Fourier PCA

Navin Goyal (MSR India), Santosh Vempala (Georgia Tech)
and Ying Xiao (Georgia Tech)

Introduction

1. Describe a learning problem.
2. Develop an efficient tensor decomposition.

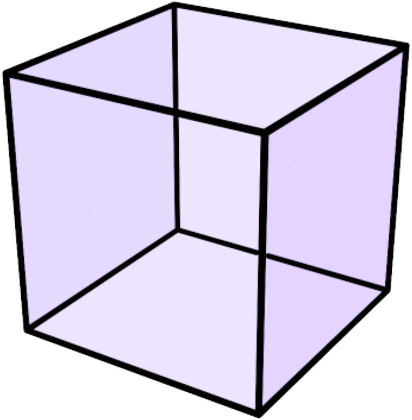
Independent component analysis

See independent samples $x = As$:

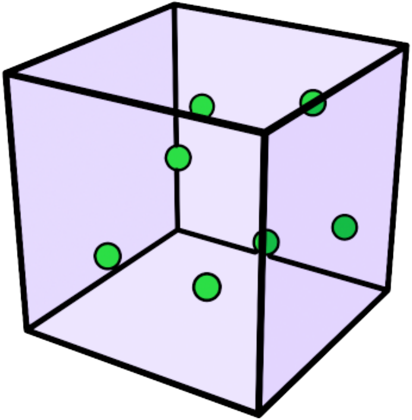
- $s \in \mathbb{R}^m$ is a random vector with independent coordinates.
- Variables s_i are not Gaussian.
- $A \in \mathbb{R}^{n \times m}$ is a fixed matrix of full row rank.
- Each column $A_i \in \mathbb{R}^m$ has unit norm.

Goal is to compute A.

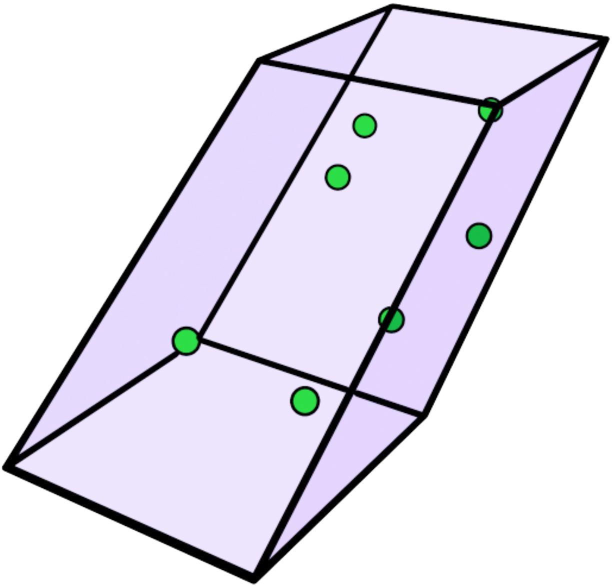
ICA: start with independent random vector s



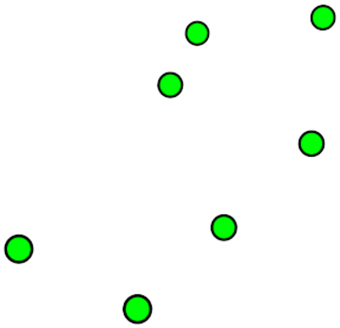
ICA: independent samples



ICA: but under the map A



ICA: goal is to recover A from samples only.



Applications

Matrix A gives n linear measurements of m random variables.

- General dimensionality reduction tool in statistics and machine learning [HTF01].
- Gained traction in deep belief networks [LKNN12].
- Blind source separation and deconvolution in signal processing [HKO01].
- More practically: finance [KO96], biology [VSJHO00] and MRI [KOHFY10].

Status

- Jutten and Herault 1991 formalised this problem. Studied in our community first by [FJK96].
- Provably good algorithms: [AGMS12] and [AGHKT12].
- Many algorithms proposed in signals processing literature [HKO01].

Standard approaches – PCA

Define a "contrast function" where optima are A_j.

- Second moment $\mathbb{E}\left((u^T x)^2\right)$ is usual PCA.
- Only succeeds when all the eigenvalues of the covariance matrix are different.
- Any distribution can be put into isotropic position.

Standard approaches – fourth moments

Define a "contrast function" where optima are A_j.

- Fourth moment $\mathbb{E}\left((u^T x)^4\right)$.
- Tensor decomposition:

$$T = \sum_{j=1}^{m} \lambda_j A_j \otimes A_j \otimes A_j \otimes A_j$$

- In case A_j are orthonormal, they are the local optima of:

$$T(v, v, v, v) = \sum_{i,j,k,l} T_{ijkl} v_i v_j v_k v_l$$

where $\|v\| = 1$.

Standard assumptions for $x = As$

All algorithms require:

1. A is full rank $n \times n$: as many measurements as underlying variables.
2. Each s_i differs from a Gaussian in the fourth moment:

$$\mathbb{E}(s_i) = 0, \qquad \mathbb{E}\left(s_i^2\right) = 1, \qquad \left|\mathbb{E}\left(s_i^4\right) - 3\right| \geq \Delta$$

Note: this precludes the underdetermined case when $A \in \mathbb{R}^{n \times m}$ is fat.

Our results

We require neither standard assumptions.

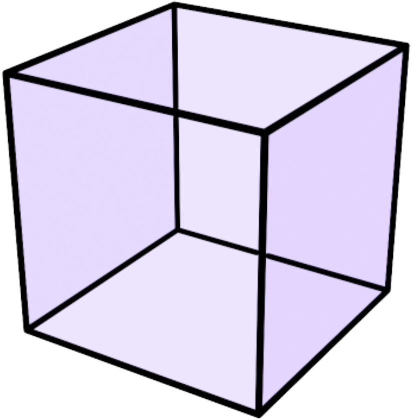
Underdetermined: $A \in \mathbb{R}^{n \times m}$ is a fat matrix where $n << m$.

Any moment: $\left|\mathbb{E}\left(s_i^r\right) - \mathbb{E}\left(z^r\right)\right| \geq \Delta$ where $z \sim N(0,1)$.

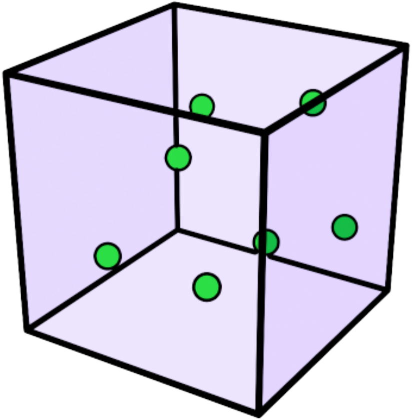
Theorem (Informal)

Let $x = As$ be an underdetermined ICA model. Let $d \in 2\mathbb{N}$ be such that $\sigma_m\left(\left[\mathrm{vec}\left(A_i^{\otimes d/2}\right)\right]_{i=1}^m\right) > 0$. Suppose for each s_i, one of its first k cumulants satisfies $\left|\mathrm{cum}_{k_i}(s_i)\right| \geq \Delta$. Then one can recover the columns of A up to ϵ accuracy in polynomial time.

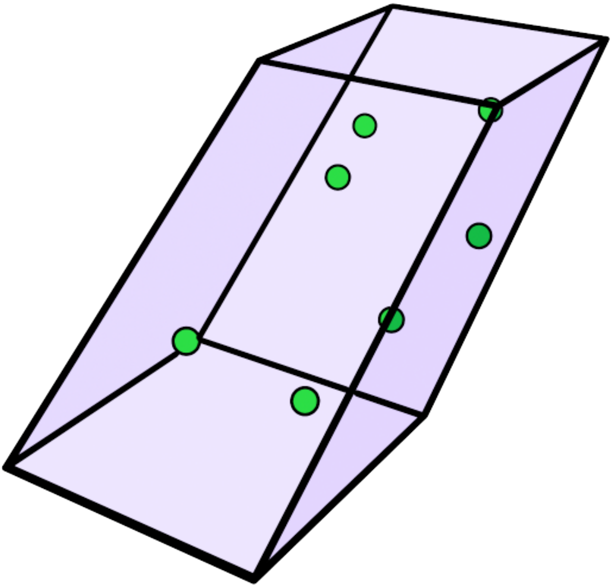
Underdetermined ICA: start with distribution over $\mathbb{R}^m$



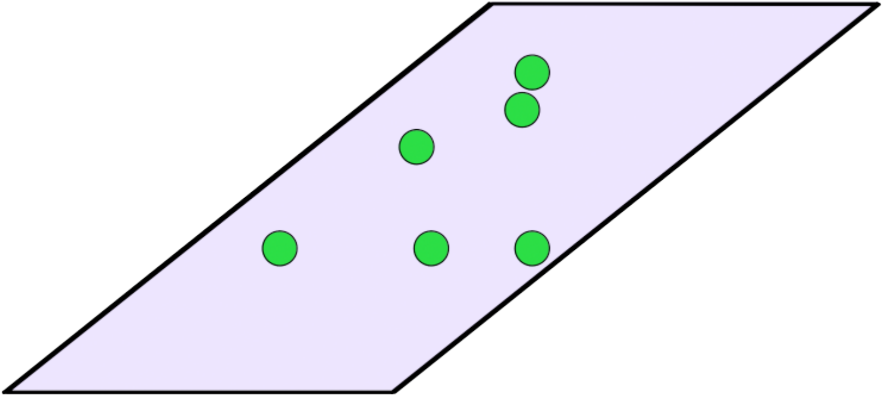
Underdetermined ICA: independent samples s



Underdetermined ICA: A first rotates/scales



Underdetermined ICA: then A projects down to $\mathbb{R}^n$



Fully determined ICA – nice algorithm

1. (Fourier weights) Pick a random vector u from $N(0, \sigma^2 I_n)$.
   For every x, compute its Fourier weight

$$w(x) = \frac{e^{iu^T x}}{\sum_{x \in S} e^{iu^T x}}.$$

2. (Reweighted Covariance) Compute the covariance matrix of the points x reweighted by $w(x)$

$$\mu_u = \frac{1}{|S|} \sum_{x \in S} w(x) x \quad \text{and} \quad \Sigma_u = \frac{1}{|S|} \sum_{x \in S} w(x)(x - \mu_u)(x - \mu_u)^T.$$

3. Compute the eigenvectors V of Σ_u.

Why does this work?

1. Fourier differentiation/multiplication relationship:

$$(f')\hat{}(u) = (2\pi i u)\hat{f}(u)$$

2. Actually we consider the log of the fourier transform:

$$D^2 \log\left(\mathbb{E}\left(\exp(iu^T x)\right)\right) = A\mathrm{diag}\left(g_j(A_j^T u)\right) A^T$$

where $g_j(t)$ is the second derivative of $\log(\mathbb{E}\left(\exp(its_j)\right))$.

Technical overview

Fundamental analytic tools:

- *Second characteristic function* $\psi(u) = \log(\mathbb{E}\left(\exp(iu^T x)\right)$.
- Estimate order d derivative tensor field $D^d\psi$ from samples.
- Evaluate $D^d\psi$ at two randomly chosen $u, v \in \mathbb{R}^n$ to give two tensors T_u and T_v.
- Perform a tensor decomposition on T_u and T_v to obtain A.

First derivative

Easy case $A = I_n$:

$$\psi(u) = \log(\mathbb{E}\left(\exp(iu^T x)\right) = \log(\mathbb{E}\left(\exp(iu^T s)\right)$$

Thus:

$$\begin{aligned}
\frac{\partial \psi}{\partial u_1} &= \frac{1}{\mathbb{E}\left(\exp(iu^T s)\right)} \mathbb{E}\left(s_1 \exp(iu^T s)\right) \\
&= \frac{1}{\prod_{j=1}^{n} \mathbb{E}\left(\exp(iu_j s_j)\right)} \mathbb{E}\left(s_1 \exp(iu_1 s_1)\right) \prod_{j=2}^{n} \mathbb{E}\left(\exp(iu_j s_j)\right) \\
&= \frac{\mathbb{E}\left(s_1 \exp(iu_1 s_1)\right)}{\mathbb{E}\left(exp(iu_1 s_1)\right)}
\end{aligned}$$

Second derivative

Easy case $A = I_n$:

1. Differentiating via quotient rule:

$$\frac{\partial^2 \psi}{\partial u_1^2} = \frac{\mathbb{E}\left(s_1^2 \exp(iu_1 s_1)\right) - \mathbb{E}\left(s_i \exp(iu_1 s_1)\right)^2}{\mathbb{E}\left(\exp(iu_1 s_1)\right)^2}$$

2. Differentiating a constant:

$$\frac{\partial^2 \psi}{\partial u_1 \partial u_2} = 0$$

General derivatives

- Key point: taking one derivative isolates each variable u_i.
- Second derivative is a diagonal matrix.
- Subsequent derivatives are diagonal tensors: only the $(i, \ldots, i)$ term is nonzero.

NB: Higher derivatives are represented by $n \times \cdots \times n$ tensors. There is one such tensor per point in $\mathbb{R}^n$.

Basis change: second derivative

When $A \neq I_n$, we have to work much harder:

$$D^2\psi_u = A\mathrm{diag}\left(g_j(A_j^T u)\right) A^T$$

where $g_j : \mathbb{R} \to \mathbb{C}$ is given by:

$$g_j(v) = \frac{\partial^2}{\partial v^2} \log\left(\mathbb{E}\left(\exp(ivs_j)\right)\right)$$

Basis change: general derivative

When $A \neq I_n$, we have to work much harder:

$$D^d \psi_u = \sum_{j=1}^{m} g(A_j^T u)(A_j \otimes \cdots \otimes A_j)$$

where $g_j : \mathbb{R} \to \mathbb{C}$ is given by:

$$g_j(v) = \frac{\partial^d}{\partial v^d} \log\left(\mathbb{E}\left(\exp(ivs_j)\right)\right)$$

Evaluating the derivative at different points u give us tensors with shared decompositions!

Single tensor decomposition is NP hard

Forget the derivatives now. Take $\lambda_j \in \mathbb{C}$ and $A_j \in \mathbb{R}^n$:

$$T = \sum_{j=1}^{m} \lambda_j A_j \otimes \cdots \otimes A_j,$$

When we can recover the vectors A_j? When is this computationally tractable?

Known results

- When $d = 2$, usual eigenvalue decomposition.

$$M = \sum_{j=1}^{n} \lambda_j A_j \otimes A_j$$

- When $d \geq 3$ and A_j are linearly independent, a tensor power iteration suffices [AGHKT12].

$$T = \sum_{j=1}^{m} \lambda_j A_j \otimes \cdots \otimes A_j,$$

- This necessarily implies $m \leq n$.

For unique recovery, require all the eigenvalues to be different.

Generalising the problem

What about two equations instead of one?

$$T_\mu = \sum_{j=1}^{m} \mu_j A_j \otimes \cdots \otimes A_j \qquad T_\lambda = \sum_{j=1}^{m} \lambda_j A_j \otimes \cdots \otimes A_j$$

Our technique will flatten the tensors:

$$M_\mu = \left[\operatorname{vec}\left(A_j^{\otimes d/2}\right)\right] \operatorname{diag}\left(\mu_j\right) \left[\operatorname{vec}\left(A_j^{\otimes d/2}\right)\right]^T$$

Algorithm

Input: two tensors T_μ and T_λ flattened to M_μ and M_λ:

1. Compute W the right singular vectors of M_μ.
2. Form matrix $M = (W^T M_\mu W)(W^T M_\lambda W)^{-1}$.
3. Eigenvector decomposition $M = PDP^{-1}$.
4. For each column P_i, let $v_i \in \mathbb{C}^n$ be the best rank 1 approximation to P_i packed back into a tensor.
5. For each v_i, output $\mathrm{re}\left(e^{i\theta^*} v_i\right) / \left\|\mathrm{re}\left(e^{i\theta^*} v_i\right)\right\|$ where $\theta^* = \mathrm{argmax}_{\theta \in [0,2\pi]}(\left\|\mathrm{re}\left(e^{i\theta} v_i\right)\right\|)$.

Theorem

Theorem (Tensor decomposition)

Let $T_\mu, T_\lambda \in \mathbb{R}^{n\times\cdots\times n}$ be order d tensors such that $d \in 2\mathbb{N}$ and:

$$T_\mu = \sum_{j=1}^{m} \mu_j A_j^{\otimes d} \qquad T_\lambda = \sum_{j=1}^{m} \lambda_j A_j^{\otimes d}$$

where $\mathrm{vec}\left(A_j^{\otimes d/2}\right)$ are linearly independent, $\mu_i/\lambda_i \neq 0$ and $\left|\frac{\mu_i}{\lambda_i} - \frac{\mu_j}{\lambda_j}\right| > 0$ for all i, j. Then, the vectors A_j can be estimated to any desired accuracy in polynomial time.

Analysis

Let's pretend M_μ and M_λ are full rank:

$$\begin{aligned} M_\mu M_\lambda^{-1} &= \left[\mathrm{vec}\left(A_j^{\otimes d/2}\right)\right]\mathrm{diag}\left(\mu_j\right)\left[\mathrm{vec}\left(A_j^{\otimes d/2}\right)\right]^T \\ &\quad \times \left(\left[\mathrm{vec}\left(A_j^{\otimes d/2}\right)\right]^T\right)^{-1}\mathrm{diag}\left(\lambda_j\right)^{-1}\left[\mathrm{vec}\left(A_j^{\otimes d/2}\right)\right]^{-1} \\ &= \left[\mathrm{vec}\left(A_j^{\otimes d/2}\right)\right]\mathrm{diag}\left(\mu_j/\lambda_j\right)\left[\mathrm{vec}\left(A_j^{\otimes d/2}\right)\right]^{-1} \end{aligned}$$

The eigenvectors are flattened tensors of the form $A_j^{\otimes d/2}$.

Diagonalisability for non-normal matrices

When can we write $A = PDP^{-1}$?

- Require all eigenvectors to be independent (P invertible).
- Minimal polynomial of A has non-degenerate roots.

Diagonalisability for non-normal matrices

When can we write $A = PDP^{-1}$?

- Require all eigenvectors to be independent (P invertible).
- Minimal polynomial of A has non-degenerate roots.
- Sufficient condition: all roots are non-degenerate.

Perturbed spectra for non-normal matrices

More complicated than normal matrices:

Normal: $|\lambda_i(A+E) - \lambda_i(A)| \leq \|E\|$.

not-Normal: Either Bauer-Fike Theorem $|\lambda_i(A+E) - \lambda_j(A)| \leq \|E\|$ for some j, or we must assume $A+E$ is already diagonalizable.

Neither of these suffice.

Generalised Weyl inequality

Lemma

Let $A \in \mathbb{C}^{n\times n}$ be a diagonalizable matrix such that $A = P\mathrm{diag}\,(\lambda_i)\,P^{-1}$. Let $E \in \mathbb{C}^{n\times n}$ be a matrix such that $|\lambda_i(A) - \lambda_j(A)| \geq 3\kappa(P)\,\|E\|$ for all $i \neq j$. Then there exists a permutation $\pi : [n] \to [n]$ such that

$$\left|\lambda_i(A+E) - \lambda_{\pi(i)}(A)\right| \leq \kappa(P)\,\|E\|\,.$$

Proof.

Via a homotopy argument (like strong Gershgorin theorem). □

Robust analysis

Proof sketch:

1. Apply Generalized Weyl to bound eigenvalues hence diagonalisable.
2. Apply Ipsen-Eisenstat theorem (generalised Davis-Kahan $\sin(\theta)$ theorem).
3. This implies that output eigenvectors are close to $\mathrm{vec}\left(A_j^{\otimes d/2}\right)$
4. Apply tensor power iteration to extract approximate A_j.
5. Show that the best real projection of approximate A_j is close to true.

Underdetermined ICA Algorithm

$x = As$ where A is a fat matrix.

1. Pick two independent random vectors $u, v \sim N(0, \sigma^2 I_n)$.
2. Form the d^{th} derivative tensors at u and v, T_u and T_v.
3. Run tensor decomposition on the pair (T_u, T_v).

Estimating from samples

$$
\begin{aligned}
&[D^4\psi_u]_{i_1,i_2,i_3,i_4} \\
&= \frac{1}{\phi(u)^4}\Big[\mathbb{E}\left((ix_{i_1})(ix_{i_2})(ix_{i_3})(ix_{i_4})\exp(iu^T x)\right)\phi(u)^3 \\
&\quad - \mathbb{E}\left((ix_{i_2})(ix_{i_3})(ix_{i_4})\exp(iu^T x)\right)\mathbb{E}\left((ix_{i_1})\exp(iu^T x)\right)\phi(u)^2 \\
&\quad - \mathbb{E}\left((ix_{i_2})(ix_{i_3})\exp(iu^T x)\right)\mathbb{E}\left((ix_{i_1})(ix_{i_4})\exp(iu^T x)\right)\phi(u)^2 \\
&\quad - \mathbb{E}\left((ix_{i_2})(ix_{i_4})\exp(iu^T x)\right)\mathbb{E}\left((ix_{i_1})(ix_{i_3})\exp(iu^T x)\right)\phi(u)^2 + \cdots
\end{aligned}
$$

At most $2^{d-1}(d-1)!$ terms. Each one is easy to estimate empirically!

Theorem

Theorem

Fix $n, m \in \mathbb{N}$ such that $n \leq m$. Let $x \in \mathbb{R}^n$ be given by an underdetermined ICA model $x = As$. Let $d \in \mathbb{N}$ such that and $\sigma_m\left(\left[\mathrm{vec}\left(A_i^{\otimes d/2}\right)\right]_{i=1}^{m}\right) > 0$. Suppose that for each s_i, one of its cumulants $d < k_i \leq k$ satisfies $|\mathrm{cum}_{k_i}(s_i)| \geq \Delta$ and $\mathbb{E}\left(|s_i|^k\right) \leq M$. Then one can recover the columns of A up to ϵ accuracy in time and sample complexity

$$poly\left(n^{d+k}, m^{k^2}, M^k, 1/\Delta^k, 1/\sigma_m\left(\left[\mathrm{vec}\left(A_i^{\otimes d/2}\right)\right]\right)^k, 1/\epsilon\right).$$

Analysis

Recall our matrices were:

$$M_u = \left[\mathrm{vec}\left(A_i^{\otimes d/2}\right)\right]\mathrm{diag}\left(g_j(A_j^T u)\right)\left[\mathrm{vec}\left(A_i^{\otimes d/2}\right)\right]^T$$

where:

$$g_j(v) = \frac{\partial^d}{\partial v^d}\log\left(\mathbb{E}\left(\exp(ivs_j)\right)\right)$$

Need to show that $g_j(A_j^T u)/g_j(A_j^T v)$ are well-spaced.

Truncation

Taylor series of second characteristic:

$$g_i(u) = -\sum_{l=d}^{k_i} \mathrm{cum}_l(s_i) \frac{(iu)^{l-d}}{(l-d)!} + R_t \frac{(iu)^{k_i-d+1}}{(k_i-d+1)!}.$$

- Finite degree polynomials are anti-concentrated.
- Tail error is small because of existence of higher moments (in fact one suffices).

Tail error

- g_j is the d^{th} derivative of $\log(\mathbb{E}\left(\exp(iu^T s)\right))$.
- For characteristic function $\left|\phi^{(d)}(u)\right| \leq \mathbb{E}\left(|x|^d\right)$.
- Count the number of terms after iterating quotient rule d times.

Polynomial anti-concentration

Lemma

Let $p(x)$ be a degree d monic polynomial over $\mathbb{R}$. Let $x \sim N(0, \sigma^2)$, then for any $t \in \mathbb{R}$ we have

$$\Pr\left(|p(x) - t| \leq \epsilon\right) \leq \frac{4d\epsilon^{1/d}}{\sigma\sqrt{2\pi}}$$

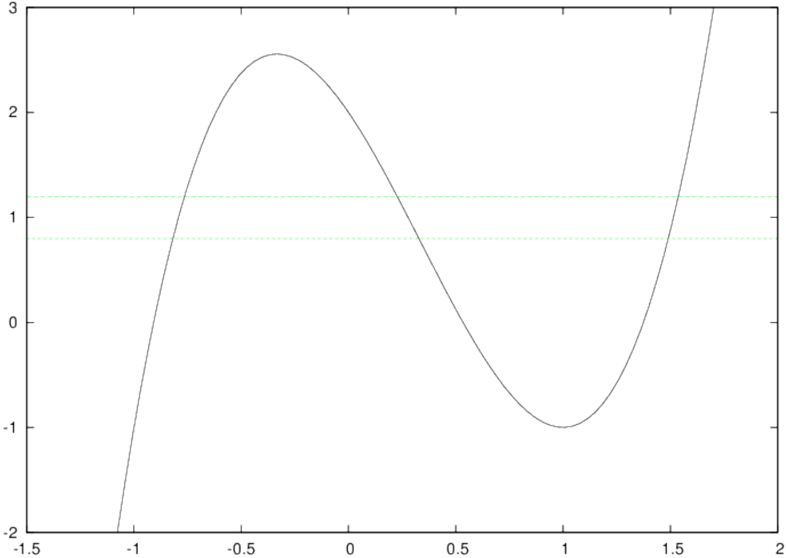
Polynomial anti-concentration

Proof.

1. For a fixed interval, a scaled Chebyshev polynomial has smallest ℓ_∞ norm (order $1/2^d$ when interval is $[-1, 1]$).
2. Since p is degree d, there are at most $d - 1$ changes of sign, hence only $d - 1$ intervals where $p(x)$ is close to any t.
3. Applying the first fact, each interval is of length at most $\epsilon^{1/d}$, each has Gaussian measure $1/\sigma\sqrt{2\pi}$.

□

Polynomial anti-concentration



Eigenvalue spacings

- Want to bound $\Pr\left(\left|\frac{g_i(A_i^T u)}{g_i(A_i^T v)} - \frac{g_j(A_j^T u)}{g_j(A_j^T v)}\right| \leq \epsilon\right)$.
- Condition on a value of $A_j^T u = s$. Then:

$$\begin{aligned}\left|\frac{g_i(A_i^T u)}{g_i(A_i^T v)} - \frac{s}{g_j(A_j^T v)}\right| &= \left|\frac{p_i(A_i^T u)}{g_i(A_i^T v)} + \frac{\epsilon_i}{g_i(A_i^T v)} - \frac{s}{g_j(A_j^T v)}\right| \\ &\geq \left|\frac{p_i(A_i^T u)}{g_i(A_i^T v)} - \frac{s}{g_j(A_j^T v)}\right| - \left|\frac{\epsilon_i}{g_i(A_i^T v)}\right|.\end{aligned}$$

Once we've conditioned on $A_j^T u$ we can pretend $A_i^T u$ is also a Gaussian (of highly reduced variance).

Eigenvalue spacings

- $A_i^T u = \langle A_i, A_j \rangle A_j^T u + r^T u$ where r is orthogonal to A_i
- Variance of remaining randomness is $\|r\|^2 \geq \sigma_m \left(\left[\mathrm{vec} \left(A_i^{\otimes d/2} \right) \right] \right)$.

We conclude by union bounding with the event that denominators are not too large, and then over all pairs i, j.

Extensions

- Can remove Gaussian noise when $x = As + \eta$ and $\eta \sim N(\mu, \Sigma)$.
- Gaussian mixtures (when $x \sim \sum_{i=1}^{n} w_i N(\mu_i, \Sigma_i)$), in the spherical covariance setting. (Gaussian noise applies here too.)

Open problems

- What is the relationship between our method and kernel PCA?
- Independent subspaces.
- Gaussian mixtures: underdetermined and generalized covariance case.

Fin

Questions?