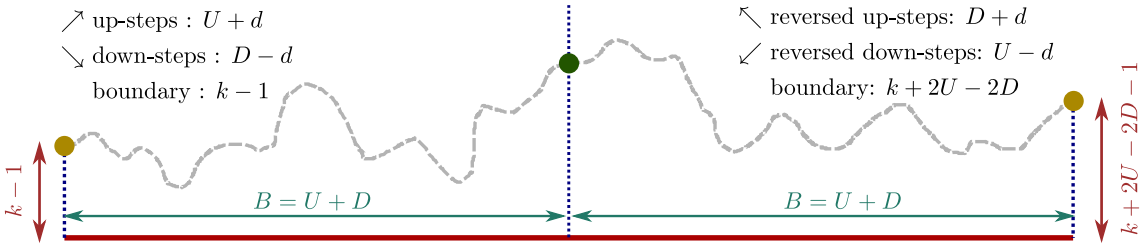

Figure 3 The $2B$-interval is split into two equal parts resulting in two separate Dyck problems. The green node (center) is the sampled height of the midpoint corresponding to a specific value of d. The path considered in both sub-intervals starts at a yellow node (left and right edges) and ends at the green node. From this perspective, the path on the right is reversed with up and down steps being swapped. A possible path is shown in gray.

negate the elements. This essentially means that the overall path in the interval is non-decreasing in height.

We will sample the height of the path $B = U + D$ steps into the interval at the midpoint (see Figure 3). This is equivalent to sampling the number of up or down steps that get assigned to the first half of the interval. We parameterize the possibilities by d and define p_d to be the probability that exactly $U + d$ up-steps and $D - d$ down steps get assigned to the first half, and therefore the second half gets exactly $U - d$ up steps and $D + d$ down steps.

$$p_d = \frac{S_{left}(d) \cdot S_{right}(d)}{S_{total}(d)}$$

Here, $S_{left}(d)$ denotes the number of possible paths in the first half (using $U + d$ up steps) and $S_{right}(d)$ denotes the number of possible paths in the second half (using $U - d$ up steps). Note that all of these paths have to respect the k-boundary constraint (cannot dip more than $k - 1$ units below the starting height). Moving forwards, we will drop the d when referring to the path counts. We (conceptually) flip the second half of the interval, such that the corresponding path begins from the end of the $2B$-interval and terminates at the midpoint (Figure 3). This results in a different starting point, and the boundary constraint will also be different. Hence, we define $k' = k + 2U - 2D$ to represent the new boundary constraint (since the final height of the $2B$-interval is $k' - 1$). Finally, S_{total} is the total number of possible paths in the interval (of length $2B$).

Frequently?

We will use the following rejection sampling lemma from [GGN10].

▶ **Lemma 1.** *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$(\log n)$-time recognizable set B such that*
   - *$1 - \sum_{i \in B} p_i$ 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.

An important point to note is that in order to apply this lemma, we must be able to compute the p_d values at least approximately. For now, we will assume that we have access to an oracle that

will compute the value for us. Lemma 30 shows how to construct such an oracle. We also use the following lemmas to bound the deviation of the path.

▶ **Lemma 2.** *Consider a contiguous* sub-path *of a simple Dyck path of length* $2n$ *where the sub-path is of length* $2B$ *comprising of* U *up-steps and* D *down-steps (with* $U + D = 2B$*). Then there exists a constant* c *such that the quantities* $|B - U|$, $|B - D|$, *and* $|U - D|$ *are all* $< c\sqrt{B \log n}$ *with probability at least* $1 - 1/n^2$ *for every possible sub-path.*

1.3.1 The Simple Case: Far Boundary

unconstrained random walks will not dip below $1 - k$ threshold whp

▶ **Lemma 3.** *Given a Dyck path sampling problem of length* B *with* U *up and* D *down steps with a bounary at* k*, there exists a constant* c *such that if* $k > c\sqrt{B \log n}$*, then the distribution of paths sampled without a boundary* $\mathsf{C}_\infty(U, D)$ *(hypergeometric sampling) is statistically* $\mathcal{O}(1/n^2)$*-close in* L_1 *distance to the distribution of Dyck paths* $\mathsf{C}_k(U + D)$*.*

By Lemma 3, the problem of sampling from $C_k(2U, 2D)$ reduces to sampling from the hypergeometric distribution $C_\infty(2U, 2D)$ when $k > \mathcal{O}(\sqrt{B \log n})$ i.e. the probabilities p_d can be approximated by:

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

This problem of sampling from the hypergeometric distribution is equivalent to the interval summable functions implemented in [GGN10] (using $\mathcal{O}(poly(\log n))$ resources).

1.3.2 Path Segments Close to Zero

A difficult case is when $k = \mathcal{O}(\sqrt{B \log n})$ and we need to compute the actual probability p_d, For the definitions in Section A.4, we see that:

$$S_{left} = C_k(U + d, D - d) \qquad S_{right} = C_{k'}(U - d, D + d) \qquad S_{total} = C_k(2U, 2D) \tag{2}$$

Here, $k' = k + 2U - 2D$, and so $k' = \mathcal{O}(\sqrt{B \log n})$ (using Lemma 2). The distribution $\mathsf{C}_k(2U, 2D)$ we wish to sample from has probabilties $p_d = S_{left} \cdot S_{right}/S_{total}$. We will use rejection sampling (Lemma 1) by constructing a different distribution q_d that approximates p_d up to logarithmic factors over the vast majority of its support (we ignore all $|d| > \Theta(\sqrt{B \log n})$ since the associated probability mass is negligible by Lemma 2). To invoke the rejection sampling lemma, we present a method to approximate the probabilities p_d in Lemma 30. The only thing left to do is to find an appropriate q_d that also has an *efficiently computable CDF*. Surprisingly, as in Section 1.3.1, we will be able to use the hypergeometric distribution for q_d,

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

However, the argument for why this q_d is a good approximation to p_d is far less straightforward.

First, we consider the case where $k \cdot k' \leq 2U + 1$. In this case, we use loose bounds for $S_{left} < \binom{B}{D-d}$ and $S_{right} < \binom{B}{U-d}$ along with the following lemma (proven in Section A).

▶ **Lemma 4.** *When* $kk' > 2U + 1$*,* $S_{total} > \frac{1}{2} \cdot \binom{2B}{2D}$*.*

Combining the three bounds we obtain $p_d < \frac{1}{2} q_d$. Intuitively, in this case the Dyck boundary is far away, and therefore the number of possible paths is only a constant factor away from the

number of unconstrained paths (see Section 1.3.1). The case where the boundaries are closer (i.e. $k \cdot k' \leq 2U + 1$) is trickier, since the individual counts need not be close to the corresponding binomial counts. However, in this case we can still ensure that the sampling probability is within poly-logarithmic factors of the binomial sampling probability. We use the following lemmas (proven in Section A).

▶ **Lemma 5.** *$S_{left} \leq c_1 \frac{k \cdot \sqrt{\log n}}{\sqrt{B}} \cdot \binom{B}{D-d}$ for some constant c_1.*

▶ **Lemma 6.** *$S_{right} < c_2 \frac{k' \cdot \sqrt{logn}}{\sqrt{B}} \cdot \binom{B}{U-d}$ for some constant c_2.*

▶ **Lemma 7.** *When $kk' \leq 2U + 1$, $S_{total} < c_3 \frac{k \cdot k'}{B} \cdot \binom{2B}{2D}$ for some constant c_3.*

We can now put these lemmas together to show that $p_d/q_d \leq \Theta(\log n)$ and invoke Lemma 1 to sample the value of d. This gives us the height of the Dyck path at the midpoint of the two given points.

▶ **Theorem 8.** *Given two positions a and b (and the associated heights) in a Dyck path of length $2n$ with the guarantee that no position between a and b has been sampled yet, there is an algorithm that returns the height of the path halfway between a and b. Moreover, this algorithm only uses $\mathcal{O}(poly(\log n))$ resources.*

Proof. If $b - a$ is even, we can set $B = (b - a)/2$. Otherwise, we first sample a single step from a to $a + 1$, and then set $B = (b - a - 1)/2$. Since there are only two possibilities for a single step, we can explicitly approximate the probabilities, and then sample accordingly. This allows us to apply the rejection sampling from Lemma 1 using $\{q_d\}$ to obtain samples from $\{p_d\}$ as defined above. □

▶ **Theorem 9.** *There is an algorithm that provides sample access to a Dyck path of length $2n$, by answering queries of the form* HEIGHT(x) *with the correctly sampled height of the Dyck path at position x using only $\mathcal{O}(poly(\log n))$ resources per query.*

Proof. The algorithm maintains a successor-predecessor data structure (e.g. Van Emde Boas tree) to store all positions x that have already been sampled. Each newly sampled position is added to this structure. Given a query HEIGHT(x), the algorithm first finds the successor and predecessor (say a and b) of x among the alredy queried positions. This provides us the quarantee required to apply Theorem 8, which allows us to query the height at the midpoint of a and b. We then binary search by updating either the successor or predecessor of x and repeat until we sample the height of position x. □

1.4 Supporting "First Return" Queries

We also support more complex queries to a Dyck path. Specifically, in addition to querying the height after t steps, we introduce a FIRST-RETURN query that allows the user to query the next time the path returns to that height (if at all i.e. only if the step from x to $x + 1$ is an up-step).

The utility of this kind of query can be seen in other interpretations of Catalan objects. For instance, if we interpret it as a well bracketed expression, FIRST-RETURN(x) returns the position of the bracket matching the one started at x. If we consider a uniformly random rooted tree, the function effectively returns the next child of a vertex (see Section 1.1).

1.4.1 Maintaining a Boundary Invariant

Notice that over a sequence of HEIGHT queries $\langle x_1, x_2, \cdots, x_m \rangle$ to the Dyck path, many different positions are revealed (possibly in adverserial locations). This partitions the path into at most $m+1$ disjoint and independent *intervals* with set boundary conditions. The first step towards finding the FIRST-RETURN from position t would be to locate the *interval* where the return occurs. Even if we had an efficient technique to filter intervals, we would want to avoid considering all $\Theta(m)$ intervals to find the correct one. In addition the inefficiency, the fact that a specific interval *does not* contain the first return causes dependencies for all subsequent samples.

We resolve this issue by maintaining the following invariant. Consider all positions that have been queried already $\langle x_1, x_2, \cdots, x_m \rangle$ (in increasing order) along with their corresponding heights $\langle y_1, y_2, \cdots, y_m \rangle$.

▶ **Invariant 10.** *For any interval $[x_i, x_{i+1}]$ where the heights of the endpoints have been sampled to be y_i and y_{i+1}, and no other position in the interval has yet been sampled, the section of the Dyck path between positions x_i and x_{i+1} is constrained to lie above $min(y_i, y_{i+1})$.*

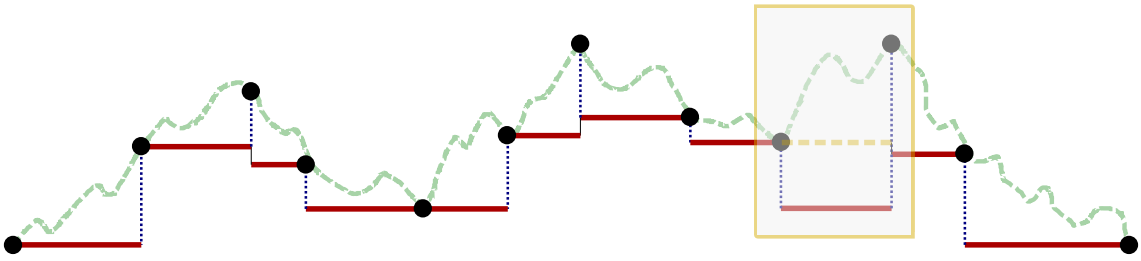

Figure 4 Error in third segment.

It's not even clear that it is always possible to maintain such an invariant. After sampling the height of a particular position x_i as y_i (with $x_{i-1} < x_i < x_{i+1}$), the invariant is potentially broken on either side of x_i. We will re-establish the invariant by sampling an additional point on either side. This proceeds as follows for the interval between x_i and x_{i+1} (see error in Figure 4):

1. Sample the lowest height h achieved by the walk between x_i and x_{i+1}.
2. Sample a position x such that $x_i < x < x_{i+1}$ and HEIGHT$(x) = h$.

Since h is the minimum height along this interval, sampling the point x suffices to preserve the invariant. Lemma 16 shows how this invariant can be used to efficiently find the interval containing the first return.

1.4.2 Sampling the Lowest Achievable Height in an Interval

For the first step, we need to sample the lowest height h of the walk between x_i and x_{i+1} (with $x_i < x_{i+1}$). Say that this interval defines a generalized Dyck problem with U up steps and D down steps with a boundary that is $k-1$ units below y_i.

Given any two boundaries k_{lower} and k_{upper} on the same interval (with $k_{lower} < k_{upper}$, we can count the number of possible generalized Dyck paths that violate the k_{upper} boundary but *not* the k_{lower} boundary as $P(k_{lower}, k_{upper}) = C_{k_{lower}}(U, D) - C_{k_{upper}}(U, D)$. We set $k_{low} = k, k_{up} = 0$, and $k_{mid} = (k_{low} + k_{up})/2$. Since we can compute $P(k_{low}, k_{up})$, $P(k_{low}, k_{mid})$, and $P(k_{mid}, k_{up})$, we can sample a single bit to decide if the "lower boundary" should move up or if the "upper boundary" should move down. We then repeat this binary search until we find $k' = k_{low} = k_{up} - 1$

and k' becomes the "mandatory boundary" (i.e. the walk reaches the height exactly $k'-1$ units below y_i but no lower.

1.4.3 Sampling First Position that Touches the *"Mandatory Boundary"*

Now that we have a "*mandatory boundary*" k, we just need to sample a position x with height $h = x_i - k + 1$. In fact, we will do something stronger by sampling the *first* time the walk touches the boundary after x_i. As before, we assume that this interval contains U up steps and D down steps.

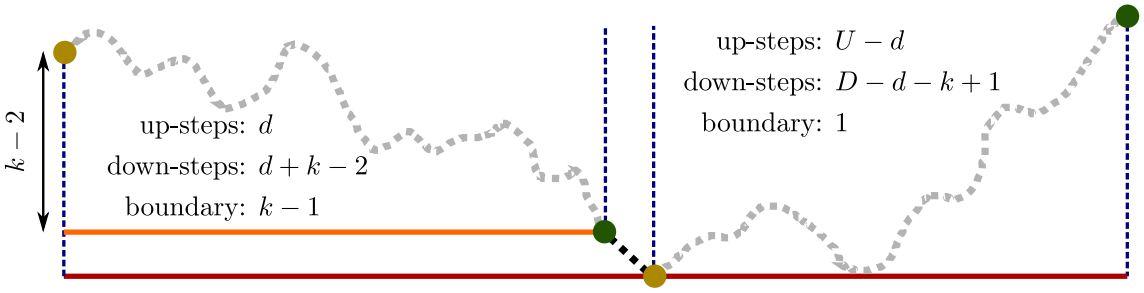

Figure 5 Zooming into the error in Figure 4

We will parameterize the position x the number of up-steps d between x_i and x (See Figure 5). implying that $x = x_i + 2d + k - 1$. Given a specific d, we want to compute the number of valid paths that result in d up-steps before the first approach to the boundary. We will calculate this quantity by counting the total number of paths to the left and right of the first approach and multiplying them together.

Since we only care about getting a asymptotic (up to poly$(\log n)$ factors) estimate of the probabilities, it suffices to estimate the number of paths asymptotically as well. As in Section 1.3.2, we define S_{left} to be the number of paths in the sub-interval before the first approach (left side of Figure 5), S_{right} to be the number of paths following the first approach, and S_{total} to be the toal number of paths that touch the "mandatory boundary" at k:

$$S_{left} = C_k(d, d+k-2) \quad S_{right} = C_1(U-d, D-d-k+1) \quad S_{total} = C_k(U,D) - C_{k-1}(U,D)$$

Our goal is to sample d from the distribution $\{p_d\}$ where $p_d = S_{left} \cdot S_{right}/S_{total}$. Using Equation 1, we obtain the following approximations for S_{left} and S_{right}.

▶ **Lemma 11.** *If $d > \log^4 n$, then $S_{left}(d) = \Theta\left(\frac{2^{2d+k}}{\sqrt{d}} e^{-r_{left}(d)} \cdot \frac{k-1}{d+k-1}\right)$ where $r_{left}(d) = \frac{(k-2)^2}{2(2d+k-2)}$. Furthermore, $r_{left}(d) = \mathcal{O}(\log^2 n)$.*

▶ **Lemma 12.** *If $U + D - 2d - k > \log^4 n$, then $S_{right}(d) = \Theta\left(\frac{2^{U+D-2d-k}}{\sqrt{U+d-2d-k}} e^{-r_{right}(d)} \cdot \frac{U-D+k}{U-d+1}\right)$ where $r_{right}(d) = \frac{(U-D-k-1)^2}{4(U+D-2d-k+1)}$. Furthermore, $r_{right}(d) = \mathcal{O}(\log^2 n)$.*

Deal with other values of d

We now consider the values of d that are outside the range of the two preceding lemmas. These values are the ones where $d < \log^4 n$ or $2d > U + D - k - \log^4 n$. Since d is the number of up steps (in the left sub-interval), $d \geq 0$ and since the length of the right sub-interval nust be non-negative, we get $U + D - 2d - k + 1 \geq 0$. Thus, we define the set

$$\mathcal{R} = \left\{ d \;\middle|\; 0 \leq d < \log^4 n \text{ or } -1 < 2d - U - D + k < \log^4 n \right\}$$

Clearly, we can bound the size of this set as $|\mathcal{R}| = \mathcal{O}(\log^4 n)$. An immediate consequence of Lemma 11 and Lemma 12 is the following.

▶ **Corollary 13.** *When $d \notin \mathcal{R}$, $S_{left}(d) \cdot S_{right}(d) = \Theta\left(\frac{2^{U+D}}{\sqrt{d(U+D-2d-k)}} \cdot e^{-r(d)} \cdot \frac{k-1}{d+k-1} \cdot \frac{U-D+k}{U-d+1}\right)$ where $r(d) = \mathcal{O}(\log^2 n)$.*

1.4.4 Estimating the CDF

We will now use these observations to construct a suitable $\{q_d\}$ that can be used to invoke the rejection sampling lemma. In addition to beign a good approximation to $\{p_d\}$, $\{q_d\}$ must be of a form that allows us to compute its CDF efficiently. First, we rewrite the approximate probability as $p_d = \Theta\left(\mathcal{K} \cdot f(d) \cdot e^{-r(d)}\right)$ where:

$$\mathcal{K} = \frac{2^{U+D}}{S_{total}} = \frac{2^{U+D}}{C_k(U,D) - C_{k-1}(U,D)} \qquad f(d) = \frac{(k-1)(U-D+k)}{\sqrt{d(U+D-2d-k)}(d+k-1)(U-d+1)}$$

Notice that $\mathcal{K}$ is a constant and $f(d)$ is a function whose integral has a closed form. Using the fact that $r(d) = \mathcal{O}(\log^2 n)$ (from Corollary 13), we obtain the following lemma:

▶ **Lemma 14.** *Given the piecewise continuous function*

$$\hat{q}(\delta) = \begin{cases} p_{\lfloor \delta \rfloor} & \text{if } \lfloor \delta \rfloor \in \mathcal{R} \\ \mathcal{K} \cdot f(\delta) \cdot exp\left(\lfloor r(\lfloor \delta \rfloor) \rfloor\right) & \text{if } \lfloor \delta \rfloor \notin \mathcal{R} \end{cases} \implies p_d = \Theta\left(\int_d^{d+1} \hat{q}(\delta)\right)$$

Proof. For $d \in \mathcal{R}$, the integral trivially evaluateds to exactly p_d. For $d \notin \mathcal{R}$, it suffices to show that $p_d = \Theta\left(\hat{q}(\delta)\right)$ for all $\delta \in [d, d+1)$. We already know that $p_d = \Theta\left(\mathcal{K} \cdot f(d) \cdot e^{-r(d)}\right)$. Moreover, for any $\delta \in [d, d+1)$, the exponential term in $\hat{q}(\delta)$ is within a factor of e of the original $e^{-r(d)}$ term. □

Now, we have everything in place to define the distribution $\{q_d\}$ that we will be sampling from. Specifically, we will define q_d and it's CDF Q_d as follows:

$$q_d = \frac{\int_d^{d+1} \hat{q}(d)}{\mathcal{N}} \qquad Q_d = \frac{\int_0^{d+1} \hat{q}(d)}{\mathcal{N}} \tag{3}$$

Here the normalizing factor $\mathcal{N}$ is $\int_0^{d_{max}+1} \hat{q}(d)$. To show that these can be computed efficiently, it suffices to show that any integral of $\hat{q}_d$ can be efficiently evaluated.

▶ **Theorem 15.** *Kinesthetics*

Needs a theorem

1.4.5 Finding the Correct Interval: `First-Return` Query

As before, consider all positions that have been queried already $\langle x_1, x_2, \cdots, x_m \rangle$ (in increasing order) along with their corresponding heights $\langle y_1, y_2, \cdots, y_m \rangle$.

▶ **Lemma 16.** *For any position x_i, assuming that Invariant 10 holds, we can find the interval $(x_{k-1}, x_k]$ that contains* FIRST-RETURN(x_i) *We do this by by setting k to be either the smallest index $f > i$ such that $y_f \leq y_i$ or setting $k = i + 1$.*

Proof. We assume the contrary i.e. there exists some $k \neq f$ and $k \neq i+1$ such that the correct interval is $(x_{k-1}, x_k]$. Since $y_f < y_i$, the position of first return to y_i happens in the range $(x_i, x_f]$. So, the only possibility is $i+1 < k \leq f-1$. By the definition of y_f, we know that both y_k and y_{k-1} are strictly larger than y_i. Invariant 10 implies that the boundary for this interval $(y_{k-1}, y_k]$ is at $\min(y_{k-1}, y_k) > y_i$. So, it is not possible for the first return to be in this interval. □

The good news is that there are only two intervals that we need to worry about. Now the challenge is to find the smallest index $f > i$ such that $y_f \leq y_i$. One solution is to maintain an interval tree over the range $[2n]$ storing the position of the boundary. Specifically, we have a balanced binary tree with $2n$ leaves with the i^{th} leaf storing the boundary at position i. Each internal node stores the minimum value amongst all the leaves in its sub-tree. In this setting, we can binary search for f by guessing a bound f' and performing a *range minimum query* over the interval $(x_i, x_{f'}]$. Overall, this requires $\mathcal{O}(\log n)$ range queries each of which makes $\mathcal{O}(\log n)$ probes to the binary tree.

cite

However, we cannot explicitly maintain or even construct this tree, and updates can be as expensive as $\Theta(n)$. To mitigate this, we start with just a root node (indicating that the initial boundary is 1 everywhere) and build the tree dynamically as needed. We perform updates using *lazy propagation* by only propagating updates down to the children (creating children if necessary) when needed. So, at any given time, some nodes in the tree may not hold the correct value, but the correct value must be present on the path to the root.

cite

This probably needs more explanation

▶ **Theorem 17.** *There is an algorithm using $\mathcal{O}(poly(\log n))$ resources per query that provides sample access to a Dyck path of length $2n$ by answering queries of the form* `First-Return`(x_i) *with the correctly sampled position y; where $y > x_i$ is the position where the Dyck path first returns to* `Height`(x_i) *after position x_i.*

Proof. We first query the interval $(x_i, x_{i+1}]$ to find a first return using Theorem 15. If a return is not found, we calculate f using . Since $x_{f-1} < x_i \leq x_f$ by definition, the interval $(x_{f-1}, x_f]$ must contain a position at height y_i. We sample a point in the middle of this interval and fix the boundary invariant by sampling another point, essentially breaking it up into $\mathcal{O}(1)$ sub-intervals each at most half the size of the original. Based on the new samples, we find the sub-interval containing the first return in $\mathcal{O}(1)$ time. We repeat up to $\mathcal{O}(\log n)$ times until the current interval size drops below the threshold $\mathcal{T}$. Then we spend $\tilde{\mathcal{O}}(\mathcal{T})$ time to brute force sample this interval and find the first return position (if it wasn't revealed in previous steps). □

what?

2 Random Coloring of a Graph

Query access

We wish to locally sample an uniformly random coloring of a graph. A q-coloring of a graph $G = (V, E)$ is a function $\sigma : V \to [q]$, such that for all $(u, v) \in E$, $\sigma_u \neq \sigma_v$. We will consider only bounded degree graphs, i.e. graphs with max degree $\leq \Delta$. Otherwise, the coloring problem becomes NP-hard.

cite

Using the technique of path-coupling, Vigoda showed that for $q > 2\Delta$, one can sample an uniformly random coloring by using a MCMC algorithm.

cite

The Markov Chain proceeds in T steps. The state of the chain at time t is given by $\mathbf{X}^t \in [q]^{|V|}$. Specifically, the color of vertex v at step t is $\mathbf{X}^t_v$.

In each step of the Markov process, a pair $(v, c) \in V \times [q]$ is sampled uniformly at random. Subsequently, if the recoloring of vertex v with color c does not result in a conflict with v's neighbors, i.e. $c \notin \{X^t_u : u \in \Gamma(v)\}$, then the vertex is recolored i.e. $X^{t+1}_v \leftarrow c$.

After running the MC for $T = \mathcal{O}(n \log n)$ steps we reach the stationary distribution (ϵ close), and the coloring is an uniformly random one.

Exact Bound: $t_{mix}(\epsilon) \leq \left(\frac{q-\Delta}{q-2\Delta}\right) n \left(\log n + \log(1/\epsilon)\right)$ cite book (Peres, Lyons)

2.1 Modified Glauber Dynamics

Now we define a modified Markov Chain as a special case of the *Local Glauber Dynamics* presented in [FG18]. The modified Markov chain procceds in epochs. We denote the initial coloring of the graph by $\mathbf{X}^0$ and the state of the coloring after the k^{th} epoch by $\mathbf{X}^k$. In the k^{th} epoch $\mathcal{E}_k$:

- Sample $|V|$ colors $\langle c_1, c_2, \cdots, c_n \rangle$ from $[q]$, where c_v is the proposed color for vertex v.
- For each vertex v, we set $\mathbf{X}_v^k$ to c_v if for all neighbors w of v, $\mathbf{X}_w^k \neq c_v$ and $\mathbf{X}_w^{k-1} \neq c_v$.

This procedure is a special case of the *Local Glauber Dynamics* presented in [FG18]. The goal in [FG18] is to find a simultaneous update rule that causes few conflicts among neighbors (and converges to the correct distribution). Notice that we *can* have adjacent nodes update in the same epoch. However for the sake of succinctness we use their update rule and avoid a tedious path coupling argument. Cite Path Coupling

We can directly use the path coupling argument from [FG18] which be briefly describe below. Given two colorings $\mathbf{X}$ and $\mathbf{Y}$, we define $d(\mathbf{X}, \mathbf{Y})$ as the number of vertices v such that $\mathbf{X}_v \neq \mathbf{Y}_v$. We define the coupling $(\mathbf{X}, \mathbf{Y}) \to (\mathbf{X}', \mathbf{Y}')$ where $\mathbf{X}$ and $\mathbf{Y}$ differ only at a single vertex v such that $\mathbf{X}_v = c_X$ and $\mathbf{Y}_v = c_Y$. Now, we pick a random permutation of the vertices along with uniformly sampled colors:

$$\langle (v_1, c_1), (v_2, c_2), \cdots, (v_n, c_n) \rangle = \langle (\pi_1, c_1), (\pi_2, c_2), \cdots, (\pi_n, c_n) \rangle$$

Now, for each (v_i, c_i) in order, we update the coloring of X and Y as follows:

- If the current color of v_i as well as c_i are both in $\{c_X, c_Y\}$, then the $\mathbf{X}$ chain picks the color c_i and the $\mathbf{Y}$ chain picks the other color.
- Otherwise, both chains pick the same color c_i for the vertex v_i.

We use the following result from [FG18] that bounds the coupled distance.

▶ **Lemma 18.** *If $q = 2\alpha\Delta$ and $d(\mathbf{X}, \mathbf{Y}) = 1$, then $\mathbb{E}[d(\mathbf{X}', \mathbf{Y}')] \leq 1 - \left(1 - \frac{1}{2\alpha}\right) e^{-3/\alpha} + \frac{1/2\alpha}{1-1/\alpha}$*

▶ **Corollary 19.** *If $q \geq 9\Delta$ and $d(\mathbf{X}, \mathbf{Y}) = 1$, then $\mathbb{E}[d(\mathbf{X}', \mathbf{Y}')] < \frac{1}{e^{1/3}}$*

▶ **Theorem 20.** *If $q \geq 9\Delta$, then the chain is mixed after $\tau_{mix}(\epsilon) = 3\left(\ln n + \ln(\frac{1}{\epsilon})\right)$ epochs.*

Proof. Starting for a maximum distance of n, the distance decreases to 1 after at most $3 \ln n$ epochs, and it takes a further $3 \ln\left(\frac{1}{\epsilon}\right)$ to reduce the distance to ϵ. □

2.2 Local Coloring Algorithm

Given query access to the adjacency matrix of a graph G with maximum degree Δ and a vertex v, the algorithm has to output the color of v after running $t = \mathcal{O}(\ln n)$ epochs of *Modified Glauber Dynamics*. We will define the number of colors as $q = 2\alpha\Delta$ where $\alpha > 1$.

The proposals at each epoch are a vector of color samples $\mathbf{C}^t \sim_{\mathcal{U}} [q]^n$. Note that these values are fully independent and as such any $\mathbf{C}_v^t$ can be sampled trivially. We also use $\mathbf{X}^t$ to denote the final

vector of vertex colors at the end of the t^{th} epoch. Finally, we define indicator variables $\boldsymbol{\chi}_v^t$ to denote if the color for vertex v was accepted at the t^{th} epoch; $\boldsymbol{\chi}_v^t = 1$ if and only if for all neighbors $w \in \Gamma(v)$, we satisfy the condition $\mathbf{C}_v^t \neq \mathbf{X}_w^{t-1}$ and $\mathbf{C}_v^t \neq \mathbf{C}_w^t$. So, the color of a vertex v after the t^{th} epoch $\mathbf{X}_v^t$ is set to be $\mathbf{C}_v^i$ where $i \leq t$ is the largest index such that $\boldsymbol{\chi}_v^i = 1$. While the proposals $\mathbf{C}_v^t$ are easy to sample, it is much less clear how we can sample the $\boldsymbol{\chi}_v^t$ values. Note that we can compute $\mathbf{X}_v^t$ quite easily if we know the values $\boldsymbol{\chi}_v^i$ for all $i \leq t$. So, we focus our attention on the query $\textsc{Accept}(v, t)$ that returns $\boldsymbol{\chi}_v^t$.

2.2.1 Naive Coloring Implementations

The general strategy to implement this is to iterate over all neighbors w of v, and for each of them check if they conflict with v's proposed color. Given a neighbor w, one naive way to do this is to iterate backwards from epoch t querying to find if w's proposal was accepted until the first accepted proposal (from the latest epoch $t' < t$) is found. At this point, if $\mathcal{C}_w^{t'} = C_v^t$, then the current color of w conflicts with v's proposal. Otherwise there is no conflict and we can proceed to the next neighbor. THis process however makes Δ recursive calls to a sub-problem that is only slightly smaller i.e. $T(t) \leq \Delta \cdot T(t-1)$. This leads to a running time upper bound of Δ^t which is superlinear for the desired $t = \Omega(\log n)$.

We can prune the number of recursive calls by only processing the neighbors w which actually proposed the color C_v^t during *some* epoch. In this case, the expected number of neighbors that have to be probed recursively is $\leq t\Delta/q$ (since the total number of neighbor proposals over t epochs is at most $t\Delta$). So, the overall runtime is upper bounded by $(t\Delta/q)^t$. For this algorithm, if we allow $q > t\Delta = \Omega(\Delta \log n)$ colors, the runtime becomes sublinear. This lower bound on q is however asymptotically worse that the sequential requirement $q > 2\Delta = \mathcal{O}(\Delta)$.

Algorithm 6 Generator

```
1:  procedure ACCEPT(v, t)
2:      c ← C_v^t
3:      for w ← Γ(v)
4:          if C_w^t = c
5:              return 0
6:          for t' ← [t, t−1, t−2, ··· , 1]
7:              if C_w^{t'} = c and ACCEPT(w, t')
8:                  flag ← 1
9:                  while t' < t − 1
10:                     t' ← t' + 1
11:                     if ACCEPT(w, t')
12:                         flag ← 0
13:                         break
14:                 if flag = 1
15:                     return 0
16:                 break
17:     return 1
```

2.2.2 Jumping Back to Past Epochs

The expected number of neighbors that need to be checked can always be $t\Delta$ in the worst case. The crucial observation is that even though these recursive calls seem unavoidable, we can aim to reduce the size of the recursive sub-problem and thus bound the number of levels of recursion. Because of the more complex structure of this epoch jumping process, the main challenge is to analyze the runtime.

Algorithm 6 shows our final procedure for sampling $\boldsymbol{\chi}_v^t$ where $c = C_v^t$ is the color proposed by v in epoch t. As before, we iterate through all neighbors w of v. The condition $c \neq \mathbf{C}_w^t$ is can easily be checked by sampling $\mathbf{C}_w^t$ in the current epoch. If no conflict is seen, the next step is to check whether $c \neq \mathbf{X}_w^{t-1}$.

To achieve this, we iterate through all the epochs in reverse order (without making recursive calls) to check whether the color c was ever proposed for vertex w. If not, we can ignore w, and otherwise let's say that the last proposal for c was at epoch t' i.e. $\mathbf{C}_w^{t'} = c$. Now, we directly "jump" to the

t'^{th} epoch and recursively check if this proposal was accepted. If the proposal $C_w^{t'}$ was not accepted, we keep iterating back until we find another candidate proposal for color c or we run out of epochs. Otherwise if $\boldsymbol{\chi}_w^{t'} = 1$ (proposal accepted), we move to epoch $t'+1$ to see if w's color was replaced. If not, we check epoch $t'+2$, $t'+3$, and so on until we reach epoch $t-1$. At this point we have seen that $\boldsymbol{\chi}_w^{t'} = 1$ (color c was accepted) and every subsequent proposal until the current epoch was rejected i.e. $\mathbf{X}_w^{t-1} = c$ and this leads to a conflict with v's current proposal for color c and hence $\boldsymbol{\chi}_v^t = 0$. If at any of the iterations, we see that a different proposal was accepted, then w does not cause a conflict and we can move on to the next neighbor. If we exhaust all the neighbors and don't find any conflicts then $\boldsymbol{\chi}_v^t = 1$.

Now we analyze the runtime of `ACCEPT` by constructing and solving a recurrence relation. We will use the following lemma to evaluate the expectation of products of relevant random variables.

▶ **Lemma 21.** *The probability that any given proposal is rejected* $\mathbb{P}[\boldsymbol{\chi}_v^t = 0]$ *is at most* $1/\alpha$*. Moreover, this upper bound holds even if we condition on all the values in* $\mathbf{C}$ *except* $\mathbf{C}_v^t$*.*

Proof. A rejection can occur due to a conflict with at most 2Δ possible values in $\{C_w^t, X_w^{t-1} | w \in \Gamma(v)\}$. Since there are $2\alpha\Delta$ colors, the rejection probability is at most $1/\alpha$. □

▶ **Definition 22.** *We define* T_t *to be a random variable indicating the number of recursive calls performed during the execution of* `ACCEPT`(v,t) *while sampling a single* $\boldsymbol{\chi}_v^t$*.*

What probes?

So, the number of probes required to check whether a color c (assigned at epoch t') was overwritten at some epoch before t is:

$$\left[T_{t'+1} + \mathcal{B}\left(\frac{1}{\alpha}\right) \cdot T_{t'+2} + \mathcal{B}\left(\frac{1}{\alpha^2}\right) \cdot T_{t'+3} + \cdots + \mathcal{B}\left(\frac{1}{\alpha^{t-t'-2}}\right) \cdot T_{t-1}\right] \tag{4}$$

▶ **Lemma 23.** *For* $\alpha > 4.5$*, the expected number of calls to the procedure* `ACCEPT` *while sampling a single* $\boldsymbol{\chi}_v^t$ *is* $\mathbb{E}[T_t] = \mathcal{O}\left(e^{1.02t/\alpha}\right)$*.*

What probes? Given graph G and q colors ...

Proof. We start with the recurrence for the expected number of probes to $\{\boldsymbol{\chi}^{t'}\}_{t' \in [t]}$ (equivalently calls to `ACCEPT`) used by the algorithm. We will use $\mathcal{B}(p)$ to refer to the Bernoulli random variable with bias p. When checking a single neighbor w, the algorithm iterates through all the epochs t' such that $\mathbf{C}_w^{t'} = c$ (in reality, only the last occurence matters, but we are looking for an upper bound). If such a t' is found (this happens with probability $1/q$ independently for each trial), there is one recursive call to $T_{t'}$. Regardless of what happens, let's say the algorithm queries $T_{t'+1}, T_{t'+2}, \cdots, T_{t-1}$ until an `ACCEPT` proposal is found. Adding an extra $T_{t'}$ term to Equation 4 and summing up over all neighbors and epochs we get the following:

$$T_t \leq \Delta \cdot \sum_{t'=1}^{t} \mathbb{P}[C_w^{t'} = c] \cdot \left[T_{t'} + T_{t'+1} + \mathcal{B}\left(\frac{1}{\alpha}\right) \cdot T_{t'+2} + \mathcal{B}\left(\frac{1}{\alpha^2}\right) \cdot T_{t'+3} + \cdots \right. \tag{5}$$

$$\left. \cdots + \mathcal{B}\left(\frac{1}{\alpha^{t-t'-2}}\right) \cdot T_{t-1}\right] \tag{6}$$

$$\leq \Delta \cdot \mathcal{B}\left(\frac{1}{q}\right) \left[\sum_{t'=1}^{t-1} T_{t'} + \sum_{t'=1}^{t-1} T_{t'} \cdot \left(1 + \mathcal{B}\left(\frac{1}{\alpha}\right) + \mathcal{B}\left(\frac{1}{\alpha^2}\right) + \cdots\right)\right] \tag{7}$$

In the second step, we just group all the terms from the same epoch together. Using Lemma 21 and the fact that $\mathbb{P}[C_w^{t'} = c]$ is independent of all other events, we can write a recurrence for the

expected number of probes.

$$\mathbb{E}[T_t] \leq \Delta \cdot \frac{1}{2\alpha\Delta}\left[\sum_{t'=1}^{t-1} T_{t'} + \sum_{t'=1}^{t-1} T_{t'} \cdot \left(1 + \frac{1}{\alpha} + \frac{1}{\alpha^2} + \cdots\right)\right] \leq \frac{1}{2\alpha} \cdot \sum_{t'=1}^{t-1} T_{t'} \cdot \left[1 + \frac{\alpha}{\alpha - 1}\right] \tag{8}$$

Now, we make the assumption that $\mathbb{E}[T_{t'}] \leq e^{kt/\alpha}$, and show that this satisfies the expectation recurrence for the desired value of k. First, we sum the geometric series:

$$\sum_{t'=1}^{t-1} \mathbb{E}[T_{t'}] = \sum_{t'=1}^{t-1} e^{kt'/\alpha} < \frac{e^{kt/\alpha} - 1}{e^{k/\alpha} - 1} < \frac{e^{kt/\alpha}}{e^{k/\alpha} - 1}$$

The expectation recurrence to be satisfied then becomes:

$$\mathbb{E}[T_t] \leq \frac{1}{2\alpha} \cdot \frac{e^{kt/\alpha}}{e^{k/\alpha} - 1} \cdot \left[1 + \frac{\alpha}{\alpha - 1}\right] = e^{kt/\alpha} \cdot \frac{2\alpha - 1}{2\alpha(\alpha - 1)(e^{k/\alpha} - 1)} = e^{kt/\alpha} \cdot f(\alpha, k)$$

We notice that for $k = 1.02$ and $\alpha > 4.5$, $f(\alpha) < 1$. This can easily be verified by checking that $f(\alpha, 1.02)$ decreases monotonically with α in the range $\alpha > 4.5$. Thus, our recurrence is satisfied for $k = 1.02$, and therefore the expected number of calls is $\mathcal{O}(e^{1.02t/\alpha})$.

Finally, we note that each probe potentially takes time $\mathcal{O}(t\Delta)$ to iterate through all the neighbors in all epochs resulting in a total runtime of $\mathcal{O}(t\Delta e^{kt/\alpha})$. □

▶ **Theorem 24.** *Given adjacency list query access to a graph with n nodes, maximum degree Δ, and $q = 2\alpha\Delta \geq 9\Delta$ colors, we can sample the color of any given node in an $(1/n$-approximate) uniformly random coloring of the graph in a consistent manner using only $\mathcal{O}(n^{6.12/\alpha}\Delta \log n)$ time space and random bits. This is sublinear for $\alpha > 6.12$ and the sampled coloring is $1/n$-close to the uniform distribution in L_1 distance.*

Proof. We compute the mixing time from Theorem 20 to obtain $\tau_{mix}(1/n) = 6 \ln n$ (this is valid since $q > 9\Delta$). Since $\alpha > 4.5$, we can invoke Lemma 23 to conclude that the number of calls to `ACCEPT` is $\mathcal{O}(n^{6.12/\alpha}\Delta \log n)$ which is sublinear for $\alpha > 6.12$. Each call to `ACCEPT`(v, t) potentially spends $t\Delta$ time looking for neighbors in each epoch before t. Since $t \leq 6 \ln n$, the overall runtime becomes $\mathcal{O}(n^{6.12/\alpha}\Delta \log n)$. □

3 Open Problems

- Degree queries for undirected random graphs?
- Faster implementation of coloring $\mathcal{O}(\text{poly}(\log n))$?
- Reduce the required value of α?
- Random walks on other networks? Ideally, any network.

References

ARVX12 Noga Alon, Ronitt Rubinfeld, Shai Vardi, and Ning Xie. Space-efficient local computation algorithms. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1132–1139. Society for Industrial and Applied Mathematics, 2012.

FG18 Manuela Fischer and Mohsen Ghaffari. A simple parallel and distributed sampling technique: Local glauber dynamics. In *32nd International Symposium on Distributed Computing (DISC 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

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.

RTVX11 Ronitt Rubinfeld, Gil Tamir, Shai Vardi, and Ning Xie. Fast local computation algorithms. *arXiv preprint arXiv:1104.1377*, 2011.

Spe14 Joel Spencer. *Asymptopia*, volume 71. American Mathematical Soc., 2014.

Sta15 Richard P Stanley. *Catalan numbers*. Cambridge University Press, 2015.

A Dyck Path Generator

▶ **Theorem 25.** *There are* $\frac{1}{n+1}\binom{2n}{n}$ *Dyck paths for length* $2n$ *(construction from [Sta15]).*

Proof. Consider all possible sequences containing $n+1$ up-steps and n down-steps with the restriction that the first step is an up-step. We say that two sequences belong to the same *class* if they are cyclic shifts of each other. Because of the restriction, the total number of sequences is $\binom{2n}{n}$ and each class is of size $n+1$. Now, within each class, exactly one of the sequences is such that the prefix sums are *strictly greater* than zero. From such a sequence, we can obtain a Dyck sequence by deleting the first up-step. Similarly, we can start with a Dyck sequence, add an initial up-step and consider all $n+1$ cyclic shifts to obtain a *class*. This bijection shows that the number of Dyck paths is $\frac{1}{n+1}\binom{2n}{n}$. □

A.1 Approximating Close-to-Central Binomial Coefficients

We start with Stirling's approximation which states that

$$m! = \sqrt{2\pi m}\left(\frac{m}{e}\right)^m\left(1+\mathcal{O}\left(\frac{1}{m}\right)\right)$$

We will also use the logarithm approximation when a better approximation is required:

$$\log(m!) = m\log m - m + \frac{1}{2}\log(2\pi m) + \frac{1}{12m} - \frac{1}{360m^3} + \frac{1}{1260m^5} - \cdots \tag{9}$$

This immediately gives us an asymptotic formula for the central binomial coefficient as:

▶ **Lemma 26.** *The central binomial coefficient can be approximated as:*

$$\binom{n}{n/2} = \sqrt{\frac{2}{\pi n}}2^n\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)$$

Cite Asymptopia

Now, we consider a "off-center" Binomial coefficient $\binom{n}{k}$ where $k = \frac{n+c\sqrt{n}}{2}$.

▶ **Lemma 27.** *Proof from [Spe14]*

$$\binom{n}{k} = \binom{n}{n/2}e^{-c^2/2}exp\left(\mathcal{O}(c^3/\sqrt{n})\right)$$

Proof. We consider the ratio: $R = \binom{n}{k}/\binom{n}{n/2}$:

$$R = \frac{\binom{n}{k}}{\binom{n}{n/2}} = \frac{(n/2)!(n/2)!}{k!(n-k)!} = \prod_{i=1}^{c\sqrt{n}/2}\frac{n/2-i+1}{n/2+i} \tag{10}$$

$$\implies \log R = \sum_{i=1}^{c\sqrt{n}/2}\log\left(\frac{n/2-i+1}{n/2+i}\right) \tag{11}$$

$$= \sum_{i=1}^{c\sqrt{n}/2} -\frac{4i}{n} + \mathcal{O}\left(\frac{i^2}{n^2}\right) = -\frac{c^2 n}{2n} + \mathcal{O}\left(\frac{(c\sqrt{n})^3}{n^2}\right) = -\frac{c^2}{2} + \mathcal{O}\left(\frac{c^3}{\sqrt{n}}\right) \tag{12}$$

$$\implies \binom{n}{k} = \binom{n}{n/2}e^{-c^2/2}exp\left(\mathcal{O}(c^3/\sqrt{n})\right) \tag{13}$$

□

A.2 Dyck Path Boundaries and Deviations

▶ **Lemma 28.** *Given a random walk of length $2n$ with exactly n up and down steps, consider a contiguous* sub-path *of length $2B$ that comprises of U up-steps and D down-steps i.e. $U + D = 2B$. Both $|B - U|$ and $|B - D|$ are $\mathcal{O}(\sqrt{B \log n})$ with probability at least $1 - 1/n^4$.*

Proof. We consider the random walk as a sequence of unbiased random variables $\{X_i\}_{i=1}^{2n} \in \{0,1\}^{2n}$ with the constraint $\sum_{i=1}^{2n} X_i = n$. Here, 1 corresponds to an up-step and 0 corresponds to a down step. Because of the constraint, X_i, X_j are negatively correlated for $i \neq j$ which allows us to apply Chernoff bounds. Now we consider a sub-path of length $2B$ and let U denote the sum of the X_is associated with this subpath. Using Chernoff bound with $\mathbb{E}[X] = B$, we get:

$$\mathbb{P}\left[|U - B| < 3\sqrt{B \log n}\right] = \mathbb{P}\left[|U - B| < 3\frac{\sqrt{\log n}}{\sqrt{B}}B\right] < e^{\frac{9 \log n}{3}} \approx \frac{1}{n^3}$$

Since U and D are symmetric, the same argument applies. □

▶ **Corollary 29.** *With high probability, every contiguous sub-path in the random walk (with U up and D down steps) satisfies the property with high probability. Specifically, if $U + D = 2B$, then $|B - U|$ and $|B - D|$ are upper bounded by $c\sqrt{B \log n}$ w.h.p. $1 - 1/n^2$ for all contiguous sub-paths (for some constant c).*

Proof. We can simply apply Lemma 28 and union bound over all n^2 possible contiguous sub-paths. □

▶ **Lemma 2.** *Consider a contiguous* sub-path *of a simple Dyck path of length $2n$ where the sub-path is of length $2B$ comprising of U up-steps and D down-steps (with $U + D = 2B$). Then there exists a constant c such that the quantities $|B - U|$, $|B - D|$, and $|U - D|$ are all $< c\sqrt{B \log n}$ with probability at least $1 - 1/n^2$ for every possible sub-path.*

Proof. As a consequence of Theorem 25, we can sample a Dyck path by first sampling a *balanced* random walk with n up steps and n down steps and adding an initial up step. We can then find the corresponding Dyck path by taking the unique cyclic shift that satisfies the Dyck constraint (after removing the initial up-step). Any interval in a cyclic shift is the union of at most two intervals in the original sequence. This affects the bound only by a constant factor. So, we can simply use Corollary 29 to finish the proof. Notice that since $|U - D| \leq |B - U| + |B - D|$, $|U - D| = \mathcal{O}(\sqrt{B \log n})$ comes for free. □

▶ **Lemma 3.** *Given a Dyck path sampling problem of length B with U up and D down steps with a bounary at k, there exists a constant c such that if $k > c\sqrt{B \log n}$, then the distribution of paths sampled without a boundary $\mathsf{C}_\infty(U, D)$ (hypergeometric sampling) is statistically $\mathcal{O}(1/n^2)$-close in L_1 distance to the distribution of Dyck paths $\mathsf{C}_k(U + D)$.*

Proof. We use $\mathcal{D}$ and $\mathcal{R}$ to denote the set of all valid Dyck paths and all random sequences respectively. Clearly, $\mathcal{D} \subseteq \mathcal{R}$. Let c be a constant satisfying Corollary 29. Since the random walk/sequence distribution is uniform on $\mathcal{R}$, and by Corollary 29 we see that at least $1 - 1/n^2$ fraction of the elements of $\mathcal{R}$ do not violate the boundary constraint. Therefore, $|\mathcal{D}| \geq (1 - 1/n^2)|\mathcal{R}|$ and so the L_1 distance between $\mathcal{U}_\mathcal{D}$ and $\mathcal{U}_\mathcal{R}$ is $\mathcal{O}(1/n^2)$. □

A.3 Computing Probabilities

Oracle for estimating probabilities:

Point to section referencing the left right/ total.

▶ **Lemma 30.** *Given a Dych sub-path problem within a global Dyck path of size $2n$ and a probability expression of the form $p_d = \frac{S_{left} \cdot S_{right}}{S_{total}}$, there exists a $\mathrm{poly}(\log n)$ time oracle that returns a $\left(1 \pm 1/n^2\right)$ multiplicative approximation to p_d if $p_d = \Omega(1/n^2)$ and returns 0 otherwise.*

Proof. We first compute a $1 + 1/n^3$ multiplicative approximation to $\ln p_d$. Using $\mathcal{O}(\log n)$ terms of the series in Equation 9, it is possible to estimate the logarithm of a factorial up to $1/n^c$ additive error. So, we can use the series expansion from Equation 9 up to $\mathcal{O}(\log n)$ terms. The additive error can also be cast as multiplicative since factorials are large positive integers.

The probability p_d can be written as an arithmetic expression involving sums and products of a constant number of factorial terms. Given a $1 \pm 1/n^c$ multiplicative approximation to $l_a = \ln a$ and $l_a = \ln b$, we wish to approximate $\ln(ab)$ and $\ln(a+b)$. The former is trivial since $\ln(ab) = \ln a + \ln b$. For the latter, we assume $a > b$ and use the identity $\ln(a+b) = \ln a + \ln(1 + b/a)$ to note that it suffices to approximate $\ln(1 + b/a)$. We define $\hat{l}_a = l_a \cdot (1 \pm \mathcal{O}(1/n^c))$ and $\hat{l}_b = l_b \cdot (1 \pm \mathcal{O}(1/n^c))$. In case $\hat{l}_b - \hat{l}_a < c \ln n \implies b/a < 1/n^c$, we approximate $\ln(a+b)$ by $\ln a$ since $\ln(1 + b/a) = \mathcal{O}(1/n^c)$ in this case. Otherwise, using the fact that $l_a - l_b = o(n^2)$, we compute:

$$1 + e^{\hat{l}_b - \hat{l}_a} = 1 + \frac{b}{a} \cdot e^{\mathcal{O}\left(\frac{l_b - l_a}{n^c}\right)} = 1 + \frac{b}{a} \cdot \left(1 \pm \mathcal{O}\left(\frac{1}{n^{c-2}}\right)\right) = \left(1 + \frac{b}{a}\right) \cdot \left(1 \pm \mathcal{O}\left(\frac{1}{n^{c-2}}\right)\right)$$

In other words, the value of c decreases every time we have a sunm operation. Since there are only a constant number of such arithmetic operations in the expression for p_d, we can set c to be a high enough constant (when approximating the factorials) and obtain the desired $1 \pm 1/n^3$ approximation to $\ln p_d$. If $\ln p_d < -3 \ln n$, we approximate $p_d = 0$. Otherwise, we can exponentiate the approximation to obtain $p_d \cdot e^{-\mathcal{O}(\ln n / n^3)} = p_d \left(1 \pm \mathcal{O}(1/n^2)\right)$. □

fix

A.4 Sampling the Height

- $d < c \cdot \sqrt{B} \log n$
- $k < c \cdot \sqrt{B} \log n \implies U - D < c \cdot \sqrt{B} \log n$
- $k' < c \cdot \sqrt{B} \log n$
- $B > \log^2 n \implies \sqrt{B} \log n < B$

▶ **Lemma 31.** *For $x < 1$ and $k \geq 1$,*

$$1 - kx < (1 - x)^k < 1 - kx + \frac{k(k-1)}{2} x^2.$$

▶ **Lemma 5.** $S_{left} \leq c_1 \frac{k \cdot \sqrt{\log n}}{\sqrt{B}} \cdot \binom{B}{D-d}$ *for some constant c_1.*

Proof. This involves some simple manipulations.

$$
\begin{aligned}
S_{left} &= \binom{B}{D-d} - \binom{B}{D-d-k} && (14)\\
&= \binom{B}{D-d} \cdot \left[1 - \frac{(D-d)(D-d-1)\cdots(D-d-k+1)}{(B-D-d+k)(B-D-d+k-1)\cdots(B-D-d+1)}\right] && (15)\\
&\leq \binom{B}{D-d} \cdot \left[1 - \left(\frac{D-d-k+1}{B-D+d+k}\right)^k\right] && (16)\\
&\leq \binom{B}{D-d} \cdot \left[1 - \left(\frac{U+d+k-(U-D+d+k-1)}{U+d+k}\right)^k\right] && (17)\\
&\leq \binom{B}{D-d} \cdot \left[1 - \left(\frac{U+d+k-\mathcal{O}(\sqrt{B\log n})}{U+d+k}\right)^k\right] && (18)\\
&\leq \Theta\left(\frac{k\sqrt{\log n}}{\sqrt{B}}\right) \cdot \binom{B}{D-d} && (19)
\end{aligned}
$$

□

▶ **Lemma 6.** *$S_{right} < c_2 \frac{k'\cdot\sqrt{logn}}{\sqrt{B}} \cdot \binom{B}{U-d}$ for some constant c_2.*

Proof.

$$
\begin{aligned}
S_{right} &= \binom{B}{U-d} - \binom{B}{U-d-k'} && (20)\\
&= \binom{B}{U-d} \cdot \left[1 - \frac{(U-d)(U-d-1)\cdots(U-d-k'+1)}{(B-U-d+k')(B-U-d+k'-1)\cdots(B-U-d+1)}\right] && (21)\\
&\leq \binom{B}{U-d} \cdot \left[1 - \left(\frac{U-d-k'+1}{B-U+d+k'}\right)^{k'}\right] && (22)\\
&\leq \binom{B}{U-d} \cdot \left[1 - \left(\frac{2D-U-d-k+1}{2U-D+k+d}\right)^{k'}\right] && (23)\\
&\leq \binom{B}{U-d} \cdot \left[1 - \left(\frac{U+k+d-(2U-2D+2d+2k-1)}{U+k+d}\right)^{k'}\right] && (24)\\
&\leq \binom{B}{U-d} \cdot \left[1 - \left(\frac{U+k+d-\mathcal{O}(\sqrt{B\log n})}{U+k+d}\right)^{k'}\right] && (25)\\
&\leq \Theta\left(\frac{k'\sqrt{\log n}}{\sqrt{B}}\right) \cdot \binom{B}{U-d} && (26)
\end{aligned}
$$

□

▶ **Lemma 32.** $S_{tot} \geq \binom{2B}{2D} \cdot \left[1 - \left(1 - \frac{k'}{2U+1}\right)^k\right]$.

change statement

Proof.

$$S_{tot} = \binom{2B}{2D} - \binom{2B}{2D-k} \tag{27}$$

$$= \binom{2B}{2D} \cdot \left[1 - \frac{(2D)(2D-1)\cdots(2D-k+1)}{(2B-2D+k)(2B-2D+k-1)\cdots(2B-2D+1)}\right] \tag{28}$$

$$\geq \binom{2B}{2D} \cdot \left[1 - \left(\frac{2D-k+1}{2B-2D+1}\right)^k\right] \tag{29}$$

$$\geq \binom{2B}{2D} \cdot \left[1 - \left(\frac{2U-(2U-2D+k-1)}{2U+1}\right)^k\right] \tag{30}$$

$$\geq \binom{2B}{2D} \cdot \left[1 - \left(\frac{(2U+1)-k'}{2U+1}\right)^k\right] \tag{31}$$

$$\geq \binom{2B}{2D} \cdot \left[1 - \left(1 - \frac{k'}{2U+1}\right)^k\right] \tag{32}$$

$$\tag{33}$$

□

Reference previous lemma

▶ **Lemma 4.** *When $kk' > 2U+1$, $S_{total} > \frac{1}{2} \cdot \binom{2B}{2D}$.*

Proof. When $kk' > 2U+1 \implies k > \frac{2U+1}{k'}$, we will show that the above expression is greater than $\frac{1}{2}\binom{2B}{2D}$. Defining $\nu = \frac{2U+1}{k'} > 1$, we see that $(1-\frac{1}{\nu})^k \leq (1-\frac{1}{\nu})^\nu$. Since this is an increasing function of ν and since the limit of this function is $\frac{1}{e}$, we conclude that

$$1 - \left(1 - \frac{k'}{2U+1}\right)^k > \frac{1}{2}$$

□

▶ **Lemma 7.** *When $kk' \leq 2U+1$, $S_{total} < c_3 \frac{k \cdot k'}{B} \cdot \binom{2B}{2D}$ for some constant c_3.*

Proof. Now we bound the term $1 - \left(1 - \frac{k'}{2U+1}\right)^k$, given that $kk' \leq 2U+1 \implies \frac{kk'}{2U+1} \leq 1$. Using Taylor expansion, we see that

$$1 - \left(1 - \frac{k'}{2U+1}\right)^k \tag{34}$$

$$\leq \frac{kk'}{2U+1} - \frac{k(k-1)}{2} \cdot \frac{k'^2}{(2U+1)^2} \tag{35}$$

$$\leq \frac{kk'}{2U+1} - \frac{k^2k'^2}{2(2U+1)^2} \tag{36}$$

$$\leq \frac{kk'}{2U+1}\left(1 - \frac{kk'}{2(2U+1)}\right) \tag{37}$$

$$\leq \frac{kk'}{2(2U+1)} \leq \frac{kk'}{\Theta(B)} \tag{38}$$

$$\tag{39}$$

□

A.5 First Return Sampling

▶ **Corollary 13.** *When* $d \notin \mathcal{R}$, $S_{left}(d) \cdot S_{right}(d) = \Theta\left(\frac{2^{U+D}}{\sqrt{d(U+D-2d-k)}} \cdot e^{-r(d)} \cdot \frac{k-1}{d+k-1} \cdot \frac{U-D+k}{U-d+1}\right)$ *where* $r(d) = \mathcal{O}(\log^2 n)$.

Proof. This follows from the fact that both $r_{left}(d)$ and $r_{right}(d)$ are $\mathcal{O}(\log^2 n)$. □

▶ **Lemma 11.** *If* $d > \log^4 n$, *then* $S_{left}(d) = \Theta\left(\frac{2^{2d+k}}{\sqrt{d}} e^{-r_{left}(d)} \cdot \frac{k-1}{d+k-1}\right)$ *where* $r_{left}(d) = \frac{(k-2)^2}{2(2d+k-2)}$. *Furthermore,* $r_{left}(d) = \mathcal{O}(\log^2 n)$.

Proof. In what follows, we will drop constant factors: Refer to Figure 5 for the setup. The left section of the path reaches one unit above the boundary (the next step would make it touch the boundary). The number of up-steps on the left side is d and therefore the number of down steps must be $d+k-2$. This inclues d down steps to cancel out the upwards movement, and $k-2$ more to get to one unit above the boundary. The boundary for this section is $k' = k-1$. This gives us:

$$S_{left}(d) = \binom{2d+k-2}{d} - \binom{2d+k-2}{d-1} \tag{40}$$

$$= \binom{2d+k-2}{d}\left[1 - \frac{d}{d+k-1}\right] = \binom{2d+k-2}{d}\frac{k-1}{d+k-1} \tag{41}$$

Now, letting $z = 2d+k-2$, we can write $d = \frac{z-(k-2)}{2} = \frac{z-\frac{k-2}{\sqrt{z}}\sqrt{z}}{2}$. Using Lemma 2, we see that $\frac{k-2}{\sqrt{z}}$ should be $\mathcal{O}(\sqrt{\log n})$. If this is not the case, we can simply return 0 because the probability associated with this value of d is negligible. Since $z > \log^4 n$, we can apply Lemma 27 to get:

$$S_{left}(d) = \Theta\left(\binom{z}{z/2} e^{\frac{(k-2)^2}{2z}} \frac{k-1}{d+k-1}\right) = \Theta\left(\frac{2^{2d+k}}{\sqrt{d}} e^{\frac{(k-2)^2}{2(2d+k-2)}} \frac{k-1}{d+k-1}\right)$$

□

▶ **Lemma 12.** *If* $U+D-2d-k > \log^4 n$, *then* $S_{right}(d) = \Theta\left(\frac{2^{U+D-2d-k}}{\sqrt{U+d-2d-k}} e^{-r_{right}(d)} \cdot \frac{U-D+k}{U-d+1}\right)$ *where* $r_{right}(d) = \frac{(U-D-k-1)^2}{4(U+D-2d-k+1)}$. *Furthermore,* $r_{right}(d) = \mathcal{O}(\log^2 n)$.

Proof. The right section of the path starts from the original boundary. Consequently, the boundary for this section is at $k' = 1$. The number of up-steps on the right side is $U-d$ and the number of down steps is $D-d-k+1$. This gives us:

$$S_{right}(d) = \binom{U+D-2d-k+1}{U-d} - \binom{U+D-2d-k+1}{U-d+1} \tag{42}$$

$$= \binom{U+D-2d-k+1}{U-d}\left[1 - \frac{D-d-k-1}{U-d+1}\right] \tag{43}$$

$$= \binom{U+D-2d-k+1}{U-d}\frac{U-D+k}{U-d+1} \tag{44}$$

Now, letting $z = U+D-2d-k+1$, we can write $U-d = \frac{z+(U-D+k-1)}{2} = \frac{z+\frac{U-D+k-1}{\sqrt{z}}\sqrt{z}}{2}$. Using Lemma 2, we see that $\frac{k-2}{\sqrt{z}}$ should be $\mathcal{O}(\sqrt{\log n})$. If this is not the case, we can simply return 0

because the probability associated with this value of d is negligible. Since $z > \log^4 n$, we can apply Lemma 27 to get:

$$S_{right}(d) = \Theta\left(\binom{z}{z/2} e^{\frac{(U-D+k-1)^2}{2z}} \frac{U-D+k}{U-d+1}\right) \tag{45}$$

$$= \Theta\left(\frac{2^{U+D-2d-k}}{\sqrt{U+D-2d-k}} e^{\frac{(U-D+k-1)^2}{2(U+D-2d-k+1)}} \frac{U-D+k}{U-d+1}\right) \tag{46}$$

□