Our results

Theorem. The sample complexity for learning mixtures of k Gaussians in $\mathbb{R}^d$ up to total variation distance $\varepsilon$ is

$$\tilde{O}\left(\frac{kd^2}{\varepsilon^2}\right)$$

(for general Gaussians)

$$\tilde{O}\left(\frac{kd}{\varepsilon}\right)$$

(for axis-aligned Gaussians)

Correspondingly, given n samples from the true distribution, the minimax risk is $\tilde{O}\left(\frac{kd}{\varepsilon}\right)$.

PAC Learning of Distributions

- Given i.i.d. samples from unknown target distribution $\mathcal{D}$, output $\mathcal{D}$ such that

$$d_{TV}(\mathcal{D}, \hat{\mathcal{D}}) = \sup_{E} |Pr[E] - Pr[E]| = \frac{1}{2} \|f_{\mathcal{D}} - f_{\hat{\mathcal{D}}}\|_1 \leq \varepsilon.$$

- $\mathcal{F}$: An arbitrary class of distributions (e.g. Gaussians)

- $k$-mix($\mathcal{F}$): $k$-mixtures of $\mathcal{F}$, i.e. $k$-mix($\mathcal{F}$) = $\{\sum_{i=1}^{k} w_i D_i : w_i \geq 0, \sum_i w_i = 1, D_i \in \mathcal{F}\}$

- Sample complexity of $\mathcal{F}$ is minimum number $m_{\mathcal{F}}(\varepsilon)$ such that there is an algorithm that, given $m_{\mathcal{F}}(\varepsilon)$ samples from $\mathcal{D}$, outputs $\mathcal{D}$ with $d_{TV}(\mathcal{D}, \hat{\mathcal{D}}) \leq \varepsilon$.

- PAC learning is not equivalent to parameter estimation where goal is to recover parameters of distribution.

Compression Framework

We develop a novel compression framework that uses few samples to build a representative family of distributions.

Compression Theorem

Compression Theorem [ABHLM '18].

If $\mathcal{F}$ is $(m(\varepsilon), t(\varepsilon))$-compressible then sample complexity for learning $\mathcal{F}$ is $\tilde{O}(m(\varepsilon) + \frac{t(\varepsilon)}{\varepsilon})$.

Compressibility of Mixtures

- Lemma. If $\mathcal{F}$ is $(m(\varepsilon), t(\varepsilon))$-compressible then $k$-mix($\mathcal{F}$) is $(\tilde{O}(km(\varepsilon)/\varepsilon), \tilde{O}(kt(\varepsilon)/\varepsilon))$-compressible.

Example: Gaussians in $\mathbb{R}$

- Claim. Gaussians in $\mathbb{R}$ are $(1/\varepsilon, 2)$-compressible.

  1. True distribution is $\mathcal{N}(\mu, \sigma^2)$; encoder draws $1/\varepsilon$ points from $\mathcal{N}(\mu, \sigma^2)$.

  2. With high probability, $\exists (X_i \approx \mu + \sigma, X_j \approx \mu - \sigma$).

  3. Encoder sends $X_i, X_j$; decoder recovers $\mu, \sigma$ approximately.

Outline of the Algorithm

Assume: (i) $\mathcal{F}$ is $(m(\varepsilon), t(\varepsilon))$-compressible; (ii) true dist. $\mathcal{D} \in \mathcal{F}$

Input: Error parameter $\varepsilon > 0$.

1. Draw $m(\varepsilon)$ i.i.d. samples from $\mathcal{D}$.

2. Encoder has at most $m(\varepsilon) t(\varepsilon) 2^{O(\varepsilon^2)}$ outputs so enumerate all $M = m(\varepsilon) t(\varepsilon) 2^{O(\varepsilon^2)}$ of decoder’s outputs, $D_1, \ldots, D_M$.

   By assumption, $d_{TV}(\mathcal{D}, \hat{\mathcal{D}}) \leq \varepsilon$ for some $i$.

3. Use tournament algorithm [DL '01] to find best distribution amongst $D_1, \ldots, D_M$; $O(\log(M)/\varepsilon^2)$ samples suffice for this step.

Sample complexity is $m(\varepsilon) + O(\log(M)/\varepsilon^2) = \tilde{O}(m(\varepsilon) + \frac{t(\varepsilon)}{\varepsilon})$.

Proof of Upper Bound

Lemma. Gaussians in $\mathbb{R}^d$ are $(O(d), \tilde{O}(d^2))$-compressible.

Sketch of lemma. Suppose true Gaussian is $\mathcal{N}(\mu, \Sigma)$.

- Encoder draws $O(d)$ points from $\mathcal{N}(\mu, \Sigma)$.

- Points give rough shape of ellipsoid induced by $\mu, \Sigma$: encoder sends points & $\tilde{O}(d^2)$ bits; decoder approximates ellipsoid.

- Decoder outputs $\mathcal{N}(\hat{\mu}, \hat{\Sigma})$.

Proof of upper bound. Combine lemma with compression theorem.

Lower Bound Technique

Theorem (Fano’s Inequality). If $\mathcal{D}_1, \ldots, \mathcal{D}_r$ are distributions such that $d_{TV}(\mathcal{D}_i, \mathcal{D}_j) \geq \varepsilon$ and $\text{KL}(\mathcal{D}_i, \mathcal{D}_j) \leq \varepsilon^2$ for all $i \neq j$ then sample complexity is $\Omega(\log(r)/\varepsilon^2)$.

- Use probabilistic method to find $2^{\Omega(d^2)}$ Gaussian distributions satisfying hypothesis of Fano’s Inequality.

- Repeat following procedure $2^{\Omega(d^2)}$ times:

  1. Start with identity covariance matrix.

  2. Choose random subspace $S_d$ of dimension $d/10$ & perturb eigenvalues by $\epsilon/\sqrt{\|S_d\|}$.

     Let $\Sigma$ be corresponding covariance matrix and $\mathcal{D}_S = \mathcal{N}(0, \Sigma)$.

Claim. If $\varepsilon = O(1/d^2)$ then $\text{KL}(\mathcal{D}_i, \mathcal{D}_j) \leq \varepsilon^2$ and $d_{TV}(\mathcal{D}_i, \mathcal{D}_j) \geq \varepsilon$ with probability $1 - \exp(-\Omega(d^2))$.

Can lift construction to get $2^{\Omega(kd^2)}$-k-mixture of d-dimensional Gaussians satisfying Fano’s Inequality.

Remark. Lower bound for axis-aligned proved by [SOAJ ‘14].

References
