LEARNING-BASED SUPPORT ESTIMATION IN SUBLINEAR TIME

Anonymous authors
Paper under double-blind review

A OMITTED PROOFS

A.1 ANALYSIS OF THE UPPER BOUND

In this subsection, we prove Theorem 1, which provides the upper bound for the sample complexity of Algorithm 1.

Proof of Theorem 1. Consider some $i \in [n]$. The predictor Π gives us an estimate $\Pi(i) \in [p_i, b \cdot p_i]$ of the unknown true probability p_i. Let j_i be the integer for which $\Pi(i) \in [\frac{b^{j_i}}{n}, \frac{b^{j_i+1}}{n}]$. We treat i as assigned by the predictor to the interval I_{j_i}. Observe that as a consequence, we have

$$p_i \in \left[\frac{b^{j_i}}{n}, \frac{b^{j_i+2}}{n}\right]. \tag{1}$$

Suppose i is observed k times in the sample. Define,

$$\tilde{S}(i) = \begin{cases} 1 + a_k \left(\frac{n}{b^j}\right)^k \cdot \frac{k!}{N^k} & \text{if } \frac{b^{j_i}}{n} \leq \frac{0.5 \log n}{N} \\ 1 & \text{if } \frac{b^{j_i}}{n} > \frac{0.5 \log n}{N} \text{ and } k \geq 1 \\ 0 & \text{if } \frac{b^{j_i}}{n} > \frac{0.5 \log n}{N} \text{ and } k = 0 \end{cases}$$

The a_k's are the Chebyshev polynomial coefficients as per Algorithm 1. Note that as we drew a total of $Pois(N)$ samples, the number of times each element i appears in the sample is distributed as $N_i = Pois(N \cdot p_i)$ times, and the values N_i are independent across different i's (this is the standard Poissonization trick). Thus the $\tilde{S}(i)$'s are also independent. Moreover, note that the value $\tilde{S}_j$ defined in Algorithm 1 satisfies $\tilde{S}_j = \sum_{i:j_i=j} \tilde{S}(i)$. Finally, by eq. (2) we have $p_i \cdot \frac{n}{b^{j_i}} \in [1, b^2]$. Thus, if $\frac{b^{j_i}}{n} \leq \frac{0.5 \log n}{N}$, we compute the expectation of $\tilde{S}(i)$ as (letting $j = j_i$ to ease notation),

$$\begin{aligned} \mathbb{E}[\tilde{S}(i)] &= 1 + \sum_{k=0}^{L} \mathbb{P}(N_i = k) \cdot a_k \left(\frac{n}{b^j}\right)^k \cdot \frac{k!}{N^k} \\ &= 1 + \sum_{k=0}^{L} \frac{e^{-Np_i}(Np_i)^k}{k!} \cdot a_k \left(\frac{n}{b^j}\right)^k \cdot \frac{k!}{N^k} \\ &= 1 + e^{-Np_i} \cdot \sum_{k=0}^{L} a_k \left(p_i \cdot \frac{n}{b^j}\right)^k \\ &= 1 + e^{-Np_i} \cdot P_L\left(p_i \cdot \frac{n}{b^j}\right) \\ &\in [1 - \varepsilon, 1 + \varepsilon], \end{aligned}$$

since $p_i \cdot \frac{n}{b^j} \in [1, b^2]$ and since $e^{-Np_i} \leq 1$. However, if $\frac{b^{j_i}}{n} > \frac{0.5 \log n}{N}$, then $\mathbb{E}[\tilde{S}(i)]$ equals the probability that i is sampled, which is $1 - (1 - p_i)^N$. But since $p_i > \frac{b^{j_i}}{n} > \frac{0.5 \log n}{N}$, we have that $1 \geq \mathbb{E}[\tilde{S}(i)] > 1 - \left(1 - \frac{0.5 \log n}{N}\right)^N > 1 - e^{(0.5 \log n/N) \cdot N} = 1 - n^{-1/2}$. We note that if i is never seen (which is possible even if $p_i \neq 0$), we do not know j_i. However, in this case, $\tilde{S}(i) = 0$ regardless of the value of j_i. Therefore, we get an estimator $\tilde{S}(i)$ such that $\tilde{S}(i) = 0$ if $p_i = 0$ and $\mathbb{E}[\tilde{S}(i)] \in [1 - \varepsilon, 1 + \varepsilon]$ otherwise, assuming $\varepsilon > n^{-1/2}$.

Next we analyze the variance of $\tilde{S}(i)$. For i with $p_i = 0$, $\tilde{S}(i) = 0$ always, so $\mathbf{Var}(\tilde{S}(i)) = 0$. Otherwise, if $\frac{b^{j_i}}{n} > \frac{0.5 \log n}{N}$, then $\tilde{S}(i)$ is always between 0 and 1, so $\mathbf{Var}(\tilde{S}(i)) \leq 1$. Finally, if

$\frac{b^{j_i}}{n} \leq \frac{0.5 \log n}{N}$, then $\mathbf{Var}(\tilde{S}(i)) \leq \mathbb{E}[(\tilde{S}(i)-1)^2]$ and we can write (again with $j = j_i$)

$$\begin{aligned}
\mathbb{E}[(\tilde{S}(i)-1)^2] &= \sum_{k=0}^{L} \mathbb{P}(N_i = k) \cdot \left(a_k \left(p_i \cdot \frac{n}{b^j} \right)^k \right)^2 \\
&= \sum_{k=0}^{L} \frac{e^{-Np_i}(Np_i)^k}{k!} \cdot \left(a_k \left(\frac{n}{b^j} \right)^k \cdot \frac{k!}{N^k} \right)^2 \\
&\leq \left(\max_{0 \leq k \leq L} a_k^2 \right) \cdot \sum_{k=0}^{L} \left(p_i \cdot \frac{n}{b^j} \right)^k \cdot \left(\frac{n}{b^j} \right)^k \cdot \frac{k!}{N^k} \\
&\leq \left(\max_{0 \leq k \leq L} a_k^2 \right) \cdot \sum_{k=0}^{L} \left(\frac{b^2 \cdot k \cdot n}{N} \right)^k, \qquad (2)
\end{aligned}$$

where for the last inequality we noted that $k! \leq k^k$, $p_i \cdot \frac{n}{b^j} \leq b^2$, and $b^j \geq 1$.

Now, it is a well known consequence of the Markov Brothers' inequality that all the coefficients of the standard degree L Chebyshev polynomial $Q_L(x)$ are bounded by $e^{O(L)}$ (Markov, 1892). Since $P_L = \sum_{k=0}^{L} a_k x^k$ is just $\varepsilon \cdot Q\left(\frac{b^2+1}{2} + x \cdot \frac{b^2-1}{2} \right)$, we have that for any fixed $b = O(1)$, the coefficients a_k are all bounded by $e^{O(L)}$ as well. Therefore, for $N = C \cdot b^2 \cdot L \cdot n^{1-1/L}$ for some constant C, we have that $b^2 \cdot k \cdot n/N \leq n^{1/L}/C$, so we can bound Equation (3) by

$$e^{O(L)} \cdot \sum_{k=0}^{L} \left(\frac{n^{1/L}}{C} \right)^k \leq \frac{n}{(C')^L}$$

for some other constant C'.

In summary, if $p_i \neq 0$, we have that $\mathbb{E}[\tilde{S}(i)] \in [1-\varepsilon, 1+\varepsilon]$ and $\mathbf{Var}(\tilde{S}(i)) \leq \varepsilon^2 n$ if we choose C to be a sufficiently large constant, since $\varepsilon = e^{-\Theta(L)}$ for a constant $1 < b \leq O(1)$. This is true even for $\frac{b^{j_i}}{n} > \frac{0.5 \log n}{N}$ since $\varepsilon^2 n > 1$. However, we know that $\tilde{S} = \sum_{j=0}^{\log_b n} \tilde{S}_j = \sum_{i=1}^{n} \tilde{S}(i)$, since by the definition of $\tilde{S}_j$, $\tilde{S}_j = \sum_{i:j_i=j} \tilde{S}(i)$. Therefore, since the estimators $\tilde{S}(i)$ are independent, we have that $\mathbb{E}[\tilde{S}] = \sum_{i=1}^{n} \mathbb{E}[\tilde{S}(i)] \in (1-\varepsilon, 1+\varepsilon) \cdot S$ and $\mathbf{Var}(\tilde{S}) = \sum_{i:p_i \neq 0} \mathbf{Var}(\tilde{S}(i)) \leq \varepsilon^2 \cdot n \cdot S$, where we use the independence of the $\tilde{S}(i)$'s. Therefore, since $S \leq n$, with probability at least 0.9, $|\tilde{S} - S| = O(\varepsilon) \cdot \sqrt{n \cdot S}$ by Chebyshev's inequality. □

Remark. *We note that our analysis actually works (and becomes somewhat simpler) even if the algorithm is modified to define* $\tilde{S}_j = \sum_{i \in [n]: \Pi(i) \in I_j} \left(1 + a_{N_i} \left(\frac{n}{b^j} \right)^{N_i} \cdot \frac{N_i!}{N^{N_i}} \right)$ *for all intervals, as opposed to* $\tilde{S}_j = \#\{i \in [n] : N_i \geq 1, \Pi(i) \in I_j\}$ *for the intervals* I_j *with* $\frac{b^j}{n} > \frac{0.5 \log n}{N}$. *However, because we can decrease the bias as well as the variance for these larger intervals, we modify the algorithm accordingly. While this does not affect the theoretical guarantees, it demonstrated an improvement in practice. This threshold was also used in Wu & Yang (2019) (the choice of leading constant* 0.5 *in the threshold term* $\frac{0.5 \log n}{N}$ *is arbitrary, and for simplicity we have adopted the value they used).*

A.2 Analysis of the Lower Bound

In this subsection, we prove Theorem 2, which proves that the sample complexity of Algorithm 1 is essentially tight.

Proof of Theorem 2. In order to prove the lower bound, we shall define two distributions P and Q for which the first k moments are matching, but their support size differs on at least εn elements. Furthermore, both distributions will be supported on $\{0\} \cup \left[\frac{k}{n}, \frac{k+1}{n}, \ldots, \frac{2k}{n} \right]$, so that a 2-factor approximation predictor does not provide any useful information to an algorithm trying to distinguish the two distributions.

We start with an overview of the definition of the two distributions and then explain how to formalize the arguments, following the Poissonnization and rounding techniques detailed in Raskhodnikova et al. (2009).

Let $\varepsilon = \left(k \cdot 2^{k-1} \cdot \binom{2k}{k}\right)^{-1}$ for some integer $k \geq 1$. Note that $\varepsilon = e^{-\Theta(k)}$. In order to define the distributions, we define $k+1$ real numbers $a_0, \ldots, a_k$ as $a_i = \frac{(-1)^i \cdot \binom{k}{i}}{2^{k-1} \cdot (k+i)}$ for every $i \in \{0, \ldots, k\}$. Suppose for now that n is a multiple of the least common multiple of $2^{k-1}, k, \ldots, 2k$, so that $a_i \cdot n$ is an integer for every i. We define P and Q as follows:

- **The distribution P:** For every a_i such that $a_i > 0$, the support of P contains $a_i \cdot n$ elements j that have probability $p_j = a_i \cdot \frac{k+i}{n}$ each.
- **The distribution Q**: For every a_i such that $a_i < 0$, the support of Q contains $-a_i \cdot n$ elements j that have probability $q_j = -a_i \cdot \frac{k+i}{n}$ each.

First we prove that P and Q are valid distributions and that all of their non-zero probabilities are greater than $1/n$.

Claim A.1. *P and Q as defined above are distributions. Furthermore, their probability values are either 0 or greater than $1/n$.*

Proof. The second part of the claim follows directly from the definition of P and Q. We continue to prove the first part, that P and Q are indeed distributions. It holds for P that

$$\sum_{a_i | a_i > 0} (n a_i) \cdot \frac{k+i}{n} = \sum_{\substack{i=0 \\ i \text{ even}}}^{k} n \cdot \frac{1}{k+i} \cdot \binom{k}{i} \cdot \frac{1}{2^{k-1}} \cdot \frac{k+i}{n} = \frac{1}{2^{k-1}} \cdot \sum_{\substack{i=0 \\ i \text{ even}}}^{k} \binom{k}{i} = 1.$$

Similarly, for Q,

$$\sum_{a_i < 0} (n(-a_i)) \cdot \frac{k+i}{n} = \sum_{\substack{i=0 \\ i \text{ odd}}}^{k} n \cdot \frac{1}{k+i} \cdot \binom{k}{i} \cdot \frac{1}{2^{k-1}} \cdot \frac{k+i}{n} = \frac{1}{2^{k-1}} \cdot \sum_{\substack{i=0 \\ i \text{ odd}}}^{k} \binom{k}{i} = 1. \qquad \square$$

We continue to prove that the first k moments of P and Q are matching, and that their support size differs by εn.

Claim A.2. *Let $a_1, \ldots, a_k$ be defined as above. Then*

- *For any $r \in \{1, \ldots, k\}$, it holds that $\sum_{i=0}^{k} a_i \cdot (k+i)^r = 0$.*
- *$\sum_{i=0}^{k} a_i = \varepsilon$.*

Proof. By plugging the a_i's as defined above,

$$\sum_{i=0}^{k} a_i \cdot (k+i)^r = \sum_{i=0}^{k} \frac{(-1)^i \binom{k}{i}}{2^{k-1} \cdot (k+i)} \cdot (k+i)^r = \frac{1}{2^{k-1}} \sum_{i=0}^{k} (-1)^i \binom{k}{i} \cdot (k+i)^{r-1}.$$

Hence, letting $r' = r-1$, it suffices to prove that $\sum_{i=0}^{k} (-1)^i \binom{k}{i} (k+i)^{r'} = 0$ for all $0 \leq r' \leq k-1$. For any fixed k, note that since $(k+i)^{r'}$ is a degree r' polynomial in i, $(k+i)^{r'}$ can be written as a linear combination of $\binom{i}{s}$ for $0 \leq s \leq r'$ with coefficients $b_0, \ldots, b_{r'}$. Therefore, we would like to prove that: $\sum_{i=0}^{k} (-1)^i \binom{k}{i} \sum_{s=0}^{r'} b_s \binom{i}{s} = 0$. Fix some s in $\{0, \ldots, r'\}$: it suffices to show that for any integer k, $\sum_{i=0}^{k} (-1)^i \binom{k}{i} \binom{i}{s} = 0$ for all $0 \leq s \leq k-1$.

Since $\binom{k}{i}\cdot\binom{i}{s}=\binom{k}{k-i,i-s,s}=\binom{k}{s}\cdot\binom{k-s}{i-s}$, by setting $k'=k-s$ and $i'=i-s$, we get

$$\sum_{i=0}^{k}(-1)^i\binom{k}{i}\binom{i}{s}=\sum_{i=0}^{k}(-1)^i\binom{k}{s}\cdot\binom{k-s}{i-s}=\binom{k}{s}\cdot\sum_{i'=0}^{k'}(-1)^{i'+s}\binom{k'}{i'}$$
$$=\binom{k}{s}\cdot(-1)^s\cdot(1-1)^{k'}=0,$$

where the last equality is since $k'=k-s\geq 1$. This concludes the proof of the first item.

We continue to prove the second item in the claim:

$$\sum_{i=0}^{k}a_i=\varepsilon. \tag{3}$$

Recall that $a_i=\frac{(-1)^i\cdot\binom{k}{i}}{2^{k-1}\cdot(k+i)}$, and that $\varepsilon=\left(k\cdot 2^{k-1}\cdot\binom{2k}{k}\right)^{-1}$, so plugging these into Equation (4) and multiplying both sides by $2^{k-1}\cdot\binom{2k}{k}$, this is equivalent to proving

$$\frac{1}{k}=\sum_{i=0}^{k}\frac{(-1)^i}{k+i}\cdot\binom{k}{i}\binom{2k}{k}. \tag{4}$$

Since $\binom{k}{i}\cdot\binom{2k}{k}=\binom{2k}{k,i,k-i}=\binom{2k}{k+i}\cdot\binom{k+i}{k}$, the right-hand side of Equation (5) equals

$$\sum_{i=0}^{k}\frac{(-1)^i}{k+i}\cdot\binom{2k}{k+i}\binom{k+i}{k}=\sum_{i=0}^{k}(-1)^i\cdot\binom{2k}{k+i}\cdot\binom{k+i-1}{k-1}\cdot\frac{1}{k}.$$

Multiplying by k, it suffices to show that

$$1=\sum_{i=0}^{k}(-1)^i\cdot\binom{2k}{k+i}\cdot\binom{k+i-1}{k-1}. \tag{5}$$

To do this, let $j=k+i$. Then, note that $\binom{k+i-1}{k-1}=\binom{j-1}{k-1}=\frac{(j-1)\cdots(j-k+1)}{(k-1)!}$ which is a degree $k-1$ polynomial in j. For $1\leq j\leq k-1$ the polynomial equals 0, but for $j=0$ the polynomial equals $(-1)^{k-1}$. Therefore, the right hand side of Equation (6) equals

$$\sum_{j=1}^{2k}(-1)^{j-k}\cdot\binom{2k}{j}\cdot\frac{(j-1)\cdots(j-k+1)}{(k-1)!}=1+(-1)^{-k}\cdot\sum_{j=0}^{2k}(-1)^j\cdot\binom{2k}{j}\cdot\frac{(j-1)\cdots(j-k+1)}{(k-1)!}.$$

As proven in the first part of this proof, for any $P(j)$ of degree $0\leq r\leq 2k-1$, $\sum_{j=0}^{2k}(-1)^j\binom{2k}{j}P(j)=0$. Therefore, the summation on the right hand side is 0, so this simplifies to 1, as required. □

The following corollary follows directly from the definition of P and Q and the previous claim.

Corollary A.1. *The following two items hold for P and Q as defined above.*

- $\mathbb{E}[P^r]=\mathbb{E}[Q^r]$ *for all* $r\in\{1,\ldots,k\}$.
- $TV(P,Q)=\varepsilon n$.

The above corollary states that indeed P and Q as defined above have matching moments for $r=1$ to k, and that they differ by εn in their support size. This concludes the high level view of the construction of the distributions P and Q. In order to finalize the proof we rely on the standard Possionization and rounding techniques.

First, by Theorem 5.3 in Raskhodnikova et al. (2009), any s-samples algorithm can be simulated by an $O(s)$-Poisson algorithm. Hence, we can assume that the algorithm takes $Poi(s)$ samples, rather than an arbitrary number s. Second, we can alleviate the assumption that the $a_i\cdot n$ values (similarly

$-a_i \cdot n$) are integral for all i, by rounding down the value in case it is not integral, and choosing $n - \sum_{i=0}^{k} \lceil a_i \cdot n \rceil$ additional elements in P so that each has probability $1/n$ (and analogously for Q). By Claim 5.5 in Raskhodnikova et al. (2009) this process increases the number of values in P and Q by at most $O(k^2)$. Hence, the distributions are now well defined, and we can rely on the following theorem from Raskhodnikova et al. (2009).

Theorem 1 (Corollary 5.7 in Raskhodnikova et al. (2009), restated.)**.** *Let P and Q be random variables over positive integers $b_1 < \ldots < b_{k-1}$, that have matching moments 1 through $k-1$. Then for any Poission-s algorithm $\mathcal{A}$ that succeeds to distinguish P and Q with high probability, $s = \Omega(n^{1-1/k})$.*

Therefore, plugging the value of ε, we get an $s = \Omega(n^{1-\Theta(\log(1/\varepsilon))})$ lower bound as claimed. □