Chapter 6

Sequential Pattern Matching – Analysis of Knuth-Morris-Pratt Type Algorithms Using the Subadditive Ergodic Theorem

Tobias Reichl

Based on an article by Mireille Règnier and Wojciech Szpankowski this report outlines the complexity analysis of Knuth-Morris-Pratt type algorithms using the Subadditive Ergodic Theorem, Martingales and Azuma's Inequality.

Using the Subadditive Ergodic Theorem we will prove the existence of a linearity constant for worst and average case. Although the Subadditive Ergodic Theorem doesn't indicate a way to compute the linearity constant, we may use Azuma's Inequality to show that the number of comparisons done is well concentrated around its mean value.

6.1 Pattern Matching

6.1.1 Conventions

Before starting we have to introduce some conventions in nomenclature: a pattern p of length m, denoted p_1^m, is matched against a text t of length n, denoted t_1^n.
We have to define some kind of counting function:

$$M(l,k) = \begin{cases} 1 & t[l] \text{ is compared to } p[k] \\ 0 & \text{otherwise} \end{cases} .$$

A position in the text is called an *alignment position* (AP) if starting from it comparisons between text and pattern are done, or more formally

$$M(AP + (k-1), k) = 1 \qquad \text{for some k.}$$

6.1.2 Defining Sequential Algorithms

We will classify algorithms by a property we call *sequentiality*.

1. **Semi-sequential:** The sequence of alignment positions used by the algorithm is non-decreasing.
2. **Strongly semi-sequential:** (1) and the comparisons $M(l_i, k_i)$ define non-decreasing text-positions l_i.
3. **Sequential:** (1) and $M(l,k) = 1 \Rightarrow t_{l-(k-1)}^{l-1} = p_1^{k-1}$, so: text-pattern comparisons $M(l,k)$ are only done as long as there is a prefix of the pattern to the left of the text position to be compared next.
4. **Strongly sequential:** (1), (2) and (3).

6.1.3 Naive / Brute Force Algorithm

In short we may outline the *naive* or *brute force algorithm* as follows:

- Every text position is an alignment position.
- The aligned pattern is matched against the text from left to right until either a mismatch occurs or the pattern is found.
- The pattern is then shifted by one and the next matching is started.

The brute force algorithm is a sequential algorithm: the APs are non-decreasing and the condition $M(l,k) = 1 \Rightarrow t_{l-(k-1)}^{l-1} = p_1^{k-1}$ holds: no more comparisons are done after a mismatch is found, so every alignment is used only as long as prefixes of the pattern are found in the text.
The sequence of text positions l_i defined by the sequence of comparisons $M(l_i, k_i)$, however, may include 'jumping backwards', i.e. if a mismatch occurs, the AP is shifted by one and comparisons again start at the beginning of the pattern.

6.1.4 Knuth-Morris-Pratt

Idea: (Morris-Pratt) Disregard APs if we already know that there cannot be a prefix of the pattern, namely the ones that safisfy $t_{l+i}^{l+k-1} \neq p_1^{k-i}$ for all i. Or equivalently $p_{1+i}^{k} \neq p_1^{k-1}$ as the already processed text has to be identical to the corresponding prefix of the pattern.
This knowledge can be obtained by a preprocessing of the pattern. The specific shifting functions can formally be described as following:

Morris-Pratt-Variant (MP):

$$S = \min\{k-1; \min\{s > 0 : p_{1+s}^{k-(s+1)}\}\}$$

Knuth-Morris-Pratt-Variant (KMP):

$$S = \min\{k; \min\{s : p_{1+s}^{k-(s+1)} \quad \text{and} \quad p_k^k \neq p_{k-s}^{k-s}\}\}$$

MP and KMP differ in the amount of information used from the pattern. Both are strongly sequential algorithms, because from the definition of the shift function it is automatic that the sequentiality condition (2) holds, there is no 'jumping backwards'.

6.1.5 Defining Complexity

The complexity in matching a pattern p against a text portion t_r^s can be defined as the number of comparisons needed:

$$c_{r,s}(t,p) = \sum_{l\in[r,s],k\in[1,m]} M(l,k) \tag{6.1}$$

Overall complexity $c_{1,n}$ is denoted as c_n. If either the text or the pattern is a realization of a random sequence we shall write C_n.
To look at KMP we have to introduce two probabilistic tools: the Subadditive Ergodic Theorem and Azuma's Inequality.

6.2 Subadditive Ergodic Theorem

6.2.1 Fekete's Theorem

Assume a deterministic sequence $\{x_n\}_{n=0}^{\infty}$ satisfies the so called *subadditivity property*, that is

$$x_{m+n} \leq x_n + x_m \tag{6.2}$$

for all integers $m, n \geq 0$. We may fix $m \geq 0$ and write

$$n = km + l \quad \Leftrightarrow \quad \frac{k}{n} = \frac{1}{m} - \frac{l}{mn} \ . \tag{6.3}$$

Then by successive application of the subadditivity property arrive at

$$x_n = x_{km+l} \leq x_m + x_m + \cdots + x_m + x_l = kx_m + x_l \ . \tag{6.4}$$

Now dividing by n and considering $n \to \infty$ resp. $k/n \to 1/m$, cf. (6.3) we get

$$\limsup_{n\to\infty} \frac{x_n}{n} \leq \inf_{m\geq 1} \frac{x_m}{m} \leq \alpha \ . \tag{6.5}$$

To complete the derivation we may use the definition of lim inf and get the following:

$$\liminf_{n\to\infty} \frac{x_n}{n} = \sup_{n\geq 0} \left\{ \inf_{k\geq n} \frac{x_n}{n} \right\} = \alpha \tag{6.6}$$

Thus we just derived the theorem of Fekete.

Theorem 6.1 (Fekete 1923). *If a sequence of real numbers satisfies the* subadditive property

$$x_{m+n} \leq x_n + x_m \tag{6.7}$$

for all integers $m, n \geq 0$, then

$$\lim_{n\to\infty} \frac{x_n}{n} = \inf_{m\geq 1} \frac{x_m}{m} \ . \tag{6.8}$$

If the subadditvity property (6.7) is replaced by the superadditvity property

$$x_{m+n} \geq x_n + x_m \tag{6.9}$$

for all integers $m, n \geq 0$, then

$$\lim_{n\to\infty} \frac{x_n}{n} = \sup_{m\geq 1} \frac{x_m}{m} \ . \tag{6.10}$$

Example 6.1 (Longest Common Subsequence). The *longest common subsequence* (LCS) problem is a special case of the *edit distance* problem. Two ergodic stationary sequences $X = X_1, X_2, \dots, X_n$ and $Y = Y_1, Y_2, \dots, Y_n$ are given, then let

$$L_n = max\{K : X_{i_k} = Y_{j_k} \text{ for } 1 \le k \le K, \quad \begin{array}{ll} \text{where} & 1 \le i_1 < i_2 < \dots < i_K \le n, \\ \text{and} & 1 \le j_1 < j_2 < \dots < j_K \le n\} \end{array}$$

be the length of the longest common subsequence. Observe that

$$L_{1,n} \ge L_{1,m} + L_{m,n} \ . \tag{6.11}$$

The LCS in the region $(1, n)$ may cross the boundary of X_1^m, Y_1^m and X_m^n, Y_m^n. Hence it may be bigger than the sum of the LCSs in each subregion $(1, m)$ and (m, n) and so $a_n = \mathbf{E}[L_{1,n}]$ is superadditive:

$$\lim_{n\to\infty} \frac{a_n}{n} = \alpha = \sup_{m\ge 1} \frac{\mathbf{E}[L_m]}{m} \ . \tag{6.12}$$

But here you can already see the cavity: Fekete's Theorem[1] only states the existence of the linearity constant, but neither tells us its value nor even how to compute it. For the LCS problem here Steele in 1982 conjectured $\alpha \approx 0.8284$.

Theorem 6.2 (DeBruijn and Erdös 1952). *The subadditivtiy property can be relaxed to include a sequence $c_n = o(n)$*

$$x_{n+m} \le x_n + x_m + c_{n+m} \tag{6.13}$$

where

$$\sum_{k=1}^{\infty} \frac{c_k}{k^2} < \infty \ . \tag{6.14}$$

Then, too

$$\lim_{n\to\infty} \frac{x_n}{n} = \inf_{m\ge 1} \frac{x_m}{m} \ . \tag{6.15}$$

6.2.2 Subadditive Ergodic Theorem

As Fekete's Theorem only applies to deterministic sequences, effort has been taken to generalize it to sequences of random variables.

Theorem 6.3 (Kingman and Liggett). *Let $X_{m,n}$ $(m < n)$ be a sequence of random variables satisfying the following properties:*

1. $X_{0,n} \le X_{0,m} + X_{m,n}$ *(subadditivity property)*
2. *For every k, $\{X_{nk,(n+1)k}, n \ge 1\}$ is a stationary sequence.*
3. *The distribution of $\{X_{m,m+k}, k \ge 1\}$ does not depend on m.*
4. $\mathbf{E}[X_{0,1}] < \infty$ *and for each n,* $\mathbf{E}[X_{0,n}] \ge c_0 n$ *where $c_0 > -\infty$.*

Then

$$\lim_{n\to\infty} \frac{\mathbf{E}[X_{0,n}]}{n} = \inf_{m\ge 1} \frac{\mathbf{E}[X_{0,m}]}{m} := \alpha \ , \tag{6.16}$$

$$\lim_{n\to\infty} \frac{X_{0,n}}{n} = X \qquad (a.s) \ . \tag{6.17}$$

[1] And the Subadditive Ergodic Theorem, as as we will see later.

Theorem 6.4 (Deriennic). *Similar to subadditivity with deterministic sequences, subadditivity with random sequences can be relaxed to include a sequence A_n*

$$X_{0,n} \leq X_{0,m} + X_{m,n} + A_n \tag{6.18}$$

such that $\lim_{n\to\infty} \mathbf{E}\left[A_n/n\right] = 0$*. Then, too*

$$\lim_{n\to\infty} \frac{X_{0,n}}{n} = X \qquad (a.s) \ . \tag{6.19}$$

6.3 Martingales and Azuma's Inequality

6.3.1 Basic Properties of Martingales

Martingale is a standard tool in probabilistic analysis. A sequence

$$Y_n = f(X_1, X_2, \ldots, X_n), \qquad n > 0 \tag{6.20}$$

is a martingale with respect to the *filtration*

$$\mathcal{F}_n = (X_1, X_2, \ldots, X_n) \tag{6.21}$$

if for all $n \geq 0$ the following hold:

1. $\mathbf{E}\left[|Y_n|\right] < \infty$ and
2. $\mathbf{E}\left[Y_{n+1} \mid X_0, X_1, \ldots, X_n\right] = \mathbf{E}\left[Y_{n+1} \mid \mathcal{F}_n\right] = Y_n$

So $\mathbf{E}\left[Y_{n+1} \mid \mathcal{F}_n\right]$ defines a random variable depending on the *k*nowledge contained in $(X_1, X_2, \ldots, X_n)$. Now let's define the *martingale difference* as

$$D_n = Y_n - Y_{n-1} \tag{6.22}$$

so that

$$Y_n = Y_0 + \sum_{i=1}^{n} D_i \qquad \Leftrightarrow \qquad \sum_{i=1}^{n} D_i = Y_n - Y_0 \ . \tag{6.23}$$

Then we may rewrite the martingale difference as

$$D_i = Y_i - Y_{i-1} = \mathbf{E}\left[Y_n \mid \mathcal{F}_i\right] - \mathbf{E}\left[Y_n \mid \mathcal{F}_{i-1}\right] \tag{6.24}$$

This is possible as the realization of the martingale sequence Y_n depends on the knowledge contained in $\mathcal{F}_i$, so the difference between neighbouring elements depends on the difference in knowledge about X_i. Now observe:

$$\mathbf{E}\left[Y_n \mid \mathcal{F}_n\right] = Y_n \qquad \text{and} \qquad \mathbf{E}\left[Y_n \mid \mathcal{F}_0\right] = \mathbf{E}\left[Y_n\right] \ .$$

Note: $\mathcal{F}_n$ completely defines Y_n, while $\mathcal{F}_0$ contains no information about Y_n.
Interestingly we are now able to rewrite the martingale diffence sum, cf. (6.23), as

$$\sum_{i=1}^{n} D_i = Y_n - \mathbf{E}\left[Y_0\right] \ . \tag{6.25}$$

And this is what we are interested in: the deviation of Y_n from its mean value. To further assess it we will now introduce Hoeffding's Inequality.

6.3.2 Hoeffding's Inequality and Azuma's Inequality

Theorem 6.5 (Hoeffding's Inequality). *Let $\{Y_n\}_{n=0}^{\infty}$ be a martingale and let there exist a constant c_n such that*

$$|Y_n - Y_{n-1}| = |D_n| \le c_n \tag{6.26}$$

Then

$$Pr\{|Y_n - Y_0| \ge x\} = Pr\left\{\left|\sum_{i=1}^{n} D_i\right| \ge x\right\} \le 2\exp\left(-\frac{x^2}{2\sum_{i=1}^{n} c_i^2}\right) \ . \tag{6.27}$$

By now, we know how to use the martingale difference sum $\sum_{i=1}^{n} D_i$ for assessing the deviation from the mean. We also know how to assess this martingale difference sum, provided D_i is bounded.
What we still need is to establish bounds on D_i.
The trick: let $\hat{X}_i$ be an independent copy of X_i. Then

$$\begin{aligned}
\mathbf{E}\left[f_n(X_1,\ldots,X_i,\ldots,X_n) \mid \mathcal{F}_{i-1}\right] &= \\
\mathbf{E}\left[f_n(X_1,\ldots,\hat{X}_i,\ldots,X_n) \mid \mathcal{F}_{i-1}\right] &= \\
\mathbf{E}\left[f_n(X_1,\ldots,\hat{X}_i,\ldots,X_n) \mid \mathcal{F}_{i}\right] &
\end{aligned}$$

because both X_i and $\hat{X}_i$ share the same distribution, but $\mathcal{F}_i$ in respect to $\mathcal{F}_{i-1}$ doesn't contain additional information about $\hat{X}_i$. Hence we may rewrite the martingale difference again as

$$\begin{aligned}
D_i &= \mathbf{E}\left[Y_n \mid \mathcal{F}_i\right] - \mathbf{E}\left[Y_n \mid \mathcal{F}_{i-1}\right] \\
&= \mathbf{E}\left[f_n(X_1,\ldots,X_i,\ldots,X_n) \mid \mathcal{F}_i\right] - \mathbf{E}\left[f_n(X_1,\ldots,X_i,\ldots,X_n) \mid \mathcal{F}_{i-1}\right] \\
&= \mathbf{E}\left[f_n(X_1,\ldots,X_i,\ldots,X_n) \mid \mathcal{F}_i\right] - \mathbf{E}\left[f_n(X_1,\ldots,\hat{X}_i,\ldots,X_n) \mid \mathcal{F}_i\right]
\end{aligned}$$

Taking into account both terms only differ in including X_i resp. $\hat{X}_i$ we are able to postulate the existence of a constant d_i with $|D_i| \le d_i$ and using Hoeffding's Inequaltity we finally arrive at Azuma's Inequality.

Theorem 6.6 (Azuma's Inequality). *Let $\{Y_n\}_{n=0}^{\infty}$ be a martingale and let there exist a constant c_n such that*

$$\left|f_n(X_1,\ldots,X_i,\ldots,X_n) - f_n(X_1,\ldots,\hat{X}_i,\ldots,X_n)\right| \le c_i \tag{6.28}$$

where $\hat{X}_i$ is an independent copy of X_i. Then

$$\begin{aligned}
& Pr\left\{\left|f_n(X_1,\ldots,X_i,\ldots,X_n) - \mathbf{E}\left[f_n(X_1,\ldots,\hat{X}_i,\ldots,X_n)\right]\right| \ge x\right\} \\
= \ & Pr\{|Y_n - \mathbf{E}[Y_n]| \ge x\} \quad \le \quad 2\exp\left(-\frac{x^2}{2\sum_{i=1} nc_i^2}\right)
\end{aligned}$$

6.4 Application to KMP

6.4.1 Establishing m-Convergence

An alignment position in the text is called *unavoidable alignment position* if for any $r \le i$ and any $l \ge i + m$ it's an alignment position when the algorithm is run on t_r^l. KMP-like algorithms share the same set of unavoidable alignment positions

$$\mathcal{U} = \bigcup_{l=1}^{n} \{U_l\} \tag{6.29}$$

where

$$U_l = \min\{\min_{1 \le k \le l} \{t_k^l \preceq p\}, l+1\} \ . \tag{6.30}$$

This equation on the one hand specifies starting positions of pattern prefixes as unavoidable alignment positions (no positions inside those prefixes, as those would be jumped over) and on the other hand specifies steps of size one if there is no pattern prefix.
Interstingly this property seems to be uniquely limited to Morris-Pratt type algorithms – e.g. the Boyer-Moore algorithm does not have this property.

An algorithm is said to be l-convergent if there exists an increasing sequence of unavoidable alignment positions $\{U_l\}_{i=1}^{n}$ satisfying

$$U_{i+1} - U_i \le l \ . \tag{6.31}$$

Thus, l-convergence indicates the maximum size ‘jumps’ for an algorithm. For example, the brute force algorithm is 1-convergent and – what we are interested in more – KMP-like algorithms are m-convergent.

Proof. Let l be a text position and let r be any text position with $r \le U_l$. Then let $\{A_i\}$ be the set of APs when the algorithm is run on T_r^m. Note: $r \in \{A_i\}$ as the algorithm inevitably aligns at the starting position.
Then we may define the last alignment position A_J before U_l as

$$A_J = \max\{A_i : A_i < U_l\} \ . \tag{6.32}$$

So we have $A_{J+1} \ge U_l$. Using an adversary argument we will show that $A_{J+1} > U_l$ cannot be true, thus $A_{J+1} = U_l$. We define

$$y = \max\{k : M(k, (k - A_J) + 1) = 1\} \ , \tag{6.33}$$

so y is the rightmost position in the text we can do a comparison at when starting at A_J. Observe: $y \le l$. Otherwise, when comparisons would be done further, $T_{A_J}^l \preceq H$ would have to hold – and this in turn contradicts the definition of U_l.
Since KMP-like algorithms are strongly sequential, the text-pattern comparisons define non-decreasing sequences of text positions. For pattern-text comparisons at text position $y+1$ the pattern cannot be aligned at A_J, it has to be aligned at the next alignment position A_{J+1} with $A_{J+1} \le y + 1 \le l + 1$.
The definition of U_l leaves two possibilities: $U_l \le l$ if there is a prefix of the pattern, or $U_l = l + 1$ if there is no prefix. The above equation $A_{J+1} \le l + 1$ together with the second possibility $U_l = l + 1$ contradicts the assumption $U_l < A_{J+1}$, so we may assume the first possibility $U_l \le l$ – this then implies that $H_{U_l}^l \preceq H$.
An occurence of the whole pattern is consistent with the available information. We – as we want to create a contradiction – may assume this is the case. As the sequence $\{A_i\}$ is non-decreasing and $A_{J+1} > U_l$ this occurence will be ‘jumped over’ and not be detected by the algorithm. Thus $A_{J+1} = U_l$ as needed.

Taking this a bit further we may combine $A_{J+1} \le y + 1$ and $y \le A_J + m - 1$ to $A_{J+1} \le y + 1 \le A_J + m + 1 - 1 = A_J + m$. So we have shown $A_{J+1} - A_J \le m$ for any pair (A_J, A_{J+1}) of APs in the text, thus KMP-like algorithms are m-convergent. $\square$

6.4.2 Establishing Subadditivity

If c_n, the number of comparisons, is subadditive we may use the Subadditive Ergodic Theorem to prove linear complexity of algorithms. To achieve this we have to show that c_n is (almost) subadditive

$$c_{1,n} \le c_{1,r} + c_{r,n} + a \ . \tag{6.34}$$

After rearranging the equation it suffices to prove the existence of an a such that

$$|c_{1,n} - (c_{1,r} + c_{r,n})| \le a \ . \tag{6.35}$$

Let U_r be the smallest unavoidable aligment position greater than r. Then we are able to split $c_{1,n} - (c_{1,r} + c_{r,n})$ into $c_{1,n} - (c_{1,r} + c_{U_r,n})$ and $c_{r,n} - c_{U_r,n}$.
For the first part we have to count either:

- S_1: comparisons done after position r with alignment positions before r. Those only contribute to $c_{1,n}$ but neither to $c_{1,r}$ (the algorithm won't compare after r) nor $c_{U_r,n}$ (the algorithm doesn't align before U_r, thus not before r, too).
- S_2: comparisons done with alignment positions between r and U_r. Those also only contribute to $c_{1,n}$ but neither to $c_{1,r}$ nor $c_{U_r,n}$ (the algorithm in those cases only aligns before r resp. after U_r).

Summing up we arrive at

$$S_1 = \sum_{AP<r} \sum_{i \ge r} M(i, i - AP + 1) \le m^2 \ . \tag{6.36}$$

This sum is bounded as there are at maximum m alignment positions before r with comparisons done after r. And for each aligment position there are at maximum m comparisons done.

$$S_2 = \sum_{r \le AP < U_r} \sum_{i \ge r} M(AP + (i-1), i) \le lm \tag{6.37}$$

This sum is bounded: because of the l-convergence of sequential algorithms there are at maximum l text positions between r and U_r, each with at maximum m comparisons done. Note: with m-convergent KMP-like algorithms this would resolve to m^2, too.
For the second part we have to count comparisons done with alignment positons before U_r (thus between r and U_r). Those contribute to $c_{r,n}$ only as $c_{U_r,n}$ starts comparing at position U_r.

$$S_3 = \sum_{r \le AP < U_r} \sum_{i \ge r} M(AP + (i-1), i) \le lm \tag{6.38}$$

This is the same sum as S_2, hence bounded for the same reasons. Finally we are able to put the parts together:

$$|c_{1,n} - (c_{1,r} + c_{r,n})| \le |S_1 + S_2 - S_3| \le m^2 + lm = a \ . \tag{6.39}$$

So by now we have show subadditivity

$$c_{1,n} \le c_{1,r} + c_{r,n} + a \tag{6.40}$$

and are able to apply the Subadditive Ergodic Theorem.

6.4.3 Applying the Subadditive Ergodic Theorem

Before continuing we have to develop some modelling assumptions about the structure of text and pattern.

- **Deterministic Model:** Both text and pattern are non-random.[2] In this case we have to maximize complexity over all possible texts.
- **Semi-Random Model:** The text is a realization of stationary and ergodic sequence, the pattern is given, thus non-random. In this case we use average complexity over all texts.
- **Stationary Model:** Both text and pattern are a realization of a stationary and ergodic sequence, so we use average complexity over all texts and patterns.

Applying the Subadditve Ergodic Theorem yields similar results for worst and average case:

$$\begin{aligned}
&\text{Deterministic Model:} && \lim_{n\to\infty} \frac{max_t(c_n(t,p))}{n} = \alpha_1(p)\\
&\text{Semi-Random Model:} && \lim_{n\to\infty} \frac{\mathbf{E}_t\left[C_n(p)\right]}{n} = \alpha_2(p)\\
&\text{Stationary Model:} && \lim_{n\to\infty} \frac{\mathbf{E}_{t,p}\left[C_n\right]}{n} = \alpha_3
\end{aligned}$$

6.4.4 Applying Azuma's Inequality

Even if we cannot determine the linearity constants α_1 to α_3, we still can show that C_n is concentrated around its mean.
We may assume that the text t is generated by a memoryless source, and C_n is a function of this random text $t = t_1, t_2, \ldots, t_n$. By flipping a single character we may change C_n by at most $2m^2$ comparisons, so C_n satisfies the condition for applying Azuma's Inequality:

$$\left|C_n\left(t_1, t_2, \ldots, t_i, \ldots, t_n\right) - C_n\left(t_1, t_2, \ldots, \hat{t}_i, \ldots, t_n\right)\right| \leq 2m^2 \tag{6.41}$$

Theorem 6.7. *Let t be a random text of length m generated by a memoryless source and let the pattern p of length m be given. Then the number C_n of comparisons made by the Knuth-Morris-Pratt algorithm is concentrated around its mean*

$$\mathbf{E}\left[C_n\right] = \alpha_2 n\left(1 + o(n)\right). \tag{6.42}$$

Equally

$$\begin{aligned}
Pr\left\{\left|C_n - \alpha_2 n\right| \geq \epsilon n\right\} &\leq 2\exp\left(-\frac{(\epsilon n)^2}{2\cdot n\cdot\left(2m^2\right)^2}\left(1 + o(n)\right)\right)\\
&= 2\exp\left(-\frac{\epsilon^2 n}{4m^4}\left(1 + o(n)\right)\right)
\end{aligned} \tag{6.43}$$

for any $\epsilon > 0$.

[2]Applying Murphy's Law we may asume text and/or pattern to be exactly what you do not want them to be...

6.5 Concluding Remarks

The Subadditive Ergodic Theorem proves the existence of the linearity constant under quite general probabilistic assumptions. The main prerequisite is the existence of so called unavoidable alignment positions, a property that seems to be uniquely limited to Knuth-Morris-Pratt like algorithms.

Although we have not been able to compute this constant, we have been able to show that the number C_n of comparisons done is concentrated around its mean value $\alpha_2 n$.