1 Learning / Testing Structured Distributions

In previous lectures, we mostly focused on distribution learning and testing where the target distribution is potentially arbitrary over a discrete domain $[n]$. For instance, we showed that $\Theta(n/\epsilon^2)$ samples are necessary and sufficient for learning a general discrete distribution over $[n]$ within total variation distance $\epsilon$.

When the set of target distributions is restricted to some structured family $\mathcal{D}$, where e.g., $\mathcal{D}$ is a subset of distributions over $[n]$ (monotone distributions, log-concave distributions, $k$-modal distributions, etc.), we can potentially obtain significantly better sample and time complexity upper bounds. In other words, we can algorithmically exploit the underlying structure of the distribution family to perform statistical estimation more efficiently.

Example 1. Suppose $p$ is a log-concave distribution over $[n]$ (or $\mathbb{R}$). The sample complexity of uniformity testing in $\ell_1$-distance is $\Theta(\epsilon^{-9/4})$, which is independent of the domain size.

1.1 Learning / Testing Univariate Structured Distributions

In this subsection, we describe a general algorithmic framework for learning and testing univariate structured distributions. The main idea of this approach is that the existence of good piecewise polynomial approximation for the distribution family $\mathcal{D}$ can be leveraged for the design of efficient learning and testing algorithms.

We say that a continuous distribution $q$ over an ordered domain (e.g., $[0,1]$ or $[n]$) is a $k$-piecewise degree-$d$ distribution if there exists a partition of the domain into $k$ disjoint intervals $I_1, \ldots, I_k$ such that $q(x) = q_j(x)$ for all $x \in I_j$, where $q_1, \ldots, q_k$ is a univariate polynomial of degree at most $d$. Let $\mathcal{P}_{k,d}$ denote the class of all $k$-piecewise degree-$d$ distribution over $[0,1]$.

Theorem 2 ([CDSS14]). Let $p$ be any probability density function over $[0,1]$. There is an algorithm that, given $k, d, \epsilon$ and $\tilde{O}\left(\frac{k(d+1)^4}{\epsilon^2}\right)$ samples from $p$, runs in time $\text{poly}(k, d + \epsilon^{-2})$.
1, 1/\epsilon) and with high constant probability outputs an \(O(k)\)-piecewise degree-\(d\) distribution \(q\) such that \(d_{TV}(p, q) \leq O(\text{opt}_{k,d}) + \epsilon\), where \(\text{opt}_{k,d} = \inf_{r \in \mathcal{P}_{k,d}} d_{TV}(p, r)\) is the error of the best \(k\)-piecewise degree-\(d\) distribution for \(p\).

It is shown in [CDSS14] that the sample complexity of the aforementioned algorithm is information-theoretically optimal in all three parameters up to logarithmic factors. Subsequent work has given an algorithm for this problem with optimal sample complexity (within a constant factor) that runs in nearly-linear time.

Now we describe how to apply Theorem 2 to obtain an efficient learning algorithm given the class \(D\) that we want to learn.

• Prove that any distribution in the class \(D\) is \(\epsilon\)-close in total variation distance to a \(k\)-piecewise degree-\(d\) distribution, for some appropriate values of \(k\) and \(d\).

• Agnostically learn the target distribution using the class of \(k\)-piecewise degree-\(d\) distributions as a hypothesis class.

As a consequence of Theorem 2, the algorithm will output an \(O(k)\)-piecewise degree-\(d\) distribution \(q\) such that \(d_{TV}(p, q) \leq O(\text{opt}_{k,d}) + \epsilon \leq 2\epsilon\). Furthermore, this framework yields near-sample optimal and computationally efficient estimators for a very broad class of structured distribution families, including \(k\)-modal distributions, log-concave distributions, and others.

For univariate structured distribution testing, recent work [DKN15] gave a general identity testing algorithm that applies to a broad range of distribution classes.

**Theorem 3.** Let \(D\) be a distribution class over \([n]\) such that the probability density functions of any two \(p, p' \in D\) cross essentially at most \(k\) times, which means that most of the total variation distance between \(p\) and \(p'\) comes from at most \(k\) different intervals, one each of which one has either \(p > p'\) or \(p < p'\). Then given an explicit distribution \(q\) over \([n]\) and samples from an unknown distribution \(p \in D\), one can test identity of \(p\) and \(q\) with \(O\left(\sqrt{k} \frac{\epsilon}{\epsilon^2}\right)\) samples.

As a direct application of Theorem 3 and applying the aforementioned piecewise polynomial approximation result from [CDSS14], they obtain identity testers for \(k\)-piecewise constant distributions (sample complexity \(O(\sqrt{k}/\epsilon^2)\)), \(k\)-piecewise degree-\(d\) polynomial distributions (\(O(\sqrt{k(d+1)}/\epsilon^2)\)), \(k\)-mixtures of log-concave distributions (\(\sqrt{k} \cdot \tilde{O}(1/\epsilon^{9/4})\)), and \(t\)-mixtures of \(k\)-modal distributions (\(O(\sqrt{kt \log n}/\epsilon^{5/2})\)).

**Remark 1.** Although this piecewise polynomial approximation framework is powerful for designing efficient learning and testing algorithms for a broad class of structured
distributions, it does not always yield information-theoretically sample optimal solutions. For example, we can not obtain a tester with optimal sample complexity $O(1/\epsilon^2)$ for testing uniformity of a monotone distribution by directly applying this method.

1.2 Learning / Testing Multivariate Structured Distributions

The problem of learning and testing structured distributions over $\mathbb{R}^d, d > 1$ has been studied in statistics and machine learning in many settings. Several natural high-dimensional distribution models have been considered, including general graphical models and its special cases of Bayesian Networks and Ising Models.

For low-dimensional settings, one may be able to handle learning and testing problems whose sample complexity is exponential in $d$. For instance, suppose $p$ is an arbitrary distribution over a $d$-dimensional hypercube $\{-1, +1\}^d$. We can directly obtain a uniformity tester for $p$ with sample complexity $O(2^{d/2}/\epsilon^2)$. However, for high-dimensional settings, such sample complexity exponential in $d$ is useless and our goal is to design both statistically and computationally efficient (polynomial in $d$) estimators and testers. Interestingly enough, for a wide variety of natural and important high-dimensional distribution families, there are learning and testing algorithms whose sample complexity is polynomial in the dimension $d$, but how to make them computationally efficient still remains open.

2 Tradeoffs between Different Criteria

In addition to sample and computational complexity, there are various other performance criteria including robustness to model misspecification, privacy, communication complexity, memory, and others.

- Privacy: Although the objective of machine learning is to extract useful information from data, it is important to protect data privacy in some specific settings. One of the most popular and powerful definitions of privacy is differential privacy.

- Communication complexity: When the underlying datasets are extremely large, it is natural to consider the setting where random samples are distributed on several computers. In this situation, the cost associated with communication between computers is often the dominant criterion.

- Robustness to model misspecification: As machine learning is applied to increasingly sensitive tasks, it has become important that the machine learning algorithms we develop are robust to potentially corrupted datasets.
Ideally, we want to design machine learning algorithms which are optimal on all these performance criteria, however some evidence suggests that there are essential trade-offs between them.

For example, consider the $d$-dimensional Gaussian distribution $\mathcal{N}(\mu, I)$ with unknown mean $\mu$. The sample complexity of testing whether $\mu = 0$ or $\|\mu\|_2 \geq \epsilon$ is $\Theta(\sqrt{d}/\epsilon^2)$. However in the setting where an adversary can arbitrarily change $\epsilon/100$-fraction of the samples we draw from $\mathcal{N}(\mu, I)$, the sample complexity is at least $\Omega(d)$.

References


