Tight Bounds on the Hardness of Learning Simple Nonparametric Mixtures

Bryon Aragam BRYON@CHICAGOBOOTH.EDU
University of Chicago
Wai Ming Tai WAIMING.TAI@CHICAGOBOOTH.EDU
University of Chicago

Editors: Gergely Neu and Lorenzo Rosasco

Keywords: Statistical learning theory, mixture models, latent variable models, nonparametric statistics, sample complexity

Suppose we are given n i.i.d. samples drawn from a distribution whose pdf f is of the form

$$f = w_1 f_1 + w_2 f_2, \quad w_1, w_2 > 0, \quad w_1 + w_2 = 1, \quad \text{and } f_1, f_2 \text{ are some (unknown) pdfs.} \tag{1}$$

We are interested in learning each component f_i and our goal is to study the sample complexity of this problem. Without any assumptions on f_i, this model is clearly nonidentifiable as there are infinitely many ways to decompose f into this form. To identify the model, we consider the following class of pdfs: Given an interval $I \subset \mathbb{R}$, let $\mathcal{G}_I$ be the set of pdfs that can be expressed as $g_0 * \nu$ where g_0 is the pdf of the Gaussian centered at 0 and ν is a pdf whose support is inside I.

Now, given two disjoint intervals I_1, I_2, we assume that each f_i is in $\mathcal{G}_{I_i}$. When these two intervals are unknown, we additionally assume that I_1, I_2 are well-separated to ensure the identifiability. Formally, we have the following problem:

> *Let P be a set of n i.i.d. samples drawn from $f = w_1 f_1 + w_2 f_2$, where f is defined as in (1) and $f_i \in \mathcal{G}_{I_i}$ for some unknown and well-separated intervals I_1, I_2. For a sufficiently small error $\varepsilon > 0$, what is the threshold τ_ε such that*
>
> - *if $n < \tau_\varepsilon$, then no algorithm taking P as the input returns two pdfs $\widehat{f}_1, \widehat{f}_2$ such that $\|f_i - \widehat{f}_i\|_1 < \varepsilon$ with probability at least $1 - \frac{1}{100}$ for some f?*
> - *if $n > \tau_\varepsilon$, then there is an algorithm that takes P as the input and returns two pdfs $\widehat{f}_1, \widehat{f}_2$ such that $\|f_i - \widehat{f}_i\|_1 < \varepsilon$ with probability at least $1 - \frac{1}{100}$ for any f?*

Our main result shows that $(\frac{1}{\varepsilon})^{\Omega(\log\log\frac{1}{\varepsilon})}$ samples are required to solve this problem. We exploit the fact that a single Gaussian centered anywhere can be approximated by a linear combination of Gaussians centered inside an interval. It can be proved by a quantitative version of Tauberian theorem that yields a fast rate of approximation with Gaussians, which may be of independent interest. To show this is tight, we also propose an algorithm that uses $(\frac{1}{\varepsilon})^{O(\log\log\frac{1}{\varepsilon})}$ samples to estimate each f_i. Unlike existing approaches to learning latent variable models based on moment-matching and tensor methods, our proof instead involves expanding each component with the Hermite function expansions and provides a delicate analysis of an ill-conditioned linear system involving these expansions.

Combining these bounds, we conclude that the optimal sample complexity of this problem properly lies in between polynomial and exponential, which is not common in learning theory.

. Extended abstract. Full version appears as [2203.15150, v3]