Learning arbitrary statistical mixtures of discrete distributions.

Summary: We study the problem of learning from unlabeled samples very general statistical mixture models on large finite sets. Specifically, the model to be learned, $\vartheta$, is a probability distribution over probability distributions $p$, where each such $p$ is a probability distribution over $[n] = \{1, 2, \ldots, n\}$. When we sample from mix, we do not observe $p$ directly, but only indirectly and in very noisy fashion, by sampling from $[n]$ repeatedly, independently $K$ times from the distribution $p$. The problem is to infer $\vartheta$ to high accuracy in transportation (earthmover) distance. We give the first efficient algorithms for learning this mixture model without making any restricting assumptions on the structure of the distribution $\vartheta$. We bound the quality of the solution as a function of the size of the samples $K$ and the number of samples used. Our model and results have applications to a variety of unsupervised learning scenarios, including learning topic models and collaborative filtering.

Keywords: Kantorovich-Rubinstein duality; approximation theory; convex geometry; mixture learning; randomized algorithms; spectral methods; transportation distance

doi:10.1145/2746539.2746584