ECE 6980
Algorithmic and Information-Theoretic Methods in Data Science

Instructor: Jayadev Acharya
Scribe: Sourbh Bhadane and Kai Zhang

Lecture #12
11th October, 2017

Recap

In the previous lecture, we studied estimating symmetric properties of distributions. We saw the following

1. Maximum likelihood estimator requires $O\left(\frac{k}{\epsilon}\right)$ samples for estimating entropy, whereas $\Theta\left(\frac{k}{\epsilon \log k}\right)$ suffice.
2. Optimal entropy estimators require estimating polynomial approximations of the entropy function.

1 Introduction

In this lecture, we will look at a recent paper [1] that proposes a simple estimator based on a modification of the maximum likelihood estimator. This approach turns out to be sample competitive for all symmetric properties.

1.1 Maximum Likelihood Principle

Maximum likelihood estimators or sequence maximum likelihood estimators (SML) are one of the most common distribution estimators dating back to Fisher. Given a sample $X^n \in \triangle_k^n$, the SML estimate assigns $\hat{p}_n = \arg\max_p p(x^n)$. It is easy to see that $\hat{p}_n(x) = \frac{N_x}{n}$ where N_x is the number of times symbol $x \in \triangle_k$ appears in X^n. Plug-in estimators based on SML are used to estimate symmetric properties of the distributions.

As seen in the previous lectures, the SML estimate is not optimal in the large-alphabet regime. One of the drawbacks of SML is that it tries to learn the entire distribution and ends up overfitting. In addition, SML estimators cannot obtain sublinear sample complexity since learning the entire distribution requires linear number of samples.

2 Profile Maximum Likelihood

The central idea behind profile maximum likelihood (PML) [2] is to output a distribution that maximizes the likelihood of a sufficient statistic, which in the case of estimating symmetric properties is the profile of the sequence. Recall that the profile of a sequence is defined as $\phi(X^n) = \{N_x, x \in \triangle_k\}$. The probability of a profile ϕ is defined as

$$p(\phi) = \sum_{X^n : \phi(X^n) = \phi} p(X^n)$$

Suppose we want to estimate the symmetric function $f(.)$ on the class of distributions $\mathcal{P}$, given the sample X^n. The PML estimation scheme is the following

1. Compute $p_\phi = \arg\max_{p \in \mathcal{P}} p\left(\phi\left(X^n\right)\right)$
2. Output $f\left(p_\phi\right)$

Example 1. *Let $X^3 = \{a, a, b\}$ be a sequence generated from a distribution over $\{a, b, c\}$. Clearly the SML distribution is $\left\{\frac{2}{3}, \frac{1}{3}, 0\right\}$. The probability of the profile of this sequence, $\{1, 2\}$ is*

$$p\left(\{1,2\}\right) = \binom{3}{1}\left(p(a)^2 p(b) + p(b)^2 p(a) + p(a)^2 p(c) + p(c)^2 p(a) + p(c)^2 p(b) + p(b)^2 p(c)\right)$$

For the special case where the support size of the PML distribution is 2, the following argument shows that the uniform distribution $\left\{\frac{1}{2}, \frac{1}{2}\right\}$ achieves the maximum.

$$\sum_{a \neq b} p(a)^2 p(b) = \sum_a p(a) \times p(a)\, p(1-a) \leq \frac{1}{4}$$

[2] shows that the above statement holds even if the support size of the PML distribution is not known. Moreover, PML can also predict new symbols. For instance, if $X^4 = \{a, b, a, c)$ and $\phi\left(X^4\right) = \{1, 1, 2\}$, it can be shown that the PML distribution is the uniform distribution over 5 elements $\left(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right)$.

3 Performance of MLE

Since PML is a maximum likelihood based approach applied to profiles of a sequence, it is worthwhile to examine what performance guarantees an ML based approach offers in general. Let $\mathcal{P}$ be a collection of distributions over $\mathcal{Z}$ and $f\colon \mathcal{P} \to \mathbb{R}$. Given a sample Z from an unknown $p \in \mathcal{P}$, we want to estimate $f(p)$. Denote the maximum likelihood distribution after observing Z as p_Z.

Theorem 2. *Suppose there is an estimator $\hat{f}\colon \mathcal{Z} \to \mathbb{R}$ such that for all $p \in \mathcal{P}$, $Z \sim p$,*

$$\Pr\left(|f(p) - \hat{f}(Z)| > \epsilon\right) < \delta,$$

then

$$\Pr\left(|f(p) - f\left(p_Z\right)| > 2\epsilon\right) < \delta|\mathcal{Z}|$$

Proof. **Case 1:** Consider $z \in \mathcal{Z}$ such that $p(z) \geq \delta$. Clearly $|f(p) - \hat{f}(z)| \leq \epsilon$ for all such z. Since $p_Z(z) \geq p(z) \geq \delta$, $|f(p_z) - \hat{f}(z)| \leq \epsilon$. Therefore, by triangle inequality, $|f(p_z) - f(p)| \leq 2\delta$ for all such z. This proves that the required guarantee is met without any error for all z such that $p(z) \geq \delta$.

Case 2: Consider $z \in \mathcal{Z}$ such that $p(z) < \delta$. Since any error in the performance occurs when $p(z) < \delta$, for $Z \sim p$,

$$\Pr(error) \leq \Pr(p(Z) < \delta) \leq \sum_{z \in \mathcal{Z}: p(z) < \delta} p(z) < \delta|\mathcal{Z}|$$

□

Although the above theorem holds for general $\mathcal{Z}$, to estimate symmetric properties of distributions we let $\mathcal{Z} = \Phi^n$, all possible profiles of length n sequences. A similar result holds for (multiplicative) approximate ML distributions [1].

3.1 Upper bound on partition size

We now present upper bounds on $|\Phi^n|$ which is equal to the partition number of n. The following compression argument gives a $2^{n \log n}$ upper bound on the partition number.

Lemma 3.

$$|\Phi^n| \leq 2^{2\sqrt{n}\log_2 n}$$

Proof. We encode each profile of an n-length sequence using $2\sqrt{n}\log_2 n$ bits. Therefore the total number of profiles of n is upper bounded by $2^{2\sqrt{n}\log n}$.

Consider a profile ϕ. For all x such that $N_x = t < \sqrt{n}$, encode the number of x such that $N_x = t$ using $\log_2 n$ bits. For x such that $N_x = t \geq \sqrt{n}$, encode each x separately using $\log_2 n$ bits. The number of such x is atmost $\sqrt{n}$ since the sum of all entries in the profile is n. □

We will use a tighter bound of $e^{3\sqrt{n}}$ due to Hardy and Ramanujan. [3]

3.2 Competitiveness of ML - Median Trick

We now establish sample competitiveness of ML based approaches using the median trick in the informal theorem below. Recall the median trick from assignment 1 problem 6 : if we can solve a problem using n independent samples with error probability atmost 0.1, then the same problem can be solved with error probability e^{-cm} using $O(nm)$ samples where c is a constant.

Theorem 4. *If $\hat{f} : \mathcal{Z} \to \mathbb{R}$ is an estimator of $f(p)$ with sample complexity n, accuracy ϵ and constant error probability, the ML estimator $f(p_Z)$ achieves accuracy 2ϵ with sample complexity $O(n^2)$ and constant error probability*

Proof. Using the median trick based estimator, Theorem 2 then implies that the error probability of the ML estimator for accuracy 2ϵ is $e^{-cm} * |\mathcal{Z}|$. For the PML based approach, using the partition number bound in the previous section, the error probability of the PML estimator is $e^{-cm} * e^{3\sqrt{nm}}$. Therefore, $O(n)$ copies are necessary to get a constant error probability implying that the PML estimator has a sample complexity of $O(n^2)$. □

Note that the above result has both drawbacks and advantages. The drawback being the sample complexity is now quadratic in n and the advantage being that the sample complexity is not dependent on k or ϵ. In the following section we provide an intuition as to how PML does better than quadratic sample complexity.

4 Optimality of PML

Recall from previous lectures (Lecture 9) that if an estimator has a small bounded difference constant we can use McDiarmid's inequality to drive the error probability down faster than the median trick. Specifically, we saw that learning a discrete distribution can be done in $\Theta\left(\frac{k}{\epsilon^2}\right)$ samples with error probability e^{-k}. Similarly, to drive drown the error probability of PML, [1] modify

existing sample-optimal estimators so that the bounded difference of the modified estimators is small. We state this in the form of the following lemma proved in [1].

Lemma 5. *For a fixed constant $\alpha > 0$, there exist PML based entropy estimators with optimal sample complexity and bounded difference constant $c\frac{n^\alpha}{n}$ where c is a positive constant.*

The above lemma holds for other symmetric properties like support size, support coverage and distance to uniformity. Given a lower bound on the bias of the PML estimator, one can show that the PML estimator estimates entropy with 4ϵ accuracy and error probability $e^{-\sqrt{n}}$.

An interesting fact shown in [1] is that the above result holds even for approximate ML distributions. The approximation can be as weak as $e^{-\sqrt{n}}$.

References

[1] J. Acharya, H. Das, A. Orlitsky, and A.T. Suresh. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *ICML*, pages 11–21, 2017.

[2] A Orlitsky, N. P. Santhanam, K Viswanathan, and J Zhang. On modeling profiles instead of values. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 426–435, 2004.

[3] G. H. Hardy and S. Ramanujan. Asymptotic formula in combinatory analysis. *Proceedings of the London Mathematical Society*, s2-17(1):75–115, 1918.