Asymptotic Behavior of Sequence Models

Flavio Chierichetti*
Dipartimento di Informatica
Sapienza University of Rome
flavio@di.uniroma1.it

Ravi Kumar
Google
Mountain View, CA
ravi.k53@gmail.com

Andrew Tomkins
Google
Mountain View, CA
atomkins@gmail.com

ABSTRACT

In this paper we study the limiting dynamics of a sequential process that generalizes Pólya's urn. This process has been studied also in the context of language generation, discrete choice, repeat consumption, and models for the web graph. The process we study generates future items by copying from past items. It is parameterized by a sequence of weights describing how much to prefer copying from recent versus more distant locations. We show that, if the weight sequence follows a power law with exponent $\alpha \in [0, 1)$, then the sequences generated by the model tend toward a limiting behavior in which the eventual frequency of each token in the alphabet attains a limit. Moreover, in the case $\alpha > 2$, we show that the sequence converges to a token being chosen infinitely often, and each other token being chosen only constantly many times.

CCS CONCEPTS

• **Mathematics of computing** → **Stochastic processes**; • **Applied computing** → *Law, social and behavioral sciences*; • **Information systems** → *Web mining*.

KEYWORDS

urn processes, power laws, copying models

ACM Reference Format:
Flavio Chierichetti, Ravi Kumar, and Andrew Tomkins. 2020. Asymptotic Behavior of Sequence Models. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3366423.3380044

1 INTRODUCTION

In this paper we are concerned with a randomized process to produce sequences over a fixed alphabet $\{1, \ldots, T\}$ of tokens. The process begins with some finite initial history of tokens, and then proceeds by randomly selecting a previous location in the history to copy from, in order to produce the next output. The most recent location has a preference weight w_1, the second most recent location has weight w_2, and so forth; more recent locations are preferred. The location from i steps ago is copied-from with probability proportional to w_i.

This simple process occurs in many settings:

- For the case with all weights uniform, $w_i = 1$, the process is exactly the classical Pólya's urn [11], with the initial contents of the urns given by the counts of each token in the history.
- If new tokens are introduced with constant probability, but the weights remain uniform, the process is exactly Simon's 1955 copying model [12] used to explain word frequencies in human language.
- If the tokens correspond to graph vertices, and new tokens are again introduced with constant probability, the model is an important special case of graph copying models [3, 8].
- If the sequence of weights w_i in the model are learned, the resulting model has been used to explain repeat consumption behavior in multiple domains [1, 4, 9].

Perhaps the most fundamental question about this model is: what happens when it runs? In this paper, we wish to understand the limiting behaviors of sequences produced by models like this, especially the following central questions:

(1) As the length of the generated sequence grows, does it reach a limiting distribution under some definition?
(2) When the limiting distribution exists, does it have positive support over the entire token set, or do certain tokens disappear forever?
(3) What can be said about the relationship between the attention weights and the limiting distribution?

1.1 Our results

We assume that our model is given some fixed prefix or history $x_1 x_2 \ldots x_h$, and then repeatedly predicts x_i given $x_1 \ldots x_{i-1}$. For $i > h$, every element x_i copies directly or indirectly from some position in the history. To study limiting distributions, we will fix some arbitrary subset of positions of interest in the history, and state our results in terms of the long-term occurrences of tokens copied from these positions. First, some definitions:

For $i > h$, let X_i be an indicator that is 1 if x_i is copied, directly or indirectly, from a "position of interest" and 0 otherwise, and let

$$Z_i = \frac{1}{i-h} \sum_{j=h+1}^{i} X_j.$$

Let $Z^* = \lim_{i \to \infty} Z_i$. We want to know when Z^* exists, and what are its properties.

The following two results are already known:

(1) For $w_i = 1$, Z^* exists and is beta-distributed (Pólya urn).
(2) For $w_i = 2^{-i}$, Z^* exists and has support only on $\{0, 1\}$ [1].

We show the following new results:

(1) For $w_i = i^{-\alpha}$ with $0 < \alpha < 1$, Z^* exists.
(2) For $w_i = i^{-\alpha}$ with $\alpha > 2$, Z^* exists and has support only on $\{0, 1\}$.

*Work done in part while visiting Google. Supported in part by the ERC Starting Grant DMAP 680153, by a Google Focused Research Award, by the PRIN project 2017K7XPAN, by BiCi — Bertinoro international Center for informatics, and by the "Dipartimenti di Eccellenza" grant awarded to the Dipartimento di Informatica at Sapienza.

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380044

To obtain such convergence results, it turns out that our assumption that the weights follow a power law is, in a sense, important. We show non-convergence results (omitted in this version) when the weights may not obey a power law but satisfy weaker analytic properties such as monotonicity or convexity.

2 RELATED WORK

Urn processes have been studied by classical mathematicians for hundreds of years, but the standard Pólya's urn formulation was developed and analyzed by Eggenberger and Pólya in 1923 [5]. Since then, a number of variations have been proposed, introducing complex rules for modifying the state of balls in urns in response to each draw. However, these processes are designed to be characterized by the state of the urns, so the numerous extensions do not typically consider rules that depend on the order of past draws.

Herbert Simon [12] introduced a sequential process to study the emergence of power laws in language, in partial response to work of Zipf [14] six years earlier. In Simon's model, during each timestep, the next token will be copied with some probability from a uniformly-selected past location, and with remaining probability, will be a previously-unseen character. The continued introduction of neologisms into the vocabulary matches observations of natural language text. The process is known to converge in the limit to a distribution over token frequencies that matches a power law, again corresponding to natural language text.

Related to Simon's model, a number of authors [3, 8] developed sequential models for the evolution of graphs, intended to reproduce the power law in-degree distribution observed for the web graph [10]. These models also produced new links by selecting existing links to copy from, using a uniform distribution.

Both Simon's copying model and the models of graph evolution relied on the introduction of new vocabulary as the model evolved, and their analysis was fundamentally structured around a growing set of tokens; hence, while the models are similar, the analytical techniques are not applicable to our domain.

More recently, Anderson et al. [1] employed copying models in the context of reconsumption of items: a user might listen to the same song many times, or eat at the same restaurant, and these decisions were shown to be well-modeled by a process that selects an item to re-consume by copying a previous consumption from the past. In their examples, and in follow-on work [4], the preference weights for the ith previous item w_i were learned through maximum likelihood estimation. In this domain as in natural language, the particular form of the resulting weights is approximated well by a power law.

Other than the Pólya's urn, the only work of which we are aware that studies the limiting behavior of processes of this form is the work of Anderson et al.[1], who show that the weight sequence $w_i = 2^{-i}$ leads to a limiting distribution in which all tokens but one eventually disappear, leaving a single "winner" who will then be copied forever. In practice, we are not aware of real-world datasets in which the weights decay exponentially, hence our interest in extending this result to power law distributions, beyond $\alpha = 0$ case of Pólya's urn.

3 BACKGROUND

Let $1 \geq w_1 \geq w_2 \geq \cdots \geq w_i \geq \cdots$ be the given list of weights. Let $T = \{1, 2, \ldots, T\}$ be a finite alphabet of tokens. Let $W_k = \sum_{i=1}^{k} w_i$ be the k-prefix sum.

Let $x_1, x_2, \ldots, x_h$ be a fixed *history*, where $h > 0$ and each $x_i \in [T]$. Given the weights and history, the model generates an infinite sequence from T according to the following rule:

$$\Pr[x_{i+1} = t] = \frac{\sum_{j=1}^{i} w_j \cdot [x_{i-j+1} = t]}{\sum_{j=1}^{i} w_j},$$

for $i \geq h$. Here $[\cdot]$ is the binary indicator function. The model thus captures the process of extending the sequence by randomly choosing a position from the past according to the weights and *copying* the token in that position. Since the weights are monotonically non-decreasing, tokens from more recent past have higher chance of being copied than tokens from the distant past.

We say that h is the end of the history. Let $i \geq h$. By definition of the process, any position i will copy from a random position less than i, according to the weights. In this case we use $c(i)$ to denote the *position* that i copies from, and $c^{(j+1)}(i)$ to denote $c(c^{(j)}(i))$, with $c^{(1)}(i) = c(i)$. We let

$$\begin{aligned} C_0(i) &= \{i\}, \\ C_1(i) &= \{i, c(i)\}, \\ C_2(i) &= \{i, c(i), c^{(2)}(i)\}, \end{aligned}$$

and so on. If $\ell(i)$ is the smallest integer such that $c^{(\ell(i))} \leq h$, we let

$$C(i) = C_{\ell(i)}(i),$$

and

$$f(i) = c^{(\ell(i)-1)}(i).$$

In other words, if we treat the sequence of copies that ended in position i as a chain, then $C(i)$ is the set of positions along this chain and $f(i)$ is the *final token* outside the history in the chain starting from i. An equivalent interpretation is to consider a walk starting at i that jumps to position $c(i)$, then jumps to position $c^{(2)}(i)$, and so on.

Definition 3.1 (Collision). We say that positions i, j, with $h \leq i < j$, *collide* if and only $(C(i) \cap C(j)) \setminus [h]$ is non-empty, i.e., equivalently if and only if $f(i) = f(j)$.

We use the following random variable to denote the collision event:

$$\xi_{i,j} \stackrel{\Delta}{=} \text{“}f(i) = f(j)\text{”}.$$

We will be using Chebyshev's inequality and the Chernoff bound.

Theorem 3.2 (Chebyshev's inequality). *Let X be a random variable with finite expectation and variance. Then, for each $c > 0$*

$$\Pr\left[|X - \mathrm{E}[X]| \leq c \cdot \sqrt{\mathrm{Var}[X]}\right] \leq c^{-2}.$$

Theorem 3.3 (Chernoff bound). *Let $X_1, \ldots, X_n$ be iid 0/1 random variables. Let $X = \sum X_i$. Then, for $\delta > 0$,*

$$\Pr[X < (1 - \delta)E[X]] \leq e^{-\delta^2 E[X]/2}.$$

Moreover,

$$\Pr[X > (1 + \delta)E[X]] \leq e^{-\min(\delta, \delta^2)E[X]/3}.$$

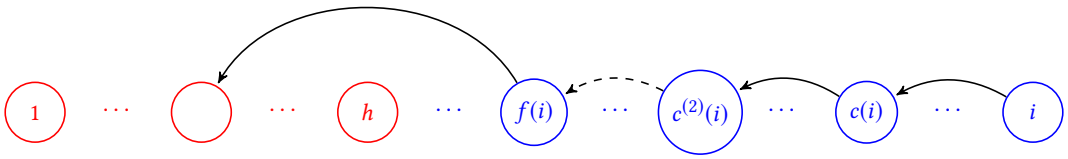

Figure 1: A walk from i. History is shown in red.

4 CONVERGENCE TO A LIMIT: $\alpha \in [0, 1)$

In this section we show our main result: convergence to a limit. We will show this in two steps: (i) $\xi_{i,j}$ vanishes as h increases assuming $w_i = i^{-\alpha}$, for some $0 \le \alpha < 1$ and (ii) a delicate covariance analysis that uses the vanishing collisions to establish convergence.

4.1 Vanishing collisions

In this section we show that $\xi_{i,j} < o_h(1)$, i.e., the collisions vanish with history size. To simplify the exposition, we first introduce the following notation. Let ξ_i be the event that the jump from i ends in some position in $\{1, 2, \ldots, \lfloor i/2 \rfloor\}$, i.e.,

$$\xi_i \stackrel{\Delta}{=} \text{“}c(i) \in [\lfloor i/2 \rfloor]\text{”}.$$

We prove something slightly more general.

Lemma 4.1. *If $\alpha \in [0, 1)$, then $\Pr[\xi_i] \ge \Omega(1)$.*

Proof. Indeed, we have that

$$\sum_{j=1}^{i} j^{-\alpha} = \Theta\left(i^{1-\alpha}\right),$$

and that,

$$\sum_{j=\lceil i/2 \rceil}^{i} j^{-\alpha} = \Theta\left(i^{1-\alpha}\right).$$

Thus, $\Pr[\xi_i] = \Theta(1)$. □

We use this to prove that, if $\alpha < 1$, then the number of steps from a generic position j to a position less than h, i.e., the length $|C(i)|$ of the chain from i is at most $O(\log j)$ with high probability.

Theorem 4.2. *Suppose that $0 < \alpha < 1$, and fix $h \ge 1$. Then, there exists a constant $c = c(\alpha)$, such that if we let $n_j = n_j(i)$, $j \le i$, be the number of positions in the range $\{h+1, h+2, \ldots, j\}$ in the chain starting in i, we have that*

$$\Pr[\exists j \in \{h+1, \ldots, i\} \mid n_j(i) \ge c \cdot \ln(j+1)] \le O(h^{-10}).$$

Proof. In each position j, the probability that the event ξ_j happens is at least $p = p(\alpha) = \Theta(1)$. Therefore, by the Chernoff bound, the probability that the event does not happen at least $\lg j$ times in $200\frac{1}{p}\ln j$ trials is at most j^{-11}. Therefore, the probability that the chain does not reach some position before h after having visited $200\frac{1}{p}\ln j$ positions is at most j^{-11}. By the union bound, we get that the probability that there exists some $j \ge h+1$ that does not reach the history after $200\frac{1}{p}\ln j$ steps is at most

$$\sum_{j \ge h+1} j^{-11} \le O(h^{-10}). \quad \square$$

Using these, we obtain the desired result on vanishing collisions.

Theorem 4.3. *Let $\alpha \in [0, 1)$ and $h < i < i'$ be given. Then, $\Pr[\xi_{i,i'}] \le o_h(1)$.*

Proof. By Theorem 4.2, with probability $1 - o_h(1)$, for each $j \ge h$, the walk from i (called i-walk) will visit at most $O(\ln(j+1))$ positions in the range $\{h+1, \ldots, j\}$. Now, consider the walk from i' (called i'-walk) and let $j > h$ be a generic position visited by the i'-walk. Assuming that j has not also been visited by the i-walk, we ask: what is the probability q that the position that the i'-walk jumps from j has not been visited by the i-walk? Observe that, with probability $1 - o(1)$, the i-walk will have visited at most $O(\ln(j+1)$ positions in the range $\{1, \ldots, j-1\}$. Then, under the conditioning, the probability that j jumps on some position visited by the i-walk is at most

$$\frac{O(\ln(j+1)) \cdot 1^{-\alpha}}{\sum_{\ell=1}^{j-1} \ell^{-\alpha}} \le O(j^{\alpha-1} \cdot \ln(j+1)).$$

Since, by Theorem 4.2, with probability $1 - o_h(1)$, the i'-walk will visit at most $O(\ln(j+1))$ positions in the range $\{j, \ldots, 2j\}$, we have

$$\begin{aligned}\Pr[\xi_{i,i'}] &\le 2 \cdot o_h(1) + \sum_{r=\lg h}^{\ln i'} O((2^r)^{\alpha-1} \cdot \ln(2^r+1)^2) \\ &\le o_h(1). \end{aligned}$$ □

4.2 Convergence via bounded covariance

In this section we establish the convergence to the limit for $w_i = i^{-\alpha}$, for $\alpha \in [0, 1)$.[1]

Fix arbitrarily a set $H \subseteq [h]$. For $k \in \{h+1, h+2, \ldots, t\}$, let X_k be 1 if position k ultimately copies from some position in H, and 0 if position k ultimately copies from some position in $[h] \setminus H$. Later we will choose H to be the set of positions in the history that contain a given token.

The idea behind the analysis is to bound the variance of the sum of X_i's. This is not immediate since the X_i's are not pairwise independent. To handle this, we focus on the correlation between X_i and X_j and show that the covariance is vanishing. Once we establish this, the variance bound is relatively easy and the convergence result follows by appealing to the Chebyshev's inequality.

First we analyze the covariance of X_i and X_j.

Lemma 4.4. *If $h < i < j$, it holds $\mathrm{Cov}[X_i, X_j] < o_h(1)$.*

[1] We point out that the result in this section does not require the weights to follow a power law — it only requires the vanishing collisions property of the chosen weights. We have proved this property for $w_i = i^{-\alpha}$, $\alpha < 1$, in the previous section.

Proof. We have that $\mathrm{Cov}[X_i, X_j] = \mathrm{E}[X_i X_j] - \mathrm{E}[X_i]\,\mathrm{E}[X_j]$. We aim to split the probability space into $\xi \oplus \overline{\xi}$, where $\xi = \xi_{i,j}$ is the event "$f(i) = f(j)$". We apply the law of total covariance, to get

$$\begin{aligned}\mathrm{Cov}[X_i, X_j] &= \left(\mathrm{E}[X_i X_j \mid \xi] - \mathrm{E}[X_i \mid \xi] \cdot E[X_j \mid \xi]\right) \cdot \Pr[\xi] + \\ &\quad + \left(\mathrm{E}[X_i X_j \mid \overline{\xi}] - \mathrm{E}[X_i \mid \overline{\xi}] \cdot E[X_j \mid \overline{\xi}]\right) \cdot \Pr[\overline{\xi}] \\ &\le \Pr[\xi] + \left(\mathrm{E}[X_i X_j \mid \overline{\xi}] - \mathrm{E}[X_i \mid \overline{\xi}] \cdot E[X_j \mid \overline{\xi}]\right) \\ &\le o_h(1) + \left(\mathrm{E}[X_i X_j \mid \overline{\xi}] - \mathrm{E}[X_i \mid \overline{\xi}] \cdot E[X_j \mid \overline{\xi}]\right),\end{aligned}$$

where the first inequality follows from $0 \le X_i, X_j, \Pr[\overline{\xi}] \le 1$, while the second inequality follows from Theorem 4.3. It remains to bound the second term.

Observe that, if $\overline{\xi}$ happens, i.e., if $f(i) \neq f(j)$, then the walks from i and from j will not meet in any position larger than h. Therefore,

$$\begin{aligned}&\mathrm{E}[X_i X_j \mid \overline{\xi}] - \mathrm{E}[X_i \mid \overline{\xi}] \cdot E[X_j \mid \overline{\xi}] \\ &= \sum_{i'=h+1}^{i} \sum_{\substack{j'=h+1 \\ j' \neq i'}}^{j} \Big(\Pr[f(i) = i' \text{ and } f(j) = j' \mid \overline{\xi}] \\ &\quad \cdot \Big(\mathrm{E}[X_i X_j \mid \overline{\xi}, f(i) = i', f(j) = j'] \\ &\qquad - \mathrm{E}[X_i \mid \overline{\xi}, f(i) = i', f(j) = j'] \\ &\qquad\quad \cdot \mathrm{E}[X_j \mid \overline{\xi}, f(i) = i', f(j) = j']\Big)\Big) \\ &= \sum_{i'=h+1}^{i} \sum_{\substack{j'=h+1 \\ j' \neq i'}}^{j} \Big(\Pr[f(i) = i' \text{ and } f(j) = j' \mid \overline{\xi}] \\ &\quad \cdot \left(\mathrm{E}[X_i X_j \mid f(i) = i', f(j) = j'] \right. \\ &\qquad - \mathrm{E}[X_i \mid f(i) = i', f(j) = j'] \\ &\qquad\quad \left. \cdot \mathrm{E}[X_j \mid f(i) = i', f(j) = j']\right).\end{aligned}$$

Now, under the conditioning $f(i) = i'$ and $f(j) = j'$ (with $i' \neq j'$), we first claim that X_i and X_j are independent. Indeed, under that conditioning, X_i is 1 if and only if $c(f(i)) \in H$, and X_j is 1 if and only if $c(f(j)) \in H$. Since $f(i) \neq f(j)$, the random position that $f(i)$ copies from (i.e., $c(f(i))$) is independent of the random position that $f(j)$ copies from (i.e., $c(f(j))$); this shows X_i and X_j are independent.

It follows that

$$\begin{aligned}&\mathrm{E}[X_i X_j \mid f(i) = i', f(j) = j'] = \\ &\quad \mathrm{E}[X_i \mid f(i) = i', f(j) = j'] \cdot \mathrm{E}[X_j \mid f(i) = i', f(j) = j'].\end{aligned}$$

Thus,

$$\begin{aligned}&\mathrm{E}[X_i X_j \mid \overline{\xi}] - \mathrm{E}[X_i \mid \overline{\xi}] \cdot E[X_j \mid \overline{\xi}] \\ &= \sum_{i'=h+1}^{i} \sum_{\substack{j'=h+1 \\ j' \neq i'}}^{j} \left(\Pr[f(i) = i' \wedge f(j) = j' \mid \overline{\xi}] \cdot 0\right) = 0.\end{aligned}$$

This establishes $\mathrm{Cov}[X_i, X_j] \le o_h(1)$. □

For $i > h$, let

$$Y_i = \sum_{k=h+1}^{i} X_k,$$

be the number of positions in the range $h+1, \ldots, i$ that ultimately copy from some position in H. We now bound the variance of this random variable using the bound on the covariance that we just established.

Lemma 4.5. *For $i > h$, it holds* $\mathrm{Var}[Y_i] < o_h(i^2)$.

Proof. Let $P = \{h+1, h+2, \ldots, i\}$. By linearity of expectation, we have that $E[Y_i] = \sum_{k \in P} X_k$. Moreover,

$$\mathrm{Var}[Y_i] = \mathrm{Var}\left[\sum_{k \in P} X_k\right] = \sum_{k \in P} \mathrm{Var}[X_k] + 2 \sum_{\{k,k'\} \in \binom{P}{2}} \mathrm{Cov}[X_k, X_{k'}].$$

Since $0 \le X_k \le 1$, we have $\mathrm{Var}[X_k] \le \frac{1}{4}$. Then,

$$\begin{aligned}\mathrm{Var}[Y_i] &\le \frac{|P|}{4} + 2 \sum_{\{k,k'\} \in \binom{P}{2}} \mathrm{Cov}[X_k, X_{k'}] \\ &\le \frac{|P|}{4} + |P|^2 o_h(1) \\ &\le \frac{i}{4} + o_h(i^2) \\ &\le \frac{i^2}{4h} + o_h(i^2) \\ &= o_h(i^2).\end{aligned}$$

□

With a bound on the variance, we apply Chebyshev's inequality to get the convergence result.

Theorem 4.6. *It holds that*

$$\Pr[|Y_i - E[Y_i]| > o_h(i)] \le o_h(1).$$

Analogously, if we let $Z_i = Y_i / i$, it holds that

$$\Pr[|Z_i - E[Z_i]| > o_h(1)] \le o_h(1).$$

For a large enough h, we can then apply the union bound on each (of the constantly many) tokens, to get the following.

Corollary 4.7. *Suppose that there are T tokens (with $T = O(1)$). Condition on the sequence of the history to be some $\sigma \in [T]^h$, and let $\overline{Z}_i = (Z_i(1), Z_i(2), \ldots, Z_i(T))$ be the vector containing in its tth position the fraction of occurrences of token t at time $i > h$. Then, $|Z_i|_1 = 1$, and Z_i converges to $E[Z_i \mid \sigma]$, i.e.,*

$$\Pr[|Z_i - E[Z_i \mid \sigma]|_\infty > o_h(1) \mid \sigma] < o_h(1).$$

That is, after having run the process for h steps, the vector of occurrences at any later step will be concentrated around its expectation.

4.3 Simulations

We run the process up until a position h, making up the history. Then, keeping the resulting history fixed, we repeatedly, and independently, run the process from that history up until time $10h$, keeping track of the final fraction of occurrences of a given token. In Figure 2, we plot the empirical distribution of the fraction of occurrences of that token, for $h = 100, 1000$. While the expectation is random (it strongly depends on what happens in the h steps of history), once we condition on the first h steps, the final distribution is more concentrated as h becomes larger.

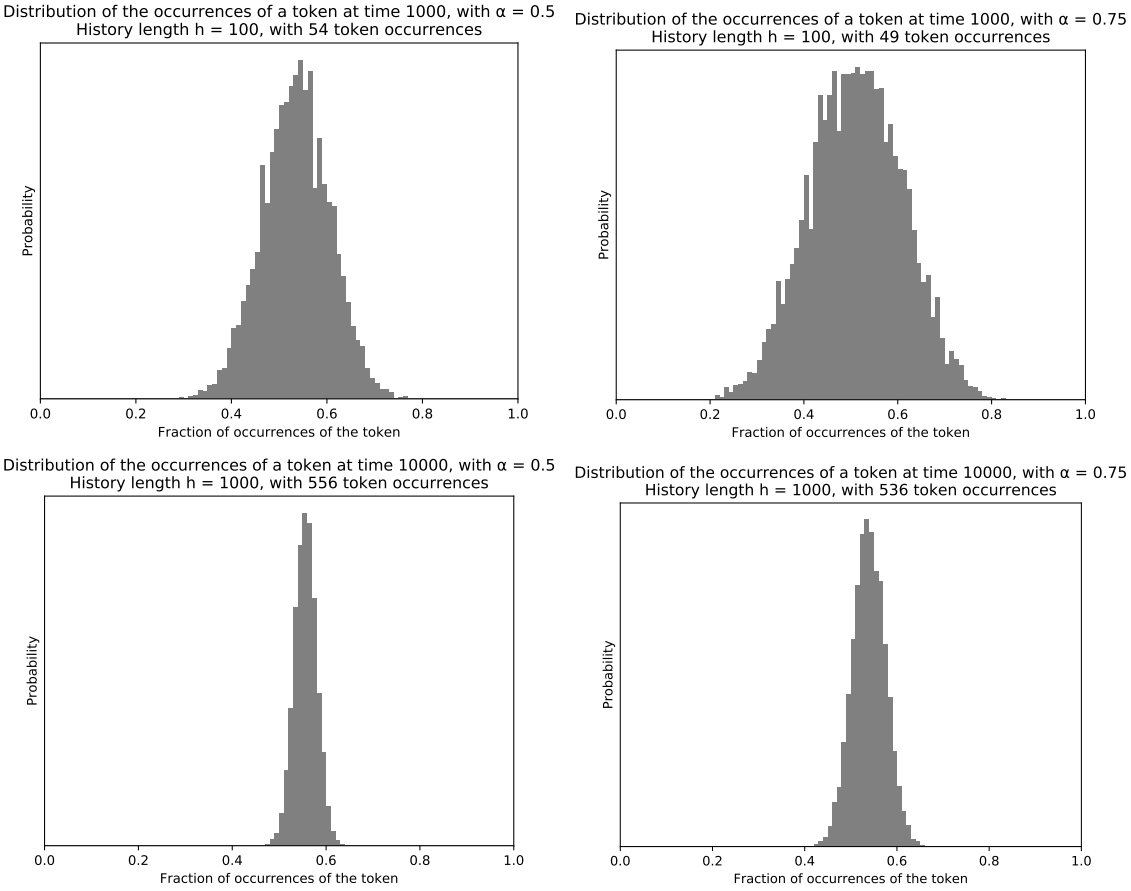

Figure 2: The empirical distributions (averaged over 10K runs) of the fraction of occurrences of a token, starting from a fixed random history of length $h \in \{100, 1000\}$ (first and second row), and continuing until time $10h$, with $\alpha \in \{0.5, 0.75\}$ (first and second column). As the length of the history increases, the distribution becomes more concentrated around its expectation.

Moreover, in Figure 3, we plot the empirical probabilities of reaching a specific position in a history of length $h = 100$, for various α's, and from various starting points.

5 SINGLE WINNER: $\alpha > 2$

In this section we show that if the weights w_i follow a power law with exponent greater than 2, then with probability 1 the process will converge to a "single-winner" limit.

Theorem 5.1. *If $w_i = i^{-\alpha}$, for $i \geq 1$, $\alpha > 2$, then the limit Z^* exists and is supported on $\{0, 1\}$ with probability* 1.

Proof. As in [1], we study the probability that, starting from a given position $i + 1$, all positions copy from some position greater than or equal i. If this happens, then all the positions greater than i will end up copying from position i, and therefore all positions greater than or equal i will end up containing the same token.

We use the same approach in [1], but generalized to $w_i = i^{-\alpha}$, as follows:

(i) Let the process go on for some number of steps i.

(ii) Fix $j \geq 1$. Then, the probability that the position $i + j$ copies from some position in $\{i, i + 1, \ldots, i + j - 1\}$ is at least

$$p_j = \frac{\sum_{k=1}^{j} k^{-\alpha}}{\zeta(\alpha)} \geq \frac{\zeta(\alpha) \cdot (1 - O(j^{1-\alpha}))}{\zeta(\alpha)} = 1 - O(j^{1-\alpha}) > 0,$$

where $\zeta(\cdot)$ is the Riemann zeta function.

(iii) Therefore, the probability that for all $j \geq 1$, position $i + j$ copies from some position in $\{i, \ldots, i + j - 1\}$ is at least

$$\begin{aligned} q_j &\geq \prod_{j=1}^{\infty}(1 - O(j^{1-\alpha})) = \prod_{j=1}^{d}(1 - O(j^{1-\alpha})) - \sum_{j=d+1}^{\infty} O(j^{1-\alpha}) \\ &\geq \prod_{j=1}^{d}(1 - O(j^{1-\alpha})) - O(d^{2-\alpha}), \end{aligned}$$

where d can be chosen arbitrarily. In fact one can show that, for each $\alpha > 2$, it is possible to choose d so that the probability q_j is at least a constant $c(\alpha) > 0$.

In other words, with constant probability (bounded away from 0 and 1), all positions greater than i will end up with the same token that is in position i.

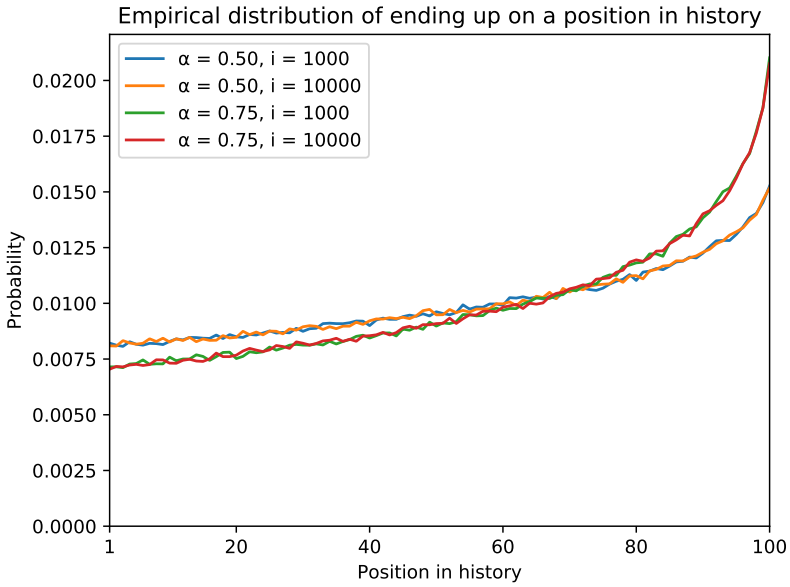

Figure 3: The empirical distributions (averaged over 1M runs) of reaching a specific position in a history of length $h = 100$, starting from positions $i \in \{1000, 10000\}$, with $\alpha \in \{0.5, 0.75\}$. The distributions have a strong dependence on α, and a weak dependence on i.

Now, consider the following process. Begin a phase at the generic position i:

(i) Let $j = 1$

(ii) While true

- Flip an independent coin with head probability at least p_j
- If it is heads, let $j = j + 1$; otherwise, break.

Observe that once a phase begins, it has constant probability of never ending. Moreover, there is a simple coupling from this process to the original one: we try to begin a streak at i when a phase begins; if the jth coin is heads, then position $i + j$ copies from some position in $\{i, i + 1, \ldots, i + j - 1\}$. Therefore, if all coins in a phase are heads (i.e., with constant probability), our process will have converged. If not, our process might have converged on the token at i or not. In any case, we run another phase on i', where the process will converge on the token on i' with at least constant probability. Since each phase converges with constant probability, our process will finally converge to a single-winning token with probability 1. □

6 FUTURE DIRECTIONS

There are a number of open questions about our model as stated:

(1) Can our results be extended for $\alpha \in [1, 2]$?
(2) Does Z^* always exist for any vector of weights $\vec{w}$?
(3) What can be said about the support of Z^*? In what situations is it supported over $[0, 1]$ or just at $\{0, 1\}$?

Additionally, there are a set of models with more complex dynamics that show some connections to our simpler model:

- Modern sequence models based on attention [2] incorporate more features of the input, and more interactions among tokens of the history; the model we study represents a very special case of an attention-weighted ML sequence model.
- There are also models that directly capture the copying of tokens from the input to the output, such as Copynet [7] and Neural Turing Machines [6].

These more complex models differ from ours in multiple respects, raising a number of questions:

(1) Can our results cover settings in which attention weights are only indirectly coupled to final probabilities of tokens? Such models may be fundamentally different, as a token may support the appearance of a different token.
(2) Can our results extend to introduce weights that are dependent on item embeddings?
(3) Can our results cover softmax normalization, rather than the normalization we use? It is easy to see that the results would be different, even for the classical Pólya case (i.e., uniform weights). With softmax and uniform weights, there seems to be a single winner with non-zero probability, which is in contrast with the classical case.
(4) Can our results extend to multiple attention heads [13]?
(5) Can our results extend to learned attention weights that are dependent on additional elements such as the context and the features of a particular attention position?

REFERENCES

[1] Ashton Anderson, Ravi Kumar, Andrew Tomkins, and Sergei Vassilvitskii. The dynamics of repeat consumption. In *WWW*, pages 419–430, 2014.
[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.
[3] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[4] Austin R. Benson, Ravi Kumar, and Andrew Tomkins. Modeling user consumption sequences. In *WWW*, pages 519–529, 2016.
[5] F. Eggenberger and G. Pólya. Uber die statistik verketteter vorgange. *Math. Mech.*, 3:279–289, 1923.
[6] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. Technical report, arXiv, 1410.5401, 2014.
[7] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. Incorporating copying mechanism in sequence-to-sequence learning. In *ACL*, 2016.
[8] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *FOCS*, pages 57–65, 2000.
[9] Ravi Kumar, Mohammad Mahdian, Bo Pang, Andrew Tomkins, and Sergei Vassilvitskii. Driven by food: Modeling geographic choice. In *WSDM*, pages 213–222, 2015.
[10] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Extracting large-scale knowledge bases from the web. In *VLDB*, pages 639–650, 1999.
[11] Hosam Mahmoud. *Pólya Urn Models.* Chapman and Hall/CRC, 2008.
[12] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425, 1955.
[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008. 2017.
[14] George Kingsley Zipf. *Human Behavior and the Principle of Least Effort.* Addison-Wesley, Reading, MA, 1949.