Permanents and modeling probability distributions

organized by
Alon Orlitsky, Narayana Santhanam, and Krishnamurthy Viswanathan

Workshop Summary

The goal of the workshop was to further recent developments on estimating a probability distribution p over a countable *alphabet* $\mathcal{A}$ if we are allowed to sample with replacement from the distribution very few times. The workshop approached the problem mostly using the *pattern* technique, described below. Given a sequence $\bar{x} = x_1 x_2 \to x_n \in \mathcal{A}^n$ of symbols drawn from a discrete alphabet $\mathcal{A}$, the pattern of $\bar{x}$, denoted $\bar{\psi}(\bar{x})$, is the sequence obtained by replacing each symbol in $\bar{x}$ by its order of appearance. For example, the pattern of $\bar{x} =$ "*abracadabra*″ is $\bar{\psi}(\bar{x}) = 12314151231$. The connection to permanents arises from the fact that when the sequence $\bar{x}$ consists of independent and identically distributed samples, the probability induced on the pattern can be written as the permanent of an appropropriately constructed matrix. In addition to patterns, the workshop also considered alternative approaches arising from the study of exchangeable random partitions and Bayesian estimation, the *large number of rare events* regime, Good-Turing estimation, and other divergence-based approaches.

The workshop featured talks on selected topics in the mornings followed by break-out sessions in which attendees split into groups to work on open problems. We first describe the talks presented and then the problems discussed in the break-out sessions.

Presentations

The presentations were on the following topics.

- An introduction to probability estimation and distribution modeling. Results derived through the pattern-based approach to probability estimation where the estimation quality is measured in terms of the *attenuation*—the rate at which we underestimate the true probability of the given sequence. These results were contrasted with those that could be obtained using extensions of the Laplace estimator.
- A more detailed lecture on probability estimation. The main focus of the lecture was a step-by-step exposition of the proof that pattern-based estimator achieves an attenuation of 1 for all observed sequences. Good-Turing and Laplace estimators were also analyzed.
- Distribution modeling where the goal is to estimate statistics that do not depend on the support set of the distribution. In particular, the pattern maximum-likelihood was considered. Since deriving the pattern maximum-likelihood is difficult in general, results on specific profiles were presented and more general results were conjectured based on them.
- A cross-section of algorithms and results on computing the permanent of a matrix, and the implications of these results to the problem of computing and thereby maximizing the probability of a pattern.

- Approaches to the probability estimation problem that are not based on patterns were also explored, including an introduction to exchangeable random partitions and a demonstration of the connection between patterns of *i.i.d.* processes and paint-box processes developed by Kingman. The talk also included subjects from Bayesian estimation suggesting the possibility that the pattern-based estimators could be written as a Baye's estimator.
- The *large number of rare events* (LNRE) setting. In this setting, all symbol probabilities are on the order of 1 over the sample size, and it captures the regime where the sample- and alphabet-sizes are comparably large. In particular, the talk focussed on estimators derived for the probability estimation problem in this setting.
- Additional talks on probability estimation based on the K-L divergence, Dirichlet processes, and sticky channels.

Open problems

The afternoon sessions were devoted entirely to open problems that either existed at the beginning of the workshop, or emerged as a result of talks and discussions. The first afternoon featured an open problem session where a number of interesting problems related to patterns and probability estimation were proposed. Subsequently the attendees split into smaller groups to tackle a subset of these problems. We describe the problems considered and the progress achieved below.

- Connections to Bayes estimation: Is it possible to recast the pattern-based diminishing attenuation estimators as a Bayesian estimators? and if not, do classical Bayesian priors on infinite alphabets, such as Dirichlet, result in a Bayesian estimator that achieves diminishing attenuation in the non-Bayesian setting? While the discussions during the workshop did not resolve the questions fully, the first question has been, interestingly, answered in the negative subsequent to the workshop. In addition, collaborations starting with the workshop have analyzed the attenuation of some Bayesian estimators, in particular, of the Ewens sampling formula.
- Consistency of pattern maximum-likelihood: Let $\hat{P}_{\bar{\psi}}$ denote the distribution that maximizes the probability of the pattern $\bar{\psi}$. It is known that this estimate is consistent for all discrete distributions over finite or countably infinite alphabets. However it is also known that no estimator is uniformly consistent for this class of distributions. Therefore a weaker uniform consistency result was desired. The following set-up was examined. For any positive integer k, let $\mathcal{P}(k)$ denote the set of all discrete distributions with support of size at most k. Let $||\cdot||_1$ denote the ℓ_1 distance. Let X^n denote an *i.i.d.* sequence of random variables, $\bar{\psi}(X^n)$ denote its pattern. For any non-decreasing function $f(n)$, the pattern maximum-likelihood is said to be uniformly consistent for $\mathcal{P}(f(n))$ if

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}(f(n))} Pr\Big(||P - \hat{P}_{\bar{\psi}}(X^n)||_1 \geq \tau\Big) = 0.$$

  From results on property testing in Computer Science, it was observed that such a result is unlikely to be true for $f(n) = \Omega(n\log n)$ and is likely to be true for $f(n) = \mathcal{O}\Big(\frac{n}{\log n}\Big)$. Some progress was made in resolving the residual ambiguity by applying techniques used for the probability estimation problem in the LNRE regime.

- Two-dimensional patterns: How does one handle memory? To this end, candidates were proposed for two-dimensional patterns that would model dependencies between two random variables. Discussions during the workshop revealed that even elementary computations in this scenario—for example, the maximum probability of patterns obtained by sampling a pair of random variables (X, Y) thrice—was non-trivial. However, discussions that continued after the workshop among the participants have resolved the elementary problems posed in the workshop, and we have moved into thinking about extending this framework in more detail.
- Largest alphabet size of pattern maximum-likelihood estimate: For a given pattern $\bar{\psi}$ let $\hat{P}_{\bar{\psi}}$ denote the distribution maximizing the pattern probability. For any distribution, P let $S(P)$ denote its support size. It was conjectured based on certain patterns whose $\hat{P}_{\bar{\psi}}$ can be derived in closed form, that for all patterns if $\hat{P}_{\bar{\psi}}$ is discrete, then

  $$S(\hat{P}_{\bar{\psi}}) \leq \frac{m^2}{2}.$$

  Attempts were made to prove this conjecture during the break-out sessions and some promising approaches were uncovered.
- Computation of pattern maximum-likelihood: Discussions were held on efficient ways to compute the probability of a pattern and thereby the pattern maximum-likelihood. These focussed on improving the existing Markov chain monte carlo techniques that are employed as for computing pattern probabilities in an expectation-maximization algorithm designed to compute $\hat{P}_{\bar{\psi}}$.