Correspondence retrieval

Alexandr Andoni [†] Daniel Hsu [†] Kevin Shi[†] Xiaorui Sun[♯]

[†]Columbia University, [♯]Simons Institute for the Theory of Computing

July 8th, 2017

Problem setup

Correspondence retrieval

- The universe has unknown vectors $x_1, \cdots, x_k \in \mathbb{R}^d$
- Sample measurement vectors $w_1, \cdots, w_n$
- For each w_i, observe the unordered set $\{w_i^T x_1, \cdots, w_i^T x_k\}$

Problem setup

Special case - phase retrieval (real-valued)

- The universe has a single unknown vector $\overline{x}$
- Sample measurement vectors $w_1, \cdots, w_n$
- For each w_i, observe $|w_i^T \overline{x}|$

This is obtained by setting $k = 2$ and $\overline{w} = \frac{1}{2}(x_1 - x_2)$

Related work

Mixture of linear regressions [YCS14] [YCS16]

- Universe has k hidden model parameters $x_1, \cdots, x_k$
- For each $i = 1, \cdots, n$, sample multinomial random variable z_i and measurement vector w_i
- Observe response-covariate pairs $\{(y_i, w_i)\}_{i=1}^n$ such that

$$y_i = \sum_{j=1}^{k} \langle w_j, x_i \rangle \, \mathbb{1}(z_i = j)$$

Algorithms

- [YCS16] show an efficient inference algorithm with sample complexity $\tilde{O}(k^{10} d)$
- Uses tensor decomposition for mixture models and alternating minimization

Main result

Theorem

Assume the following conditions:

- $n \geq d + 1$
- $w_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ *for* $i = 1, \cdots, n$
- $x_1, \cdots, x_k$ *are linearly dependent with condition number* $\lambda(X)$

Then there is an efficient algorithm which solves the correspondence retrieval using n *measurement vectors.*

Introduces a nonstandard tool in this area - the LLL Lattice Basis Reduction algorithm.

Comparison with related work

Mixture of linear regressions

- Each sample vector w_i corresponds to k samples in the mixture model
- Previous result: $\tilde{O}(k^{10}d)$ samples
- Our result: $k(d+1)$ samples

Real-valued phase retrieval

- Previous result: $2d-1$ measurement vectors can recover all possible hidden $\overline{x}$ [BCE08]
- Our result: $d+1$ measurement vectors suffice to recover any single hidden $\overline{x}$ with high probability

Main idea - reduction to Subset Sum

Subset sum

Given integers $\{a_i\}_{i=1}^{n}$ and a target sum M, determine if there are $z_i \in \{0,1\}$ such that

$$\sum_{i=1}^{n} z_i a_i = M$$

Complexity

- Subset Sum is NP-hard in the worst case, but easy in the average case where the a_i's are uniformly distributed [LO85]
- We extend this to the case where $\sum_{i=1}^{n} z_i a_i$ just needs to satisfy anti-concentration inequalities at every point

Lattices

Definition (Lattice)

Given a collection of linearly independent vectors $b_1, \cdots, b_m \in \mathbb{R}^d$, a lattice $\Lambda\mathbf{B}$ over the basis $\mathbf{B} = \{b_1, \cdots, b_m\}$ is the $\mathbb{Z}$-module of $\mathbf{B}$ as embedded in $\mathbb{R}^d$

$$\Lambda\mathbf{B} = \left\{ \sum_{i=1}^{m} z_i b_i \; : \; z_i \in \mathbb{Z} \right\}$$

Shortest vector problem

Given a lattice basis $\mathbf{B} \subset \mathbb{R}^d$, find the lattice vector $\mathbf{B}z \in \Lambda\mathbf{B}$ s.t.

$$z = \underset{z \in \mathbb{Z} - \{\mathbf{0}\}}{\arg\min} \|\mathbf{B}z\|_2^2$$

Shortest Vector Problem

Hardness of approximation

Shortest vector problem is NP-hard to approximate to within a constant factor.

LLL Lattice Basis Reduction[LLL82]

There is an efficient approximation algorithm for solving the Shortest Vector Problem.

- Approximation factor: $2^{d/2}$
- Running time: $\mathrm{poly}(d, \log \lambda(\mathbf{B}))$

Proof Overview

1. Reduce the correspondence retrieval problem to the shortest vector problem in a lattice with basis B: $\arg\min_{z \in \mathbb{Z}^{dk+1}} \|Bz\|_2$
2. Show that the coefficient vector z with 1's in the correct correspondences produces a lattice vector of norm $\sqrt{d+1}$
3. Show that for a fixed, incorrect z, with high probability $\|Bz\|_2 \geq 2^{(dk+1)/2}\sqrt{d+1}$ over the randomness of the w_i's
4. Under appropriate scaling and a union bound argument, every incorrect z produces a lattice vector with norm at least $2^{(dk+1)/2}\sqrt{d+1}$

Proof Overview

1. Reduce the correspondence retrieval problem to the shortest vector problem in a lattice with basis B: $\arg\min_{z \in \mathbb{Z}^{dk+1}} \|Bz\|_2$
2. Show that the coefficient vector z with 1's in the correct correspondences produces a lattice vector of norm $\sqrt{d+1}$
3. Show that for a fixed, incorrect z, with high probability $\|Bz\|_2 \geq 2^{(dk+1)/2}\sqrt{d+1}$ over the randomness of the w_i's
4. Under appropriate scaling and a union bound argument, every incorrect z produces a lattice vector with norm at least $2^{(dk+1)/2}\sqrt{d+1}$

Proof Overview

1. Reduce the correspondence retrieval problem to the shortest vector problem in a lattice with basis B: $\arg\min_{z \in \mathbb{Z}^{dk+1}} \|Bz\|_2$
2. Show that the coefficient vector z with 1's in the correct correspondences produces a lattice vector of norm $\sqrt{d+1}$
3. Show that for a fixed, incorrect z, with high probability $\|Bz\|_2 \geq 2^{(dk+1)/2}\sqrt{d+1}$ over the randomness of the w_i's
4. Under appropriate scaling and a union bound argument, every incorrect z produces a lattice vector with norm at least $2^{(dk+1)/2}\sqrt{d+1}$

Proof Overview

1. Reduce the correspondence retrieval problem to the shortest vector problem in a lattice with basis B: $\arg\min_{z \in \mathbb{Z}^{dk+1}} \|Bz\|_2$
2. Show that the coefficient vector z with 1's in the correct correspondences produces a lattice vector of norm $\sqrt{d+1}$
3. Show that for a fixed, incorrect z, with high probability $\|Bz\|_2 \geq 2^{(dk+1)/2}\sqrt{d+1}$ over the randomness of the w_i's
4. Under appropriate scaling and a union bound argument, every incorrect z produces a lattice vector with norm at least $2^{(dk+1)/2}\sqrt{d+1}$

Recap

- We defined a new observation model which is loosely inspired by mixture models and which also generalizes phase retrieval
- We show that this observation model admits exact inference with lower sample complexity than either of the above two models
- We describe an algorithm based on a completely different technique - the LLL basis reduction algorithm

Recap

- We defined a new observation model which is loosely inspired by mixture models and which also generalizes phase retrieval
- We show that this observation model admits exact inference with lower sample complexity than either of the above two models
- We describe an algorithm based on a completely different technique - the LLL basis reduction algorithm

Thanks for listening!