Correspondence retrieval

Alexandr Andoni[†], Daniel Hsu[†], Kevin Shi[†], Xiaorui Sun[♯],
[†]Columbia University, [♯]Simons Institute for the Theory of Computing

Problem setup

- ▷ Hidden 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 $\{w_i^T x_1, \cdots, w_i^T x_k\}$
- ▷ Goal is to recover the unknown vectors

Related problems

Phase retrieval (real-valued)

- ▷ Hidden vector $\overline{x}$
- ▷ Sample measurement vectors $w_1, \cdots, w_n$
- ▷ For each w_i, observe $|w_i^T \overline{x}|$
- ▷ Equivalent under $k = 2$ and $\overline{w} = \frac{1}{2}(x_1 - x_2)$

Mixture of linear regressions

- ▷ k hidden model parameters $w_1, \cdots, w_k$
- ▷ For each $i = 1, \cdots, n$, sample multinomial random variable z_i
- ▷ Observe response-covariate pairs $\{(y_i, x_i)\}_{i=1}^n$ such that $y_i = \sum_{j=1}^k \langle w_j, x_i \rangle \mathbb{1}(z_i = j)$

Prior work

Phase retrieval

- ▷ $2d - 1$ measurement vectors are sufficient to recover all possible hidden vectors $\overline{x}$
- ▷ For all frames of $2d - 2$ measurement vectors, the mapping from observations to hidden vectors is ambiguous

Mixture of linear regressions

- ▷ There is an efficient inference algorithm with sample complexity $\tilde{O}(k^{10} d)$
- ▷ Algorithm uses tensor decomposition for mixture models

Main result

Theorem 1.
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.*

- ▷ Each measurement corresponds to k measurements in the mixture of linear regressions model, for a total sample complexity of $k(d + 1)$
- ▷ Running time is dominated by the running time of the LLL algorithm on a basis of norm $2^{O(d^2 k^2)} / \lambda(X)$

Algorithm

Algorithm 1 Lattice algorithm for correspondence retrieval

input Data $(w_i, \mathcal{M}_i)$ for $i \in [d+1]$, parameter $\beta > 0$.
output Set of points $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_k\}$

Let $y_{i,1}, y_{i,2}, \ldots, y_{i,k}$ be an arbitrary ordering the elements of $\mathcal{M}_i$, for each $i \in [d+1]$.

Define $a = (a_{i,j} : i \in [d], j \in [k]) \in \mathbb{R}^{dk}$ by

$$a_{i,j} \ := \ \langle w_{d+1}, \tilde{w}_i \rangle y_{i,j} \,,$$

where $\tilde{w}_i$ is the i-th column of W^{-1}.

for $t = 1, 2, \ldots, k$ **do**

Construct basis

$$B^{(t)} = \begin{bmatrix} b_0^{(t)} & b_{1,1}^{(t)} & \cdots & b_{d,k}^{(t)} \end{bmatrix} := \begin{bmatrix} I_{dk+1} \\ \beta y_{d+1,t} \,|\, -\beta a^\top \end{bmatrix} \in \mathbb{R}^{(dk+2) \times (dk+1)} .$$

Let $L^{(t)}(\hat{z}_0, \hat{z}) := \hat{z}_0 b_0^{(t)} + \sum_{i,j} \hat{z}_{i,j} b_{i,j}^{(t)} \in \Lambda(B^{(t)})$ for $(\hat{z}_0, \hat{z}) \in \mathbb{Z} \times \mathbb{Z}^{dk}$ be the vector returned by LLL as an approximate solution to Shortest Vector Problem for $\Lambda(B^{(t)})$.

Let $\hat{x}_t$ be a solution to the system of linear equations (in $x \in \mathbb{R}^d$)

$$\langle w_i, x \rangle \ = \ y_{i,j} \,, \qquad (i,j) \in [d] \times [k] \,.\, \hat{z}_{i,j} \neq 0 \,,$$

end for

return $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_k$.

Main proof idea

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

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

Lemma 1 (Average case analysis).
Suppose the Subset Sum instance specified by source numbers $\{a_i\}_{i \in \mathcal{I}} \subset \mathbb{R}$ *and target sum* $t \in \mathbb{R}$ *satisfies the following properties.*

- ▷ *There is a subset* $\mathcal{S}^\star \subseteq \mathcal{I}$ *such that* $\sum_{i \in \mathcal{S}^\star} a_i = t$
- ▷ *There exists* $\varepsilon > 0$ *such that*

$$|z_0 \cdot t - \sum_{i \in \mathcal{I}} z_i \cdot a_i| \geq \varepsilon$$

for each (z_0, z) *with bounded norm that is not an integer multiple of* $(1, \chi^\star)$, *where* $\chi^\star \in \{0,1\}^{\mathcal{I}}$ *is the characteristic vector for* $\mathcal{S}^\star$

Then the LLL lattice basis reduction algorithm returns $\chi^\star$ *as the solution*

Lattice tools

Definition 2 (Lattice).
Given a collection of linearly independent vectors $\mathbf{B} := \{b_1, \cdots, b_m \in \mathbb{R}^d\}$, a lattice $\Lambda\mathbf{B}$ over the basis $\mathbf{B}$ is the $\mathbb{Z}$-module of $\mathbf{B}$ embedded in $\mathbb{R}^d$

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

Definition 3 (Shortest vector problem).
Given a lattice basis $\mathbf{B} \subset \mathbb{R}^d$, output a lattice vector $\mathbf{B}z \in \Lambda\mathbf{B}$ where

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

Lemma 2 (LLL Lattice Basis Reduction).
There is an efficient approximation algorithm for solving the Shortest Vector Problem with

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

Reduction to Subset Sum

For each y in $\{y_{d+1,1}, \cdots, y_{d+1,k}\}$:

- ▷ $t := y$
- ▷ $a_{ij} := w_{d+1}^T \tilde{w}_i y_{i,j}$ where $\tilde{w}_i$ is the ith column of W
- ▷ Output subset sum instance $t, \{a_{ij}\}_{i=1,j=1}^{d,k}$

With high probability over the w's, a subset sum solution chooses exactly one a_{ij} for each i, thus identifying the missing correspondences

Reduction to Shortest Vector Problem

$$(1, \mathcal{X}^\star) = \arg\min_{z_0, z} \left\| \begin{bmatrix} I_{dk+1} \\ \beta t \,|\, -\beta a^\top \end{bmatrix} \begin{bmatrix} z_0 \\ z \end{bmatrix} \right\|^2$$

$$= \arg\min_{(z_0, z)} \|(z_0, z)\|^2 + \beta^2 \left(z^T a - t \right)^2$$

Correct correspondence:

- ▷ $z^T a - t = 0$
- ▷ $\|(z_0, z)\| = \sqrt{d+1}$

Proof sketch

Lemma 3.
There is an $\epsilon > 0$ *such that for each incorrect integer coefficient vector* (z_0, z), *with probability* $1 - \delta$, $|z^T a - t| \geq \epsilon$

Lemma 4.
There are at most $\left(2 \cdot 2^{(dk+1)/2} \cdot \sqrt{d+1} + 1\right)^{dk+1}$ *possible integer coefficient vectors* (z_0, z) *with norm less than* $2^{(dk+1)/2}\sqrt{d+1}$

- ▷ β can be set to make δ as small as needed
- ▷ Apply a union bound over the number of possible coefficient vectors from **Lemma 4** to the high probability bound from **Lemma 3**
- ▷ The only vector with norm less than $2^{(dk+1)/2}\sqrt{d+1}$ is the correct solution vector
- ▷ The approximation factor of LLL Lattice Basis Reduction now guarantees finding the correct solution vector

andoni@cs.columbia.edu, djhsu@cs.columbia.edu, kshi@cs.columbia.edu, xiaoruisun@cs.columbia.edu