How do I prove this prop?

<u>Contradiction</u> If not.

Sequence of bad pts $X_1, X_2, X_3, \ldots, X_t$

$t > 8k/k'$ s.t. each X_i is far from span $X_1 \ldots X_{i-1}$

---

for each X_i I have $P_{i1}, P_{i2}, \ldots, P_{ik'}$ ← small singular vectors of A_{x_i}

$$P_{ij} \cdot (A_{x_i} q) \le \eta |q|_2$$

$|P_{ij}| = 1$

W.

$B_{i,j} = q_i(x) \cdot p_{ij}(y) \qquad q_i(x) = x \cdot x_i$

Gram matrix determinant of the B_{ij}'s

$[\langle B_{ij}, B_{i'j'} \rangle]$

reasonably big $\quad$ list B_{ij}'s in order $\quad B_{11}\ B_{12} \cdots B_{1k'}\ B_{21} \cdots$

then each B_{ij} has a reasonably large component orthogonal to previous ones.

$p_{ij} \perp p_{i,j'}$

B_{ij} vs $B_{i',j}$

X_i had a component of size $\geq \gamma$ orthogonal to $X_1, \dots, X_{i-1}$

$$\langle pq, p'q' \rangle \approx \langle p, p' \rangle \langle q, q' \rangle$$

B_{ij} has a comp of size $\geq \gamma$ orthogonal to any $q_{j'}(x) p(y)$ for $j' < i$

B_{ij} has a γ-sized comp. orthogonal to previous B_{ij}'s.

$\Rightarrow$ if we apply G-S. to B_{ij}'s we get comps of size $\geq \gamma$.

$\Rightarrow$ det (Gram Matrix) $\geq \gamma^{\cdot \cancel{\# B_{ij}} \cancel{(2k/k!) \cdot k!} \; 2k}$.

Upper Bound

Claim ~~B_T~~ If $P \in W$ then

$$\langle P^{(x,y)}, B_{ij} \rangle \text{ small.}$$

$$B_{ij} = q_i(x) P_{ij}(y)$$

Pick on basis for x.

$$P(x,y) = \overbrace{(x \cdot x_i)}^{q_i} P_i(y) + \underbrace{(x^{\text{orth}}) \ldots}_{\text{orthogonal}}$$

$$\langle P(x,y), B_{ij} \rangle \approx \langle P(x_i, y), \underbrace{P_{ij}(y)}_{\text{small}} \rangle$$

$$\langle P(x_i, y), P_{ij}(y) \rangle = A_{x_i} P$$

small ~~coverts~~ singular of A_{x_i}

$$\leq \eta \cdot |P|$$

$M = \begin{bmatrix} | \\ B_{ij} \\ | \end{bmatrix}$ We note M is $2k$-dimension.

W has codim $\le k$.

$\Rightarrow \exists$ k-dim'l subspace a. s.t.

$Ma \in W$.

Not hard to show that $M^T M$ has no super large e.v.

For such a's. $M^T M a = M^T (Ma)$

$\det(M^T M)$

$= \begin{bmatrix} \langle B_{ij}, Ma \rangle \end{bmatrix}$

$\le (2k\eta)^k$.

$|M^T M a| \le 2k\eta |a|$.

$\le \eta \cdot |Ma|$

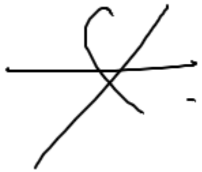
$\wedge$

$2k \cdot |a|$.

$\Rightarrow$ $M^T M$ has at least k eigenvalues $\le 2k\eta$

$$(\gamma)^{2k} < \det(M^T M) < (2k\eta)^k.$$

as long as η is small enough.



Note: This can be made computationally efficient.

---

learning mixtures of Spherical Gaussians in TV distance.

(0) reduce to k dimensions.

(1) rough clustering — reduces to case when $|M_i| = \mathrm{Poly}(k)$

(2) Compute 1'st ~~2d~~ 2d parameter moments to error $\left(\frac{\varepsilon}{k}\right)^{O(d)}$ (moments of M_i)

(4) define $Q(p) \simeq \sum w_i \, p^2(\mu_i)$ ← deg d polys.

Let U be space of small singular vectors of Q.

for any i w/ w_i not too small and any $p \in U$

$|p(\mu_i)|$ small.

(5) Use Thm compute a cover S.

every i with $w_i \geq \varepsilon/k$ has μ_i within ε of

some elt of S.

$|S| = (\varepsilon/k)^{O(d^2 k^{1/d})}$.

$$X = \sum \cancel{\mu} \; w_i N(\mu_i, I).$$

$$\approx \sum w_i N(\mu_i', I)$$

$$\uparrow$$

$$S.$$

explicit set of distributions $X_1, \ldots, X_{\underline{n}}$.

X is ϵ-close to a ~~mix~~ mixture of the X_is

<u>Claim</u> learn X given samples from X and X_i's.

Point is compute MLE. $\to$ ~~EP~~. Convex program.

$X + \varepsilon\left(\frac{x_1 + \cdots + x_n}{n}\right)$

Runtime $N = (k/\varepsilon)^{O(d^2 k^{1/d})}$

preprocessing required $(k/\varepsilon)^{O(d)}$ samples.

$d = \lg(k)$ runtime $(k/\varepsilon)^{O(\lg^2(k))}$

TV dimension linear no separation.

Existing algorithms If $|\mu_i - \mu_j| \gg \sqrt{\lg(k)}$

then in quasipoly time/samples learn parameters to error ε.

If $|\mu_i - \mu_j| \gg \sqrt{\lg(k)}$

and if you have a $c$ow for the setof means (of size N).

Then in $\mathrm{poly}(N/\epsilon)$ time & $\mathrm{poly}(N/\epsilon)$ samples,

you can learn the means.

$(k/\epsilon)^{O(\lg k)}$ samples $(k/\epsilon)^{O(\lg^2 k)}$ time.

~~$k = \lg \epsilon$~~

$d = \lg k$.

We have a trade off.

$|\mu_i - \mu_j| \gg k^{1/d}$ ← $\underline{\text{Gaussian}}$ $(k/\epsilon)^{O(d)}$ samples $(k/\epsilon)^{O(d^2 k^{1/d})}$ time.

1. Preprocessing R dim $|\mathcal{M}|$'s not too big.
2. $\boxed{\text{Computed parameter moments}}$
3. Compute a Cover
4. Use brute force.

Mixtures of linear regressions

$X \sim N(0, I)$; $Y = \beta \cdot X + \underline{\text{noise}}$.

$\beta_1 \ldots \beta_k$ weights $w_1, \ldots, w_k$.

w/ prob w_i $Y = \beta_i \cdot X + \text{noise}$.

(approximate) ~~compute~~ parameter moments of β_i's $\sum w_i \beta_i^{\otimes d}$.

Non-negative linear combinations of ReLUs

$$\mathrm{ReLU}(x) = \max(0, x)$$

$$F(x) = \sum a_i \,\mathrm{ReLU}(x \cdot \underline{w_i}) \qquad (a_i \geq 0)$$

$(x, F(x))$ + noise

approximate F.

$$\sum a_i \underline{v_i}^{\otimes d}$$

$$\hookrightarrow \underline{\sum a_i \, p^2(v_i)}$$

$x_i^2 = 1$

$$\mathbb{R}^n = \underset{x}{\mathbb{R}^{n'}} \times \underset{y}{\mathbb{R}^{n-n'}}$$

$$(x, y)$$

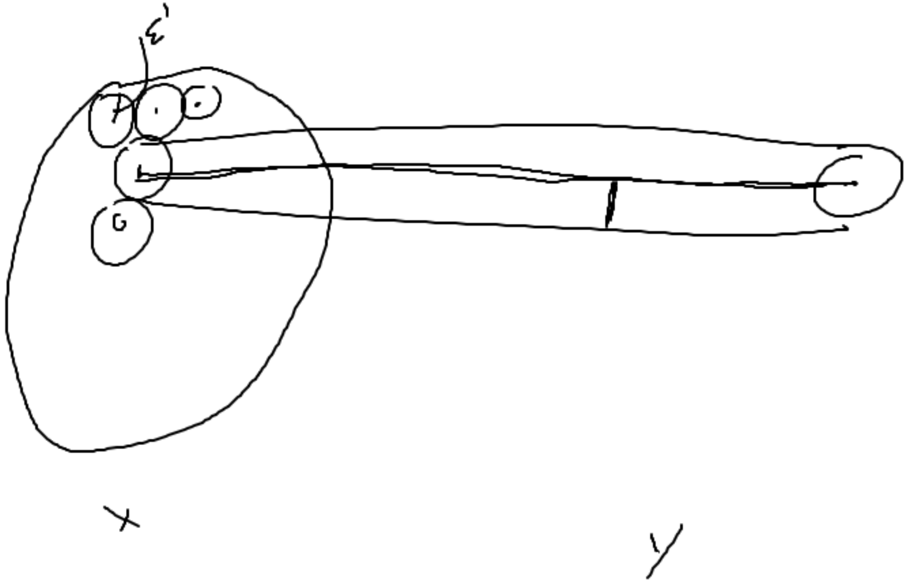
n' as smallish

$W = \{$ polys in U that are degree 1 in x & degree $d-1$ in $y\}$

$\operatorname{codim}(W) \leq k$.

if U is codim k in V.

then $U \cap W$ is codim $\leq k$ in $V \cap W$.



ε' very small

small enough

If I change δ by a bit...

It is enough to cover

this center line. $\{x_0\}\times\mathbb{R}^{n-n'}$

$\{x_i\}$

$\{x_i\}\times B(0,R)$

$\cap$

$\mathbb{R}^{n-n'}$

$A_{x_i} : W \longrightarrow$ { non deg $d-1$ pols in y }

$\cup$

deg 1, deg $d-1$ pols in U.

$P(x,y) \longrightarrow P(x_i, y)$

S_{x_i} = { non vanishing pols w/ $x = x_i$ }

η not too small

$n' \times (\cdots)$.

If $q = A_{x_i} P$. $|q| > \eta |p|$.

(x_i, y)

then S_{x_i} must nearly vanish on q.

nearly vanishing

U_{x_i} = span of the singular ~~vector~~ vectors of A_{x_i} w/ singular value $\geq \eta$.

$\delta |p| = (\delta/\eta) |q|$

$S_{x_i} \subset$ { pts that nearly vanish on U_{x_i} }

recursive version of original problem.

Say x_i is good if $\mathrm{Codim}(U_{x_i}) \le k'$

If x_i is good then the cylinder can be covered by a set of size.

$$f(d-1, k', n-n', R, \tilde{\varepsilon}, \tilde{\delta}/\eta)$$

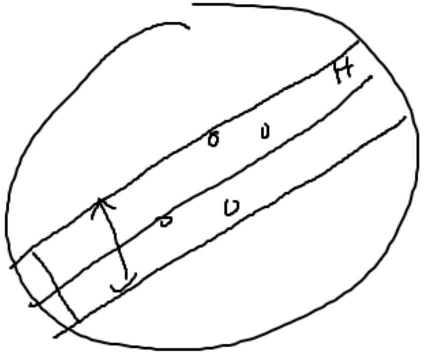
What about the not good pts

$$\delta \; \cancel{\ll} \; \varepsilon^{d} (\ldots) \qquad \eta \approx \mathrm{poly}(\varepsilon)$$

Then I can cover all of the cylinders over all of the good points w/ an appropriate # of covers.

$$f(\ldots) \le (\ldots)^{d^2 k^{1/d}} \qquad k' \approx k^{\frac{d-1}{d}}$$

Prop There exists a subspace $H < \mathbb{R}^{n'}$ of dimension $\leq 2k/k'$ s.t. all of the bad pts are close to H.



Cover $\underbrace{H \times \mathbb{R}^{n-n'}}$ as long as $n' \gg k/k' \gtrsim k^{1/d}$.

subspace of dim $\mathbb{R}^{n-n'+2k/k'}$ Then I should be fine.

$+ f(d, k, R, \tilde{\varepsilon}, \delta, \underline{n-n'+2k/k'})$