Beyond Clustering

What are the minimal assumptions we need to learn the parameters?

Assumption #1: $\varepsilon \leq w \leq 1-\varepsilon$

If not, we might never get a sample from one of the components, even w/ $\frac{1}{\varepsilon}$ samples

Assumption #2: $d_{TV}(F_1, F_2) \geq \varepsilon$

If not, we might only get samples from the overlap region and we can't learn the mixing wt

Theorem [Kalai, Moitra, Valiant] There is a polynomial time/sample complexity algorithm for learning* a mixture of two Gaussians under Assumption 1&2 up to ε — i.e.

* output $\hat{F} = \hat{w}_1 \hat{F}_1 + \hat{w}_2 \hat{F}_2$ s.t. $\exists$ a permutation $\pi: \{1,2\} \to \{1,2\}$ and for all $i \in \{1,2\}$ we have

$$d_{TV}(F_i, \hat{F}_{\pi(i)}) \leq \varepsilon \text{ and } |w_i - \hat{w}_{\pi(i)}| \leq \varepsilon$$

Outline

① reduce the d-dimensional problem to a series of 1-dimensional problems

② Give provable guarantees for Pearson's sixth moment test

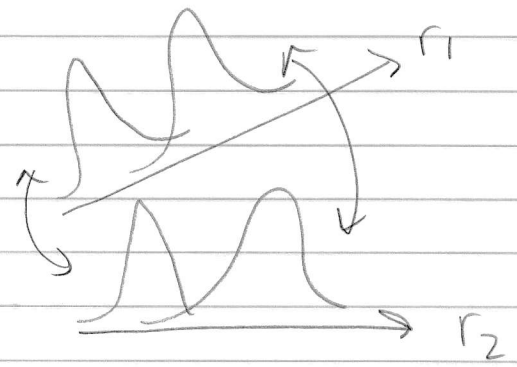
For ① the key idea is

"learning the parameters of the projection gives linear constraints on the high-dim. parameters"

Fact: $\text{Proj}_r[F(x)] = w_1 N(r^T\mu_1, r^T\Sigma_1 r) + w_2 N(r^T\mu_2, r^T\Sigma_2 r)$

i.e each sample $x \sim F$ is mapped to $r^T x$

There are $\Theta(d^2)$ parameters so we need at least $\Theta(d^2)$ projections, but consider



Main Issue: How do you pair components up along different projections?

def: we say $F(x)$ is in isotropic position if

① $\mathop{\mathbb{E}}_{x\sim F}[x] = 0$

② $\mathop{\mathbb{E}}_{x\sim F}[xx^T] = I$

Lemma: If $\mathop{\mathbb{E}}_{x\sim F}[xx^T]$ is full rank, there is an affine transformation that puts F in isotropic position

Proof: Let $\mu = \mathop{\mathbb{E}}_{x\sim F}[x]$ and consider

$$\mathop{\mathbb{E}}_{x\sim F}[(x-\mu)(x-\mu)^T] = M$$

Then by full rankness $M \succ 0$ and so by the Cholesky decomposition

$$M = BB^T$$

for invertible B. then set

$$y = B^{-1}(x-\mu) \qquad \blacksquare$$

Now let's define a parameter distance in 1-d

$$d_p(\mathcal{N}(\mu_1,\sigma_1^2), \mathcal{N}(\mu_2,\sigma_2^2)) = |\mu_1-\mu_2| + |\sigma_1^2-\sigma_2^2|$$

Isotropic Projection lemma: If F is in isotropic position and satisfies Assumption 1 & 2 then whp for random r we have

$$d_p(\text{Proj}_r[F_1], \text{Proj}_r[F_2]) \geq \text{poly}\left(\frac{1}{d}, \varepsilon\right) \triangleq \varepsilon_3.$$

Now if the directions are ε_2-close, pairing is easy because the projected means/variances change by much less than ε_3.

Main Issue: How do the errors in our estimates of the parameters of 1-d projections propagate to the high-dim estimates?

Condition Number lemma: The condition number of the linear system

$$\left\{ \begin{bmatrix} r_i^T & 0 \\ 0 & \text{vec}(r_i r_i^T) \end{bmatrix} \begin{bmatrix} \mu \\ \text{vec}(\Sigma) \end{bmatrix} = \begin{bmatrix} \hat{\mu}_{r_i} \\ \hat{\sigma}^2_{r_i} \end{bmatrix} \right\}_i$$

↑ unknowns

Is at most $\text{poly}\left(n, \frac{1}{\varepsilon_2}\right)$

Now we can estimate the parameters in 1-d up to accuracy $\varepsilon_1 \ll \varepsilon_2$ to get ε-error in high-dim