Proving Concentration Measures for the Sliding Window Problem

A Project Report Submitted in Partial Fulfillment of the Requirements for the Degree of
Bachelor of Technology

by

Akshay Kumar, 10060
Aman Sharma, 10068
Shivam Bansal, 10686

under the guidance of

Dr. Satyadev Nandakumar

Department of Computer Science & Engineering
Indian Institute of Technology, Kanpur
November2013

Certificate

It is certified that the work contained in the project entitled **Proving Concentration Measures for the Sliding Window Problem** has been carried out under my supervision and this work has not been submitted elsewhere for degree.

Dr Satyadev Nandakumar
Assistant Professor
Department of Computer Science & Engineering

Acknowledgement

We would like to express our deep sense of gratitude to **Dr Satyadev Nandakumar**, for his invaluable help and guidance during the course of the project. We are grateful to him for having given us the support and confidence.

Abstract

In an unpublished manuscript, Alan Turing used an unproven lemma to give a construction of absolutely normal numbers. A proof for the weaker version of the lemma was provided by Becher et al in 2007 and they showed that the construction still holds. In this paper, we provide a proof of a lemma using Talagrand's concentration inequality which is stronger than that proved by Becher et al but weaker than Turing's hypothesis.

Introduction

Given a string $a = (a_0 a_1 \cdots a_{3n-1})$ is given and a pattern $p = (p_1 p_2 p_3)$ where the alphabet is the set $\{0, 1\}$, consider the following two sets:

1. $S_1 = \{i | a_{3i} a_{3i+1} a_{3i+2} = p_1 p_2 p_3\}$
2. $S_2 = \{i | a_i a_{i+1} a_{i+2} = p_1 p_2 p_3\}$

We are interesting in finding the expected size of S_1 and S_2 and also the probability that the size of S_1 or S_2 deviates from the expected size.

For the expected size,

For problem 1, define a random variable X_i as follows:

$$X_i = \begin{cases} 1 & \text{if } i \in S_1 \\ 0 & \text{otherwise} \end{cases}$$

Obviously, $0 \leq i \leq (n-1)$. Also, let X be the total number of matches. Then

$$X = \sum_{i=0}^{n-1} X_i \Rightarrow E[X] = E[\sum_{i=0}^{n-1} X_i] \Rightarrow E[X] = \sum_{i=0}^{n-1} E[X_i] = \sum_{i=0}^{n-1} \frac{1}{8} = \frac{n}{8}$$

The above equation follows by applying Linearity of Expectations and from the fact that $E[X_i] = \frac{1}{8}$. Hence, $E[X] = \frac{n}{8}$.

For the second problem, define the random variable Y$_i$ in a similar way.

$$Y_i = \begin{cases} 1 & \text{if } i \in S_2 \\ 0 & \text{otherwise} \end{cases}$$

Obviously, $0 \leq i \leq (3n-3)$. Also, let Y be the total number of matches. Then

$$Y = \sum_{i=0}^{3n-3} Y_i \Rightarrow E[Y] = E[\sum_{i=0}^{3n-3} Y_i] \Rightarrow E[Y] = \sum_{i=0}^{3n-3} E[Y_i] = \sum_{i=0}^{3n-3} \frac{1}{8} = \frac{3n-2}{8}$$

Here also, the above equation follows by applying Linearity of Expectations and from the fact that $E[Y_i] = \frac{1}{8}$. Hence, $E[Y] = \frac{3n-2}{8}$.

Next comes the bigger question, a bound on the probability of deviation from the expected value of n?

For the first question, it can be found using Chernoff's Bound. Note that the variables X_i's defined above are Bernoulli Random Variables independent of each other. Also, $\mu = \frac{n}{8}$ (already shown above).

Using Chernoff's Bound

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

$$Pr[X \geq (1+\delta)\mu] \leq e^{-\mu\delta^2/4}$$

Hence,

$$Pr[|X-\mu| \geq \delta\mu] = Pr[(X \leq (1-\delta)\mu) \cup (X \geq (1+\delta)\mu)]$$
$$= Pr[X \leq (1-\delta)\mu] + Pr[X \geq (1+\delta)\mu] \leq e^{-\mu\delta^2/2} + e^{-\mu\delta^2/4} \leq 2e^{-\mu\delta^2/2}$$

In our case, $\mu = \frac{n}{8}$. This gives us an inverse exponential bound.

However, for the second case, there is no direct method to get a concentration bound. Chernoff's Bound won't work because Y_i's are not independent. There is no trivial way to get a concentration bound of the deviation of Y. Turing claimed a similar kind of bound for Y as well but didn't prove it. In 2007, [1] proved a lower bound for the same inequality and showed that the proof where this inqquality was used is still valid with the weaker inequality. The method used by used was very complicated and involved rigorous combinatorics arguments. In this paper, we give a stronger bound on the inequality using application of Talagrand's Inqeuality.

Turing's Lemma

Definition Let $t \in \mathbb{N},\ r \in \mathbb{N}\ \&\ \gamma \in \{0,1\}^r$. Then,

1. $S(w,\gamma)$ is the number of occurences of γ in w
2. $P(t,\gamma,n,R) = \{w \in \{0,\cdots,t-1\}^R : S(w,\gamma) = n\}$
3. $N(t,\gamma,n,R) = \#P(t,\gamma,n,R)$

The function N returns the number of R length strings that have n occurrences of γ. This is not a trivial function due to the possible overlapping of different occurences of γ. For example, if $\gamma = 11$ it occurs once in 1100, twice in 0111 and three times in 1111. Hence the event of γ matching the r length substring at position i is not independent of the event that γ matches(or not matches) the r length substring at position $i-r+1\cdots i+r-1$. Hopefully, if we only consider the exact number of occurences of a given digit, the expression for N becomes simple: in the scale of t, there are only $(t-1)^{R-n}$ R-length words with exactly n occurences of the digit d in fixed places. Hence, the number of words of length R in the base t with exactly n occurrences of the $digit\ d$ in some places is

$$N(t,d,n,R) = \binom{R}{n}(t-1)^{R-n}$$

Obviously,

$$\textstyle\sum_{0\leq n\leq R} N(t,d,n,R) = t^R$$

Unproved Turing's Lemma. Let $t \in \mathbb{N},\ r \in \mathbb{N}\ \&\ \gamma \in \{0,1\}^r$, and let $\delta \in \mathbb{R}$ be such that $\delta\frac{t^r}{R} < 0.3$. Then,

$$\textstyle\sum_{|n-R/t^r|>\delta} N(t,\gamma,n,R) < 2t^R e^{-\frac{\delta^2 t^r}{4R}}$$

$$Pr[|n-\tfrac{R}{t^r}| > \delta] < 2e^{-\frac{\delta^2 t^r}{4R}}$$

Becher et al gave a substitution for the unproved Turing's Lemma which is

Lemma. Let $t \in \mathbb{N},\ r \in \mathbb{N}\ \&\ \gamma \in \{0,1\}^r$, and let ε be such that $\frac{6}{\lfloor \frac{R}{r} \rfloor} \leq \varepsilon \leq \frac{1}{t^r}$. Then,

$$\textstyle\sum_{|n-R/t^r|\geq\varepsilon R} N(t,\gamma,n,R) < 2t^{R+2r-2} r e^{-\frac{t^r\varepsilon^2 R}{6r}}$$

As already mentioned, the above result involved fairly complicated combinatorial agruments.

Before going into the exact statement of Talagrand's Inequality, it's imperative to define Convex Distance.

Convex Distance

- $\forall\, r \in [-1,1]^N$ & $\alpha \in \{0,1\}^N$
  We say that α **supports** r if
$$r_i \neq 0 \Rightarrow \alpha_i = 1 \quad i = 1m \cdots, N$$
$$(\alpha_i = 0 \Rightarrow r_i = 0 \quad \forall i)$$
- $A, X \subseteq [-1,1]^N$. The **Combinatorial Support**
$$u_A(X) = \{\alpha \in \{0,1\}^N | \exists x \in X - A \text{ s.t. } \alpha \text{ supports } x\}$$
- **Combinatorial Hull**
$$V_A(X) = \text{ Convex Hull of } u_A(X)$$
- **Convex Distance**
  $d_c(X, A)$ is the distance of the combinatorial hull $V_A(x)$ from origin.

Talagrand's Inequality

In its purest form, the inequality is :
Let $\Omega = \Omega_1 \times \Omega_2 \times \cdots \Omega_n$ be a probability measure product space. If $A \subseteq \Omega$, then for any $t \geq 0$,
$$Pr[A].Pr[\bar{A}_t] \leq e^{-t^2/4}$$
where A_t is the annulus of radius t around the figure A and $\bar{A}_t$ is its complement.
$$A_t = \{x \in \Omega : (A, x) \leq t\}$$
In the above equation, is the Talagrand's Convex Distance not the normal Euclidean Distance.

The above inequality can also be stated as follows:
Let $X_0, X_1, \cdots, X_N$ be random variables and $F : \mathbb{R}^N \to \mathbb{R}$ be a function such that the following holds:

1. $\forall\, i \in \{1, 2, , N\}\ |X_i| \geq 1$
2. X_i are mutually independent
3. F is convex *i.e.*
   Let $\vec{r_1}, \vec{r_2} \in \mathbb{R}^N$. F is convex if
$$F\Big(\frac{\vec{r_1} + \vec{r_2}}{2}\Big) \leq \frac{F(\vec{r_1}) + F(\vec{r_2})}{2}$$

4. F is co-ordinate wise $1-$Lipschitz *i.e.*

$$\forall\, i = 1, \cdots, N \quad |F(\vec{X}_{-i}, x) - F(\vec{X}_{-i}, y)| \leq |x - y|$$

keeping all variables except i^{th} intact.

Then the following result holds:

1. $\exists\, c > 0$ s.t. $\forall \lambda$

$$P[\omega : |F(\omega) - MF| \geq \lambda] \leq ce^{-c\lambda^2}$$

where M is the median of F.

2. $\exists\, c > 0$ s.t. $\forall \lambda$

$$P[\omega : |F(\omega) - EF| \geq \lambda] \leq ce^{-c\lambda^2}$$

where E is the expectation of F.

Proof :

It suffices to show that for any convex set $A \subseteq D^N$ (unit disk in $N-$dimensions)

$$0. \quad Ee^{cd^2(X,A)} \leq \frac{1}{P[\omega : X(w) \in A]}$$

It suffices to show that $0 \Rightarrow 1$ and $1 \Rightarrow 2$.

$1 \Rightarrow 2$ is straightforward because mean and expectation of a function don't differ much which can be reflected in the constant on RHS.

We need to show

$$P[F(\vec{X}) \leq x].P[F(\vec{X}) \geq y] \leq e^{-c|x-y|^2}$$

The first term $P[F(\vec{X}) \leq x]$ can be visualized as a set A which is defined as follows:

$$A : \{\vec{z} \in \mathbb{R}^N : F(\vec{z}) \leq x\}$$

Since$F$ is convex, hence A is also convex.

$\Rightarrow$ We need to prove

$$e^{c|x-y|^2} P[F(\vec{x}) \geq y] \leq \frac{1}{P[x \in A]}$$

If we show that

$$e^{c|x-y|^2} P[F(\vec{x}) \geq y] \leq E[e^{cd_m^2(X,A)}]$$

Then using 0 *i.e* $E[e^{cd_m^2(X,A)}] \leq \frac{1}{P[x \in A]}$, we get

$$e^{c|x-y|^2} P[F(\vec{x}) \geq y] \leq \frac{1}{P[x \in A]}$$

Hence, we need to show

$$e^{c|x-y|^2} P[F(\vec{x}) \geq y] \leq E[e^{cd_m^2(X,A)}]$$

Note that

$$e^{c|x-y|^2} P[F(\vec{x}) \geq y] \leq E[e^{cd_m^2(X,A)}] + e^{c|x-y|^2} P[F(\vec{x}) \geq y] = E[e^{c(x-y)^2}]$$

If F is $1-$Lipschitz,

$$|F(\vec{x}) - F(\vec{y})| \leq |\vec{x} - \vec{y}|$$
$$\Rightarrow |\vec{x} - \vec{y}| \leq |F^{-1}(\vec{x}) - F^{-1}(\vec{y})|$$

Hence, we can say that $E[e^{c(x-y)^2}] \leq E[e^{cd_n^2(X,A)}]$.

This implies

$$e^{c|x-y|^2} P[F(\vec{x}) \geq y] \leq E[e^{cd_m^2(X,A)}]$$

This completes the proof of Talagrand's Inequality.

Certifiable Functions

Let $f(x_1, \cdots, x_n)$ be a real valued function on a product space $\Omega = \prod_i \in [n]\Omega_i$. Function f is r-certifiable if for every $x = (x_1, \cdots, x_n) \in \Omega$, there exists a set of indices $J(x) \subseteq [n]$ s.t.

- $|J(x)| \leq r \times f(x)$
- if y agrees with x on the co-ordinates in J(x),then $f(y) \geq f(x)$

The set $J(x)$ is said to be a certificate for $J(x)$

For example,let f be the number of heads in n coin tosses. We can consider the following certificate for f
$J(x) = \{i : x_i = 1\}$. Then $J(x) \leq f(x)$ and whenever y agrees with x on elements in J(x). Hence, f is 1-certifiable.

Talagrand's Inequality for Certifiable Functions

Let $f : \Omega \to \mathbb{R}$ be r-certifiable and suppose it is 1-lipschitz with constant c(changing any co-ordinate changes the value of the function by atmost c).Then for all $t > 0$

$$Pr[f > E[f] + t] \leq 2 \cdot e^{-\frac{t^2}{4c^2 r(E(f)+t)}} \tag{1}$$

and

$$Pr[f < E[f] - t] \leq 2 \cdot e^{-\frac{t^2}{4c^2 rE(f)}} \tag{2}$$

where $E[f]$ is the expected value of f.

Example : Longest Increasing Subsequence

Given a sequence $a := (a_1, a_2, \cdots, a_n)$, the longest increasing problem problem is to find a subsequence of the given sequence such that the elements of the subsequene are in sorted order, from lowest to highest and the subsequence is as long as possible *i.e.* a set of indices $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ such that $x_{i_1} \leq x_{i_2} \leq \cdots \leq x_{i_k}$. It was shown by [2] that the expeccted length of Longest Increasing Subsequence tends to $2\sqrt{n}$ as n approaches infinity.

We are interested in calculating the concentration bounds on expected length of Longest Increasing Subsequence. Let $I(x)$ denote this value for a sequence x. Let the set of corresponding indices in Longest Increasing Subsequence be denoted by $J(x)$. Clearly, following properties hold about $J(x)$:

1. $I(x) = |J(x)|$
2. $|J|$ is a certificate for $I(x)$
3. I is $1-$Lipschitz

Hence, if $X_1, \cdots, X_n$ are uniformly independently in $[0, 1]$, then for $I = I(X_1, \cdots, X_n)$,

$$Pr[I > M[I] + t] \leq 2e^{-t^2/4(M[I]+t)} \quad Pr[I < M[I] - t] \leq 2e^{-t^2/4M[I]}$$

Substituting the value of $M(I) = 2\sqrt(n)$ in the above inequations, if $t = O(n^{\frac{1}{4}})$, we get

$$Pr[|I - M[I]| > t] < ploy(1/e)$$

Hence, I is actually confined to a very small interval of size $O(n^{\frac{1}{4}})$.

Our application of Talagrand's inequality

Let us consider the application of Talagrand's inequality to our problem.
Let f be the number of matches of the r-length substring γ in the R-length string w.
$J(w) = \{i \cdots i + r - 1 | w[i \cdots i + r - 1] = \gamma[1 \cdots r]\}$
This certificate stores all the matched positions in the string w.
Now,$|J(w)| \leq r \cdot f(w)$ with the maximum occuring when all matched indices are distinct(no overlapping).
Also, if w' agrees with w on positions $J(w)$, then the string w' will have atleast as many matches of γ than w.Hence,$J(w)$ is a certificate for w. By direct application of Talagrand's inequality.

$$Pr[|f - E[f]| > \delta] \leq 2 \cdot (e^{-\frac{\delta^2}{4c^2 r(E(f)+\delta)}} + e^{-\frac{\delta^2}{4c^2 rE(f)}}) \tag{3}$$

Since $E[f] = \frac{R}{t^r}$.
Also, f is r-lipschitz since, changing value at a particular index can change the number of matches by atmost the length of the pattern string γ i.e. r .
Also,$\frac{\delta \cdot t^r}{R} = \frac{\delta}{E[f]} < 0.3$
Hence, $\delta < 0.3 \cdot E[f]$

Hence,

$$Pr[|f - E[f]| > \delta] \leq 4 \cdot (e^{-\frac{\delta^2}{4r^2 r(1.3 \cdot E(f))}})$$

$$Pr[|f - E[f]| > \delta] \leq 4 \cdot (e^{-\frac{\delta^2}{5.2r^3 \cdot E(f)}})$$

$$Pr[|n - \frac{R}{t^r}| > \delta] < 4e^{-\frac{\delta^2 t^r}{5.2 \cdot r^3 \cdot R}}$$

References

1. Becher, VerÃşnica, Santiago Figueira, and Rafael Picchi. "TuringâĂŹs unpublished algorithm for normal numbers." Theoretical Computer Science 377.1 (2007): 126-138.

2. Odlyzko, A. M., and E. M. Rains. "On longest increasing subsequences in random permutations." Contemporary Mathematics 251 (2000): 439-452.

3. "Concentration of Measure for the analysis of Randomized Algorithms"
http://www.users.di.uniroma1.it/ ale/Papers/master.pdf