Graded Homework 1, exercise 5

Marco Milanta

October 25, 2021

Forbidden Words (20 points)

We are given n bits and m bit strings (forbidden words) $S_1, S_2, \ldots, S_m$. We say an n bit string is legal if it does not contain any S_i as a consecutive substring. For example, let $n = 5$, and $S_1 = 010$, then 11101 is a legal string, but 10101 is not because $x_2x_3x_4 = 010$. Moreover, suppose that $x_1 \cdots x_l$ is a forbidden word, then $\overline{x_1} \cdots x_k$ for $1 \leq k \leq l$ are not forbidden words where $\overline{x_1} = 1$ if $x_1 = 0$, and $\overline{x_1} = 0$ if $x_1 = 1$. In your algorithm, you can call a procedure LegalGenerator that generates a uniformly random legal string given its length and a set of forbidden words. Devise a fully polynomial randomized approximation scheme (FPRAS) for estimating the total number of legal n bit strings.

Hint. Can you say anything in the case $k > l$?

Solution

Let's define L_i be the number of legal bit strings of length n according to the forbidden words $S_1, \ldots S_i$. L_0 will just be the number of bit strings of length n, which is 2^n. L_m is what we need to approximate. It's easy to see that

$$L_m = \underbrace{\frac{L_m}{L_{m-1}}}_{q_m} \underbrace{\frac{L_{m-1}}{L_{m-2}}}_{q_{m-1}} \cdots \underbrace{\frac{L_2}{L_1}}_{q_2} \underbrace{\frac{L_1}{L_0}}_{q_1} \underbrace{L_0}_{2^n}.$$

Now, the idea is to approximate each of the q_i by sampling.

Estimate q_i: To estimate q_i we call LegalGenerator generating strings legal according to $S_1, \ldots, S_{i-1}$. Then we check weather the string is also legal according to S_i. We now compute the probability of such an event to occur

$$\Pr\left(\text{legal according to } S_i \mid \text{legal according to } S_1, \ldots, S_{i-1}\right) = \frac{L_i}{L_{i-1}} = q_i$$

Now, we iterate this process k times, and the ratio between the number of success and k will be an approximation of q_i. Assuming $q_i \geq 1/2$, we can use Chernoff bound to show that we get that

$$\hat{q}_i := \frac{\#\text{successes}}{k} \in [q_i(1-\tilde{\epsilon}), q_i(1+\tilde{\epsilon})] \qquad \text{with prob. } 1 - \tilde{\delta}$$
$$k \geq \frac{3}{q_i \tilde{\epsilon}^2} \log\left(\frac{2}{\tilde{\delta}}\right)$$

If we take $k \geq \frac{6}{\tilde{\epsilon}^2} \log\left(\frac{2}{\tilde{\delta}}\right)$, assuming that $q_i \geq 1/2$ (proven later), we can guarantee the same error with the same probability.

Combine errors: The number of samplings we need to run to approximate $q_i \forall i$ is just $\frac{6m}{\tilde{\epsilon}^2} \log\left(\frac{2}{\tilde{\delta}}\right)$. This is polynomial in both $m, \frac{1}{\tilde{\epsilon}}$ and $\frac{1}{\tilde{\delta}}$. Let's impose that the probability to have at least one of the q_i outside of the error boundary to be no more that δ:

$$\delta \geq \Pr(\text{at least one error}) \geq \sum_{i=1}^{m} \tilde{\delta} = m\tilde{\delta}.$$

Therefore, by taking $\tilde{\delta} = \frac{\delta}{m}$ we achieved the bound. Now, let's look at the total error assuming that all of q_i are inside the required interval:

$$\begin{aligned}\hat{L}_m := 2^n \prod_{i=1}^{m} \hat{q}_i &\in \left[(1-\tilde{\epsilon})^m \underbrace{2^n \prod_{i=1}^{m} q_i}_{L_m}, (1+\tilde{\epsilon})^m \underbrace{2^n \prod_{i=1}^{m} q_i}_{L_m}\right] \\ &\subseteq [(1-2m\tilde{\epsilon})L_m, (1+2m\tilde{\epsilon})L_m]\end{aligned}$$

Therefore, if we want the error to be less than ϵ, it's enough to take $\tilde{\epsilon} = \frac{\epsilon}{2m}$. To conclude we have an algorithm that estimates L_m with $\hat{L}_m$ with a multiplicative error of ϵ with probability no less than $1-\delta$ that runs in $O\left(\frac{6(2m)^2}{\epsilon^2} \log\left(\frac{2m}{\delta}\right)\right)$. Now we only need to prove that $q_i \geq 1/2 \forall i$.

Prove that $q_i \geq 1/2$: We can rewrite q_i as:

$$q_i = \frac{L_i}{L_{i-1}} = \frac{L_i}{\underbrace{\#\{\text{ legal according to } S_1, \dots, S_{i-1} \text{ and not to } S_i\}}_{\tilde{L}_i} + L_i}.$$

If we can show that $L_i \geq \tilde{L}_{i-1}$, then we have

$$q_i = \frac{L_i + \tilde{L}_i - \tilde{L}_i}{\tilde{L}_i + L_i} = \underbrace{1 - \underbrace{\frac{\tilde{L}_i}{\tilde{L}_i + L_i}}_{\leq 1/2}}_{\geq 1/2},$$

which is enough to conclude the proof. Now, I show that for every element in $\tilde{L}_i$ there is a unique corrispondent element in L_i. Let such element be

$$e = x_1 x_2 \cdots x_k x_{k+1} \cdots x_{k+|S_i|} \cdots x_j x_{j+1} \cdots x_{j+|S_i|}$$