Learning Mixtures of Spherical Gaussians.

$$X = \sum_{i=1}^{k} w_i N(\mu_i, I) \qquad \subset \mathbb{R}^n.$$

Given samples from X and you want to learn X.

(either: learn $\hat{X}$ s.t. $d_{TV}(X, \hat{X}) < \varepsilon$,

learn $\hat{\mu}_i$ s.t. (up to a perm.) $|\mu_i - \hat{\mu}_i| < \varepsilon$)

Note: Learning Non-spherical Gaussians has $n^{\Omega(k)}$ lower bounds in the SQ model

Natural approach is method of moments.

[as long as $|\mu_i|$ not too big].

Easy approximate $\|$ st d moments of X for any constant d.

Question What are these moments?

$$X = D * G$$

D → discrete distr. over means.

G ← $N(0, I)$

$$\mathbb{E}[X^{\otimes t}] = \sum_i (\text{Sym}(\mathbb{E}D^{\otimes i} \otimes \underline{G^{\otimes t-i}}))$$

Not hard to deconvolve.

approximate $\mathbb{E}[D^{\otimes i}]$

Q Can we learn D from its low order moments?

A No. Consider 1-d. problem.

first d moments. $d+1$ parameters

law X has $2k$ parameters.

$\Rightarrow$ Unless $d \geq 2k$. moments don't determine D.

D, D' match moments, (even up to small distance)

$N \cdot D$, $N \oplus D'$ also match moments.

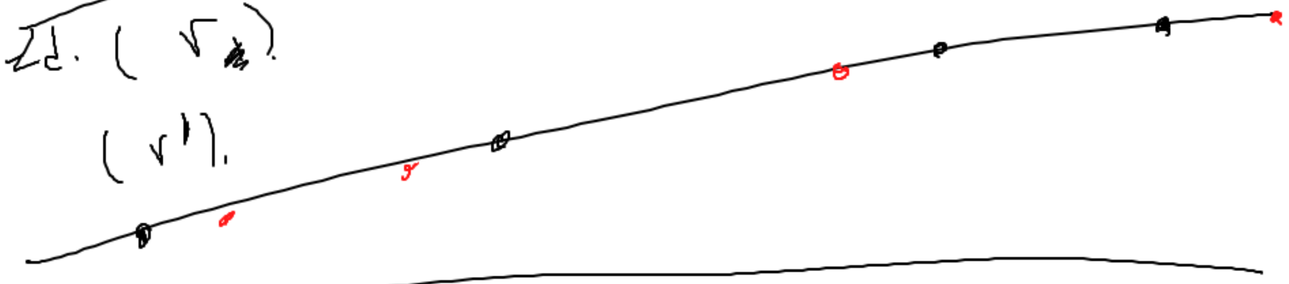
embedd into n dimensions

$x^d - \sigma_1 x^{d-1} + \dots \sigma_d \rightarrow \Sigma r_i^t$ for $t<d$

coef s of p = sym polys in roots

symply of deg 2d. (r_i)

" sym — — (r').

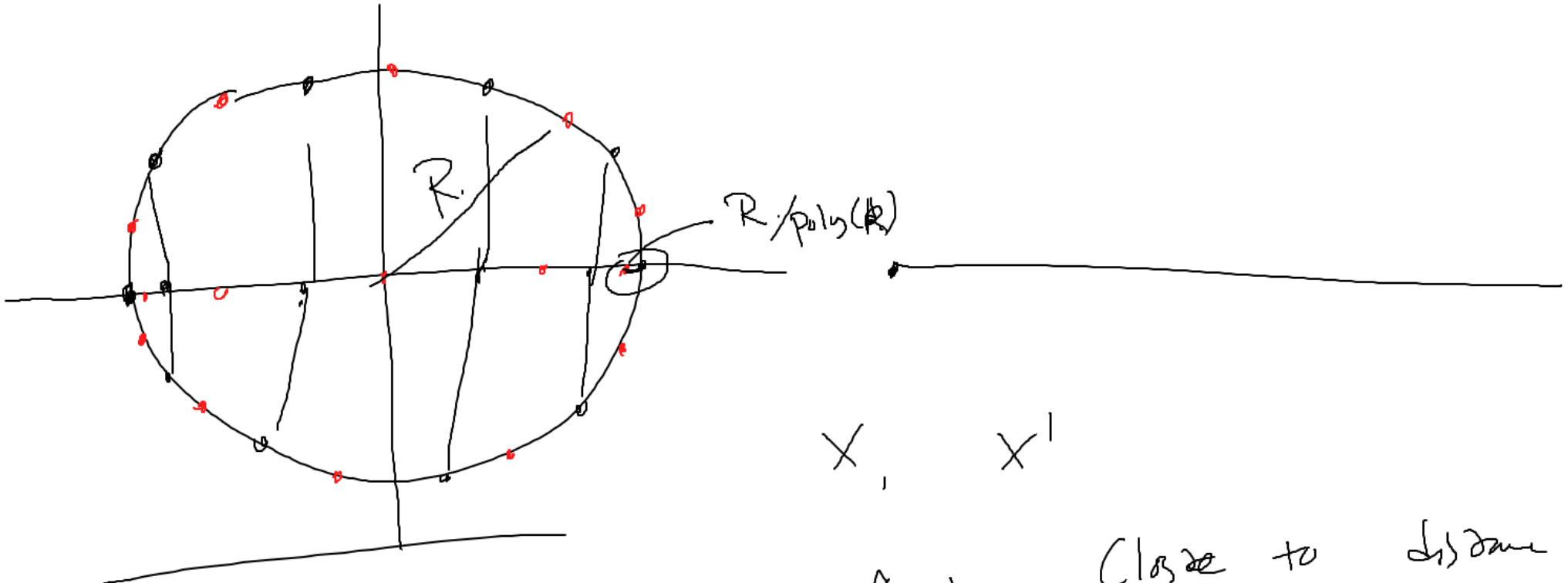


coef s of p' = — — —

Pick p, p' two degree d polys that differ by a constant. (w/ d real roots).

& let means of D be roots of p

means of D' be roots of p'

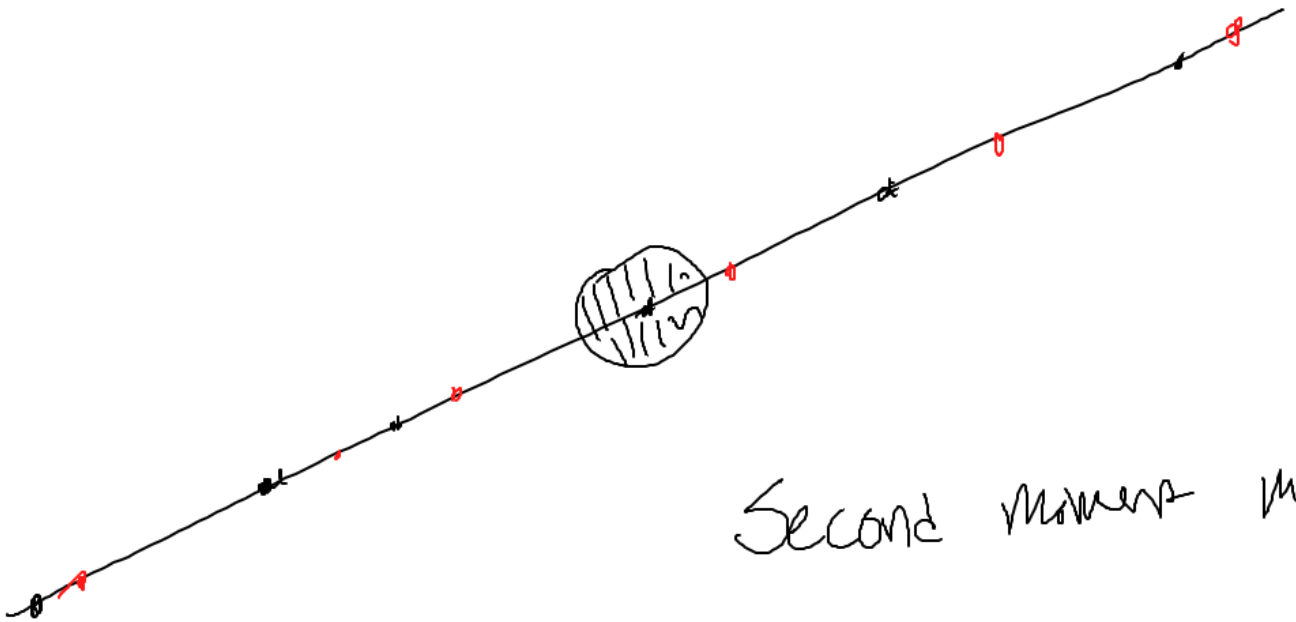


X, X'

Any 1-d X approximate X by empirical distribution $*N(0,\varepsilon)$

~~Arb.~~ Close to distance 1 in dTV.

but first 4 moments match.



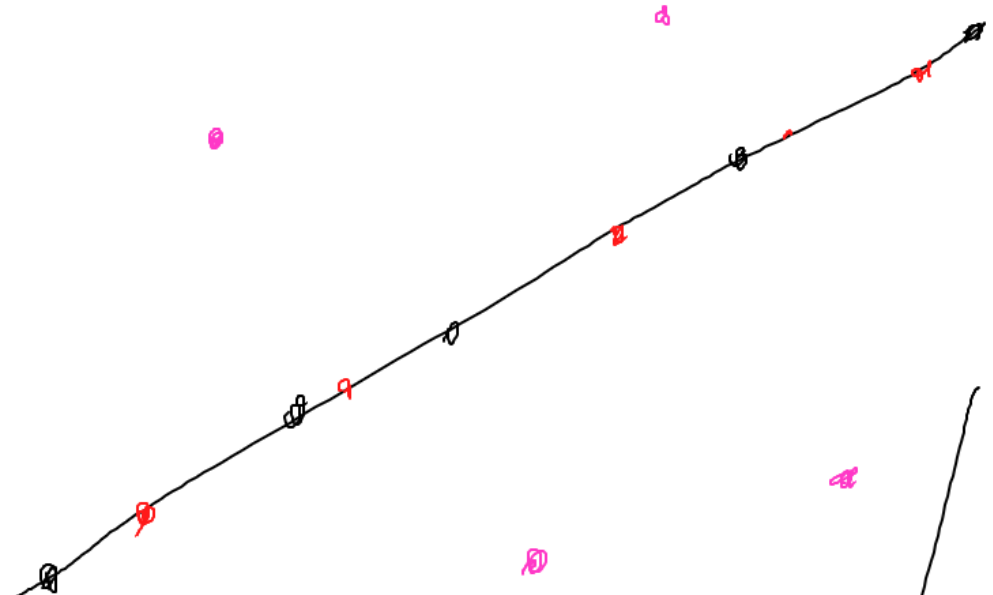
Second moment matrix of $f(X)$

$= I +$ Second moment of (D)

Supported on a ~~subspace~~ k subspace of dim.

~~Span~~ rank $\leq k$

Span $\subset$ span of M_i's.



reduce to
k-dim
problem.

---

to prove lower bounds
(say in SQ)

WMT is measures that
are spherically symmetric.

Impossible

$X = D * G.$

low deg moments of X $\longrightarrow$ low deg moments of D

$\longrightarrow$ $\sum_{i=1}^{R} w_i q(\mu_i)$ for any low degree poly. q

$q(x) = p^2(x)$ for some p

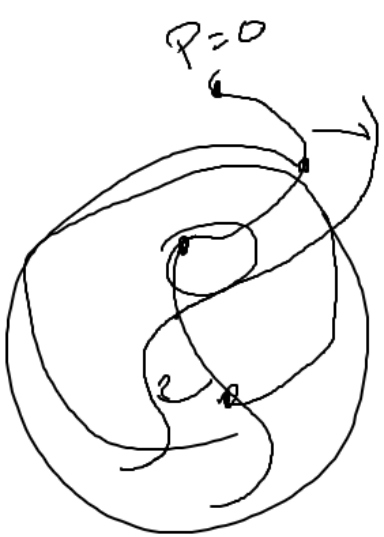
$= \sum w_i p^2(\mu_i)$

$= 0$ iff $\boxed{p(\mu_i) = 0}$ for all i.

Are the low degree picks p s.t.

$P(\mu_i) = 0$ for all i?

Yes



Linear transformation

$\{$ deg $\leq d$ poly $\}$ $\xrightarrow{\mathcal{T}}$ $\mathbb{R}^{\binom{n+d}{d}}$ $\to$ $(P(\mu_1), P(\mu_2), \ldots, P(\mu_k)) \in \mathbb{R}^k$

Kernel of $\mathcal{T}$.

Kernel has codim $\leq k$.

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

$\binom{n'+d}{d} > k$ if $n' > d\,k^{1/d}$.

Compute monomials of x → Compute D_i → Compute space of low degree vanishing polynomials P_i (exactly)

V = Variety defined by these polynomials

= Set of pts that these polys vanish on.

$M_i \in V$. intuition V small

Proof: $\dim(\text{deg-}d \text{ polys on } V) \leq$ Codim of space of defining polys.

$\leq R$.

$\binom{\dim(V)+d}{d} \leq \dim(\text{deg-}d \text{ polys on } V)$

$\Rightarrow \dim(V) \leq d R^{1/d}$.

R

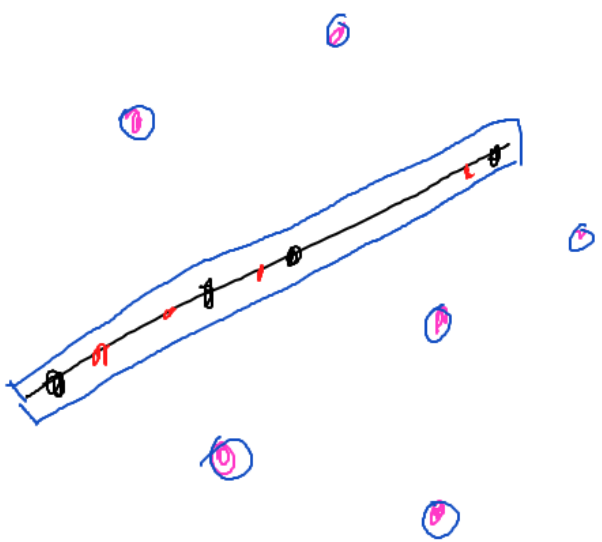
Idea "project" onto V.

do something exhaustive $\exp(\dim(U))$

$d = \lg(R)$

Computing d inner products

n^d time.



reduce to $d k^{1/d}$ $\times$ d. dimensional vectors

exhaustive then takes $\approx 2^d$ time.

quasi-poly alg

Technical problems

1. V is low dimension but is it "simple"?
2. How do we project onto V?
3. How do we compute on V?
4. How do I deal with approx. defining polynomials?

Idea replace V w/ a point cloud.

~~Set~~ Vector span $\mathcal{U}$ of degree d polys $\mathbb{R}$

$\mathrm{Codim}(\mathcal{U}) \leq k$.

we'll get back to this

$$S = \{ x \in \mathbb{R}^n : |x| < R, \quad |p(x)| < \delta |p| \text{ for all } p \in \mathcal{U} \}$$

Want for $\varepsilon > 0$ to find a small ε-cover of S.

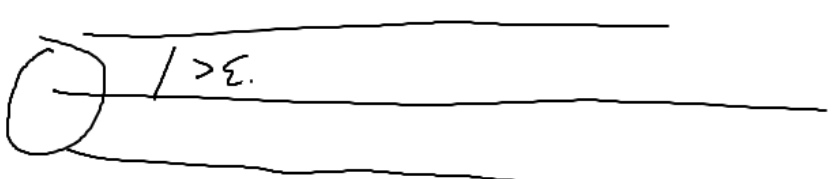
~~Ex~~ **Thrm** If $\varepsilon > \delta^{\frac{1}{2d}} \mathrm{poly}(R\, n\, d\, k)$ then there exists an ε-cover of size at most

$$(2(R/\varepsilon) \cdot dkn)^{O(d^2 k^{1/d})}$$

Expect cover size should be $\geq$

$O(R/\varepsilon)^{\dim(V)}$ $\hookleftarrow$ $dk^{1/d}$ we get $d^2k^{1/d}$

$$\varepsilon > \delta^{1/2d}(\cdots).$$



Technicalities

(1) U C nom. degree-d polynomials.

(2) $\delta < (\varepsilon/R d)^{cd}$ $\qquad P(x) = A \cdot x^{\otimes d}$ $\qquad |P| := |A|_2$.

$|P|$ is spherically symmetric.

Define $f(\varepsilon, R, d, k, n, \delta)$ = ~~Adaptive~~ worst case comm size.

Prove a recursive upper bound on f.

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

(x, y)

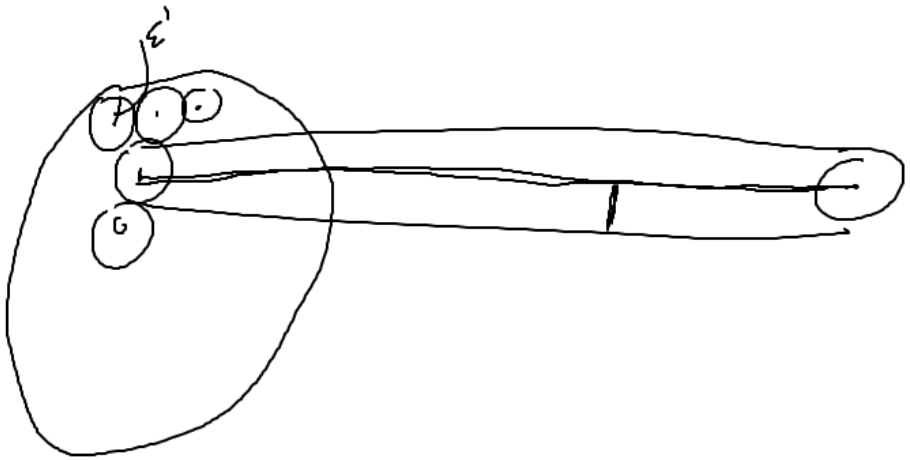
n' as small-ish

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

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

if U is codim $\geq k$ in V.

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



ε' very small

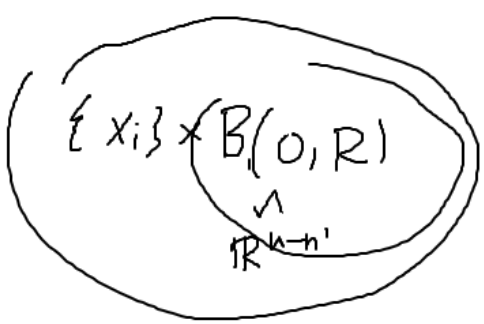
small enough

If I change δ by a bit...

It is enough to cover

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

$\{x_i\}$



$A_{x_i} : W \longrightarrow \{$ nbhd deg $d-1$ polys in $y\}$

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

$\cup$

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

$S_{x_i} = \{$ nbhd vanishing pts w/ $x = x_i\}$

η not too small

$n' \times (\cdot -)$.

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

(x_i, y) nearly vanishing

Then S_{x_i} must nearly vanish on q.

U_{x_i} = span of the singular 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 $\operatorname{Codim}(U_{X_i}) \leq k'$

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

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

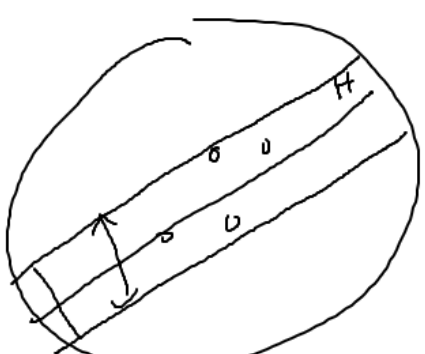
What about the not good pts

$$\delta \leq \varepsilon^{d.}(\ldots) \qquad \eta \approx \text{poly}(\varepsilon)$$

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

$$f(\ldots) \leq (\ldots)^{d^2 k^{kd.}} \qquad k' \approx k^{\frac{d-1}{d.}}$$

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



Cor $H \times \mathbb{R}^{n-n'}$ as long as $n' >> k/k' \sim k^{1/d}$.

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

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

How do I prove this prop?

Contradiction — If not.

Sequence of bad pts $x_1, x_2, x_3, \ldots, x_t$

$t > 2k/k'$ 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}

$p_{ij} \cdot (A_{x_i} q) \leq \eta \|q\|_2$ (q: w.)

$\|p_{ij}\| = 1$.

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

Gram matrix determinants of the $B_{i,j}$'s

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

---

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

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

$P_{ij} \perp P_{ij'}$

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

X_i had a component of size $\geq \gamma$ orthogonal to

$$X_1, \ldots, X_{i-1}$$

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

B_{ij} has a comp of size $\geq \gamma$ orthogonal to <u>any</u>

$$q_{i'}(x)\, p(y) \quad \text{for } i' < i$$

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) $\geq \gamma^{2k}$.

Upper Bound

Claim

If $P \in W$ then

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

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

Pick an basis for x.

$P(x,y) = (x \cdot x_i)^{q_i} P_i(y) + (x^{\perp} \cdot \ldots)$ orthogonal

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

$= A_{x_i} P$

small

Singular of A_{x_i}

$\leq \eta \cdot |P|$

$M = [B_{ij}]$ We note M is $2R$-dimensional.

W has $\text{Codim} \leq R$.

$\Rightarrow \exists$ R-dim'l subspace a. s.t. $Ma \in W$.

We bound the singular values of M^TM by the Singular Lemma.

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

$= [\langle B_{ij}, Ma \rangle]$

$|M^TMa| \leq 2kn|a|$.

$\Rightarrow M^TM$ has at least R eigenvalues $\leq 2kn$

$\leq n \cdot |Ma|$

$\leq 2k \cdot |a|$

$\det(M^TM) \leq (2kn)^R$.

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

as long as η is small enough.

$\not\subset$.

Note: This Can Can be made computationally efficient.

---

learning Mixtures of Spherical Gaussians in TV distance.

(0) reduce to k dimensions.

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

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

(4) ← deg d polys.

define $Q(p) \triangleq \sum w_i \, p^2(\mu_i)$

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.

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

$$X = \sum_{i} w_i N(\mu_i, I).$$

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

$$\Uparrow$$

$$S.$$

Explicit set of distributions $X_1, \ldots, X_N$.

X is ϵ-close to a mixture of the X_i's

Claim: learn X given samples from X and X_i's.

Point is Compute MLE. → NP. Convex program.

$X + \varepsilon\left(\frac{X_1 + \cdots + X_N}{N}\right)$

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

Previous Alg required $(k/\varepsilon)^{O(d)}$ Samples.

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

TV distance lower no separation.

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

then it efficiently learns Parameters to error ε.

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

and if you can learn a coin for the set of means (of size N).

Then in $\text{Poly}(N, n/\varepsilon)$ time & $\text{poly}(n/\varepsilon)$ samples, you can learn the means.

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

~~$k = \lg k$~~.

$d = \lg k$.

We can a trade off.

$|\mu_i - \mu_j| \gg k^{1/d}$. ← Gaussians

$(k/\varepsilon)^{O(d)}$ samples $(k/\varepsilon)^{O(d^2 k^{1/d})}$ time.

(1) Reprocessing R dm M's not too big.

(2) Computed Parameter moments

(3) Compute a Cover

(4) Use brute force.

Mixtures of linear regressions

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

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

w/ prob w_i $y = \beta_i \cdot x + \text{noise}$.

(continuous)
~~Compute~~ parameters of β_i's $\sum w_i \beta_i^{\otimes d}$.

Non-negative linear combinations of ReLUs

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

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

$(x, F(x))$ — graph

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

approximate F.

$\hookrightarrow \sum a_i p^2(v_i)$

$$\chi_i^2 = 1.$$

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

(x, y)

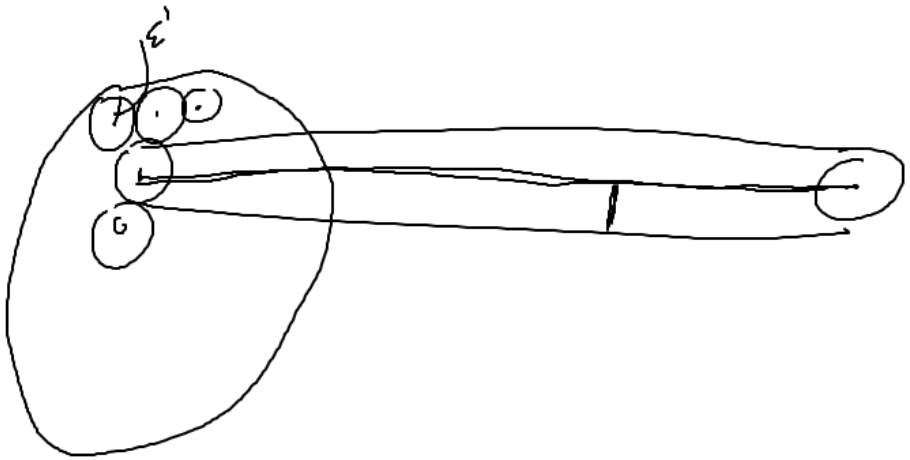
n' as small-ish

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

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

if U is codim $\geq k$ in V.

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



ε' very small

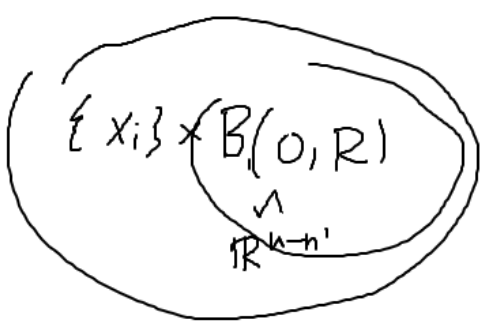
small enough

If I change δ by a bit...

It is enough to cover

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

$\{x_i\}$



$A_{x_i} : W \longrightarrow \{$ nbhd deg $d-1$ polys in $y\}$

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

$\cup$

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

$S_{x_i} = \{$ nbhd vanishing pts w/ $x = x_i\}$

η not too small

$n' \times (\cdot -)$.

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

(x_i, y) nearly vanishing

Then S_{x_i} must nearly vanish on q.

U_{x_i} = span of the singular 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 $\operatorname{Codim}(U_{X_i}) \leq k'$

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

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

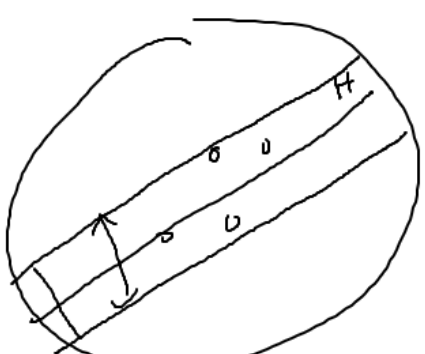
What about the not good pts

$$\delta \leq \varepsilon^{d.}(\ldots) \qquad \eta \approx \text{poly}(\varepsilon)$$

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

$$f(\ldots) \leq (\ldots)^{d^2 k^{kd.}} \qquad k' \approx k^{\frac{d-1}{d.}}$$

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



Cor $H \times \mathbb{R}^{n-n'}$ as long as $n' >> k/k' \sim k^{1/d}$.

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

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