High-Dimensional Linear Regression and Phase Retrieval without Sparsity: Lattices and Integer Relation Approach

Eren C. Kızıldağ, joint work with David Gamarnik

Massachusetts Institute of Technology

kizildag@mit.edu

2019 IEEE International Symposium on Information Theory

July 10, 2019

Overview

High-Dimensional Linear Regression

Setup:

High-Dimensional Linear Regression

Setup:

- $\beta^* \in \mathbb{R}^p$ (unknown) **feature vector**, p number of features..

High-Dimensional Linear Regression

Setup:

- $\beta^* \in \mathbb{R}^p$ (unknown) **feature vector**, p number of features..
- Observe n **noiseless** linear samples $Y = X\beta^* \in \mathbb{R}^n$ of β^*.

High-Dimensional Linear Regression

Setup:

- $\beta^* \in \mathbb{R}^p$ (unknown) **feature vector**, p number of features..
- Observe n **noiseless** linear samples $Y = X\beta^* \in \mathbb{R}^n$ of β^*.
- Measurement matrix $X \in \mathbb{R}^{n \times p}$, with random entries.

High-Dimensional Linear Regression

Setup:

- $\beta^* \in \mathbb{R}^p$ (unknown) **feature vector**, p number of features..
- Observe n **noiseless** linear samples $Y = X\beta^* \in \mathbb{R}^n$ of β^*.
- Measurement matrix $X \in \mathbb{R}^{n \times p}$, with random entries.

Goal:

Given (Y, X), recover β^* **accurately** and **efficiently** (in polynomial time).

High-Dimensional Linear Regression

Setup:

- $\beta^* \in \mathbb{R}^p$ (unknown) **feature vector**, p number of features..
- Observe n **noiseless** linear samples $Y = X\beta^* \in \mathbb{R}^n$ of β^*.
- Measurement matrix $X \in \mathbb{R}^{n \times p}$, with random entries.

Goal:

Given (Y, X), recover β^* **accurately** and **efficiently** (in polynomial time).

Question:

What is the **smallest** number n of measurements necessary?

High-Dimensional Linear Regression

High-Dimensional Linear Regression

Question:

What is the **smallest** number *n* of measurements necessary to recover β^*?

High-Dimensional Linear Regression

Question:

What is the **smallest** number n of measurements necessary to recover β^*?

- **Without any assumption:** p.

High-Dimensional Linear Regression

Question:

What is the **smallest** number n of measurements necessary to recover β^*?

- **Without any assumption:** p.
- Linear system in p unknowns, **underdetermined** if $n < p$.

High-Dimensional Linear Regression

Question:

What is the **smallest** number n of measurements necessary to recover β^*?

- **Without any assumption:** p.
- Linear system in p unknowns, **underdetermined** if $n < p$.
- Many practical applications (MRI imaging, natural language processing, genomics): $n \ll p$.

High-Dimensional Linear Regression

Question:

What is the **smallest** number n of measurements necessary to recover β^*?

- **Without any assumption:** p.
- Linear system in p unknowns, **underdetermined** if $n < p$.
- Many practical applications (MRI imaging, natural language processing, genomics): $n \ll p$.

Focus:

High-dimensional regime: $n \ll p$ and $p \to +\infty$.

High-Dimensional Linear Regression

Problem of recovering $\beta^* \in \mathbb{R}^p$ from $Y = X\beta^* \in \mathbb{R}^n$ is ill-posed if $n \ll p$, without extra assumptions.

High-Dimensional Linear Regression

Problem of recovering $\beta^* \in \mathbb{R}^p$ from $Y = X\beta^* \in \mathbb{R}^n$ is ill-posed if $n \ll p$, without extra assumptions.

Question:

Is it possible to make the problem well-posed in the regime $n \ll p$, by imposing any structural assumptions on β^*?

High-Dimensional Linear Regression

Problem of recovering $\beta^* \in \mathbb{R}^p$ from $Y = X\beta^* \in \mathbb{R}^n$ is ill-posed if $n \ll p$, without extra assumptions.

Question:

Is it possible to make the problem well-posed in the regime $n \ll p$, by imposing any structural assumptions on β^*?

- (Arguably) Most popular assumption in literature: **Sparsity**.

High-Dimensional Linear Regression

Problem of recovering $\beta^* \in \mathbb{R}^p$ from $Y = X\beta^* \in \mathbb{R}^n$ is ill-posed if $n \ll p$, without extra assumptions.

Question:

Is it possible to make the problem well-posed in the regime $n \ll p$, by imposing any structural assumptions on β^*?

- (Arguably) Most popular assumption in literature: **Sparsity**.
- Meaning: $s = |\{i : \beta_i^* \neq 0\}|$ is small, compared to p.

High-Dimensional Linear Regression

Problem of recovering $\beta^* \in \mathbb{R}^p$ from $Y = X\beta^* \in \mathbb{R}^n$ is ill-posed if $n \ll p$, without extra assumptions.

Question:

Is it possible to make the problem well-posed in the regime $n \ll p$, by imposing any structural assumptions on β^*?

- (Arguably) Most popular assumption in literature: **Sparsity**.
- Meaning: $s = |\{i : \beta_i^* \neq 0\}|$ is small, compared to p.
- Polynomial-time algorithms (such as LASSO [Wainwright '09], OMP [Fletcher et al. '11], and Basis Pursuit) exist.

High-Dimensional Linear Regression

Problem of recovering $\beta^* \in \mathbb{R}^p$ from $Y = X\beta^* \in \mathbb{R}^n$ is ill-posed if $n \ll p$, without extra assumptions.

Question:

Is it possible to make the problem well-posed in the regime $n \ll p$, by imposing any structural assumptions on β^*?

- (Arguably) Most popular assumption in literature: **Sparsity**.
- Meaning: $s = |\{i : \beta_i^* \neq 0\}|$ is small, compared to p.
- Polynomial-time algorithms (such as LASSO [Wainwright '09], OMP [Fletcher et al. '11], and Basis Pursuit) exist.
- Essentially, (efficient) recovery of β^* is possible if:

$$n > s \log \frac{p}{s}.$$

High-Dimensional Linear Regression

High-Dimensional Linear Regression

- Efficient recovery for $s-$sparse β^* (namely, $|\{i : \beta_i^* \neq 0\}| < s$) possible if: $n > s \log \frac{p}{s}$.

High-Dimensional Linear Regression

- Efficient recovery for $s-$sparse β^* (namely, $|\{i : \beta_i^* \neq 0\}| < s$) possible if: $n > s \log \frac{p}{s}$.
- Number of samples $n \to +\infty$ as $p \to +\infty$.

High-Dimensional Linear Regression

- Efficient recovery for $s-$sparse β^* (namely, $|\{i : \beta_i^* \neq 0\}| < s$) possible if: $n > s \log \frac{p}{s}$.
- Number of samples $n \to +\infty$ as $p \to +\infty$.
- **Question:** What happens if $n = o(s \log(p/s))$?

High-Dimensional Linear Regression

- Efficient recovery for $s-$sparse β^* (namely, $|\{i : \beta_i^* \neq 0\}| < s$) possible if: $n > s \log \frac{p}{s}$.
- Number of samples $n \to +\infty$ as $p \to +\infty$.
- **Question:** What happens if $n = o(s \log(p/s))$?
  - Basis Pursuit/BPDN **fail** to recover **sparse & binary** β^* [Donoho-Tanner '10]

High-Dimensional Linear Regression

- Efficient recovery for $s-$sparse β^* (namely, $|\{i : \beta_i^* \neq 0\}| < s$) possible if: $n > s \log \frac{p}{s}$.
- Number of samples $n \to +\infty$ as $p \to +\infty$.
- **Question:** What happens if $n = o(s \log(p/s))$?
  - Basis Pursuit/BPDN **fail** to recover **sparse & binary** β^* [Donoho-Tanner '10]
  - LASSO **fails** to solve **support recovery** problem, if $n = o(s \log p)$ [Wainwright '09]

High-Dimensional Linear Regression

- Efficient recovery for $s-$sparse β^* (namely, $|\{i : \beta_i^* \neq 0\}| < s$) possible if: $n > s\log\frac{p}{s}$.
- Number of samples $n \to +\infty$ as $p \to +\infty$.
- **Question:** What happens if $n = o(s\log(p/s))$?
  - Basis Pursuit/BPDN **fail** to recover **sparse & binary** β^* [Donoho-Tanner '10]
  - LASSO **fails** to solve **support recovery** problem, if $n = o(s\log p)$ [Wainwright '09]
  - Evidence of **computational hardness** for $n < s\log(p/s)$ [Gamarnik-Zadik '17]

High-Dimensional Linear Regression

- Efficient recovery for $s-$sparse β^* (namely, $|\{i : \beta_i^* \neq 0\}| < s$) possible if: $n > s \log \frac{p}{s}$.
- Number of samples $n \to +\infty$ as $p \to +\infty$.
- **Question:** What happens if $n = o(s \log(p/s))$?
  - Basis Pursuit/BPDN **fail** to recover **sparse & binary** β^* [Donoho-Tanner '10]
  - LASSO **fails** to solve **support recovery** problem, if $n = o(s \log p)$ [Wainwright '09]
  - Evidence of **computational hardness** for $n < s \log(p/s)$ [Gamarnik-Zadik '17]
  - Other models (Tree-Sparsity [He-Carin '09], Block-Sparsity [Eldar et al. '10], Generative Model [Bora et al. '17]): Work for $n < p$, but not for n very small.

High-Dimensional Linear Regression

- Efficient recovery for $s-$sparse β^* (namely, $|\{i : \beta_i^* \neq 0\}| < s$) possible if: $n > s \log \frac{p}{s}$.
- Number of samples $n \to +\infty$ as $p \to +\infty$.
- **Question:** What happens if $n = o(s \log(p/s))$?
    - Basis Pursuit/BPDN **fail** to recover **sparse & binary** β^* [Donoho-Tanner '10]
    - LASSO **fails** to solve **support recovery** problem, if $n = o(s \log p)$ [Wainwright '09]
    - Evidence of **computational hardness** for $n < s \log(p/s)$ [Gamarnik-Zadik '17]
    - Other models (Tree-Sparsity [He-Carin '09], Block-Sparsity [Eldar et al. '10], Generative Model [Bora et al. '17]): Work for $n < p$, but not for n very small.

Question:

Can we address $n = o(s \log(p/s))$ regime? Any hope for $n = O(1)$ regime?

A Related Recent Work (by Gamarnik and Zadik)

A Related Recent Work (by Gamarnik and Zadik)

- Let $\beta^* \in \mathbb{R}^p$, $Y = X\beta^* + W \in \mathbb{R}^n$, $W \sim N(0, \sigma^2)$, and $Q \in \mathbb{Z}^+$.

A Related Recent Work (by Gamarnik and Zadik)

- Let $\beta^* \in \mathbb{R}^p$, $Y = X\beta^* + W \in \mathbb{R}^n$, $W \sim N(0, \sigma^2)$, and $Q \in \mathbb{Z}^+$.
- **Structural Assumption:** $\beta_i^* = K_i / Q$, where $K_i \in \mathbb{Z}$, and $|K_i| \leqslant R$.

A Related Recent Work (by Gamarnik and Zadik)

- Let $\beta^* \in \mathbb{R}^p$, $Y = X\beta^* + W \in \mathbb{R}^n$, $W \sim N(0, \sigma^2)$, and $Q \in \mathbb{Z}^+$.
- **Structural Assumption:** $\beta_i^* = K_i/Q$, where $K_i \in \mathbb{Z}$, and $|K_i| \leqslant R$.
- Application: Half of $\beta_i^* \in \mathbb{Z}$ for linear GPS models (namely, $Q = 1$) [Boyd-Hassibi '98].

A Related Recent Work (by Gamarnik and Zadik)

- Let $\beta^* \in \mathbb{R}^p$, $Y = X\beta^* + W \in \mathbb{R}^n$, $W \sim N(0, \sigma^2)$, and $Q \in \mathbb{Z}^+$.
- **Structural Assumption:** $\beta_i^* = K_i/Q$, where $K_i \in \mathbb{Z}$, and $|K_i| \leqslant R$.
- Application: Half of $\beta_i^* \in \mathbb{Z}$ for linear GPS models (namely, $Q = 1$) [Boyd-Hassibi '98].
- **Thm [Gamarnik-Zadik '18]:** Recovery of β^* (w.h.p. as $p \to +\infty$, in $\mathrm{poly}(p, n, Q, R)$ time) even with $n = 1$ is possible, provided σ small.

A Related Recent Work (by Gamarnik and Zadik)

- Let $\beta^* \in \mathbb{R}^p$, $Y = X\beta^* + W \in \mathbb{R}^n$, $W \sim N(0, \sigma^2)$, and $Q \in \mathbb{Z}^+$.
- **Structural Assumption:** $\beta_i^* = K_i/Q$, where $K_i \in \mathbb{Z}$, and $|K_i| \leqslant R$.
- Application: Half of $\beta_i^* \in \mathbb{Z}$ for linear GPS models (namely, $Q = 1$) [Boyd-Hassibi '98].
- **Thm [Gamarnik-Zadik '18]:** Recovery of β^* (w.h.p. as $p \to +\infty$, in $\mathrm{poly}(p, n, Q, R)$ time) even with $n = 1$ is possible, provided σ small.
- Algorithm motivated from random subset-sum problem in cryptography; and based on LLL lattice basis reduction algorithm.

This Work

This Work

Question 1:

Is it possible to make the problem (given $Y = X\beta^* \in \mathbb{R}^n$, infer $\beta^* \in \mathbb{R}^p$) well-posed when $n = 1$, and β^* has irrational entries?

This Work

Question 1:

Is it possible to make the problem (given $Y = X\beta^* \in \mathbb{R}^n$, infer $\beta^* \in \mathbb{R}^p$) well-posed when $n = 1$, and β^* has irrational entries?

Question 2:

Is there an efficient algorithm to recover β^* when $n = 1$? In other words, is it possible to ensure no statistical-computational gap?

This Work

Question 1:

Is it possible to make the problem (given $Y = X\beta^* \in \mathbb{R}^n$, infer $\beta^* \in \mathbb{R}^p$) well-posed when $n = 1$, and β^* has irrational entries?

Question 2:

Is there an efficient algorithm to recover β^* when $n = 1$? In other words, is it possible to ensure no statistical-computational gap?

- **Structural Assumption:** β^* supp. on $\mathcal{S}$ with $|\mathcal{S}| = \mathrm{poly}(p)$, rationally independent.

This Work

Question 1:

Is it possible to make the problem (given $Y = X\beta^* \in \mathbb{R}^n$, infer $\beta^* \in \mathbb{R}^p$) well-posed when $n = 1$, and β^* has irrational entries?

Question 2:

Is there an efficient algorithm to recover β^* when $n = 1$? In other words, is it possible to ensure no statistical-computational gap?

- **Structural Assumption:** β^* supp. on $\mathcal{S}$ with $|\mathcal{S}| = \mathrm{poly}(p)$, rationally independent.
- **Answer:** Yes, to both. $\Rightarrow$ No **statistical-computational gap**.

This Work

Question 1:

Is it possible to make the problem (given $Y = X\beta^* \in \mathbb{R}^n$, infer $\beta^* \in \mathbb{R}^p$) well-posed when $n = 1$, and β^* has irrational entries?

Question 2:

Is there an efficient algorithm to recover β^* when $n = 1$? In other words, is it possible to ensure no statistical-computational gap?

- **Structural Assumption:** β^* supp. on $\mathcal{S}$ with $|\mathcal{S}| = \mathrm{poly}(p)$, rationally independent.
- **Answer:** Yes, to both. $\Rightarrow$ No **statistical-computational gap**.
- Algorithmic connection to subset-sum and integer relation detection problems; and to lattices.

Preliminaries

Rational Independence:

Set $\mathcal{S} = \{a_1, \dots, a_R\} \subset \mathbb{R}$ is **rationally independent**, if $\forall q_1, \dots, q_R \in \mathbb{Q}$:
$\sum_{i=1}^{R} q_i a_i = 0 \Rightarrow q_i = 0$, for all i.

Preliminaries

Rational Independence:

Set $\mathcal{S} = \{a_1, \dots, a_R\} \subset \mathbb{R}$ is **rationally independent**, if $\forall q_1, \dots, q_R \in \mathbb{Q}$: $\sum_{i=1}^R q_i a_i = 0 \Rightarrow q_i = 0$, for all i.

Integer Relation Detection Problem:

Rational Independence:

Set $\mathcal{S} = \{a_1, \ldots, a_R\} \subset \mathbb{R}$ is **rationally independent**, if $\forall q_1, \ldots, q_R \in \mathbb{Q}$: $\sum_{i=1}^{R} q_i a_i = 0 \Rightarrow q_i = 0$, for all i.

Integer Relation Detection Problem:

- Given $\mathbf{b} \in \mathbb{R}^n$, find an $\mathbf{x} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ such that $\langle \mathbf{x}, \mathbf{b} \rangle = 0$.

Preliminaries

Rational Independence:

Set $\mathcal{S} = \{a_1, \dots, a_R\} \subset \mathbb{R}$ is **rationally independent**, if $\forall q_1, \dots, q_R \in \mathbb{Q}$: $\sum_{i=1}^{R} q_i a_i = 0 \Rightarrow q_i = 0$, for all i.

Integer Relation Detection Problem:

- Given $\mathbf{b} \in \mathbb{R}^n$, find an $\mathbf{x} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ such that $\langle \mathbf{x}, \mathbf{b} \rangle = 0$.
- $\mathbf{x} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ called an **integer relation** for $\mathbf{b} \in \mathbb{R}^n$.

Preliminaries

Rational Independence:

Set $\mathcal{S} = \{a_1, \ldots, a_R\} \subset \mathbb{R}$ is **rationally independent**, if $\forall q_1, \ldots, q_R \in \mathbb{Q}$: $\sum_{i=1}^{R} q_i a_i = 0 \Rightarrow q_i = 0$, for all i.

Integer Relation Detection Problem:

- Given $\mathbf{b} \in \mathbb{R}^n$, find an $\mathbf{x} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ such that $\langle \mathbf{x}, \mathbf{b} \rangle = 0$.
- $\mathbf{x} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ called an **integer relation** for $\mathbf{b} \in \mathbb{R}^n$.
- Well studied (**Example:** Apply Euclidean algorithm when $\mathbf{b} \in \mathbb{R}^2$).

Preliminaries

Rational Independence:

Set $\mathcal{S} = \{a_1, \ldots, a_R\} \subset \mathbb{R}$ is **rationally independent**, if $\forall q_1, \ldots, q_R \in \mathbb{Q}$:
$\sum_{i=1}^{R} q_i a_i = 0 \Rightarrow q_i = 0$, for all i.

Integer Relation Detection Problem:

- Given $\mathbf{b} \in \mathbb{R}^n$, find an $\mathbf{x} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ such that $\langle \mathbf{x}, \mathbf{b} \rangle = 0$.
- $\mathbf{x} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ called an **integer relation** for $\mathbf{b} \in \mathbb{R}^n$.
- Well studied (**Example:** Apply Euclidean algorithm when $\mathbf{b} \in \mathbb{R}^2$).
- **PSLQ alg.:**

Preliminaries

Rational Independence:

Set $\mathcal{S} = \{a_1, \ldots, a_R\} \subset \mathbb{R}$ is **rationally independent**, if $\forall q_1, \ldots, q_R \in \mathbb{Q}$:
$\sum_{i=1}^{R} q_i a_i = 0 \Rightarrow q_i = 0$, for all i.

Integer Relation Detection Problem:

- Given $\mathbf{b} \in \mathbb{R}^n$, find an $\mathbf{x} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ such that $\langle \mathbf{x}, \mathbf{b} \rangle = 0$.
- $\mathbf{x} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ called an **integer relation** for $\mathbf{b} \in \mathbb{R}^n$.
- Well studied (**Example:** Apply Euclidean algorithm when $\mathbf{b} \in \mathbb{R}^2$).
- **PSLQ alg.:**
  - Let $\mathbf{m} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ be a relation for $\mathbf{b}$ with smallest $\|\mathbf{m}\|$.

Preliminaries

Rational Independence:

Set $\mathcal{S} = \{a_1, \ldots, a_R\} \subset \mathbb{R}$ is **rationally independent**, if $\forall q_1, \ldots, q_R \in \mathbb{Q}$: $\sum_{i=1}^{R} q_i a_i = 0 \Rightarrow q_i = 0$, for all i.

Integer Relation Detection Problem:

- Given $\mathbf{b} \in \mathbb{R}^n$, find an $\mathbf{x} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ such that $\langle \mathbf{x}, \mathbf{b} \rangle = 0$.
- $\mathbf{x} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ called an **integer relation** for $\mathbf{b} \in \mathbb{R}^n$.
- Well studied (**Example:** Apply Euclidean algorithm when $\mathbf{b} \in \mathbb{R}^2$).
- **PSLQ alg.:**
    - Let $\mathbf{m} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ be a relation for $\mathbf{b}$ with smallest $\|\mathbf{m}\|$.
    - **Thm (Ferguson et al. '99)** PSLQ provably returns a relation for $\mathbf{b} \in \mathbb{R}^n$ after $O(n^3 + n^2 \log \|\mathbf{m}\|)$ arithmetic operations over reals.

Preliminaries

Rational Independence:

Set $\mathcal{S} = \{a_1, \ldots, a_R\} \subset \mathbb{R}$ is **rationally independent**, if $\forall q_1, \ldots, q_R \in \mathbb{Q}$: $\sum_{i=1}^{R} q_i a_i = 0 \Rightarrow q_i = 0$, for all i.

Integer Relation Detection Problem:

- Given $\mathbf{b} \in \mathbb{R}^n$, find an $\mathbf{x} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ such that $\langle \mathbf{x}, \mathbf{b} \rangle = 0$.
- $\mathbf{x} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ called an **integer relation** for $\mathbf{b} \in \mathbb{R}^n$.
- Well studied (**Example:** Apply Euclidean algorithm when $\mathbf{b} \in \mathbb{R}^2$).
- **PSLQ alg.:**
  - Let $\mathbf{m} \in \mathbb{Z}^n \setminus \{\mathbf{0}\}$ be a relation for $\mathbf{b}$ with smallest $\|\mathbf{m}\|$.
  - **Thm (Ferguson et al. '99)** PSLQ provably returns a relation for $\mathbf{b} \in \mathbb{R}^n$ after $O(n^3 + n^2 \log \|\mathbf{m}\|)$ arithmetic operations over reals.
- One of top 10 algorithms of past century by *IEEE Computer Society*.

Preliminaries

Preliminaries

Subset-Sum Problem

Subset-Sum Problem

- **Given:** $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{Z}^n$ and $Y \in \mathbb{Z}$; find an $S \subseteq [n]$: $\sum_{i \in S} X_i = Y$.

Subset-Sum Problem

- **Given:** $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{Z}^n$ and $Y \in \mathbb{Z}$; find an $S \subseteq [n]$: $\sum_{i \in S} X_i = Y$.
- Interpretations:

Subset-Sum Problem

- **Given:** $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{Z}^n$ and $Y \in \mathbb{Z}$; find an $S \subseteq [n]$: $\sum_{i \in S} X_i = Y$.
- Interpretations:
  - **Regression:** $\beta^* \in \{0,1\}^n$, $Y = \langle \mathbf{X}, \beta^* \rangle$. Given $(Y, \mathbf{X})$, recover β^*.

Preliminaries

Subset-Sum Problem

- **Given:** $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{Z}^n$ and $Y \in \mathbb{Z}$; find an $S \subseteq [n]$: $\sum_{i \in S} X_i = Y$.
- Interpretations:
  - **Regression:** $\beta^* \in \{0,1\}^n$, $Y = \langle \mathbf{X}, \beta^* \rangle$. Given $(Y, \mathbf{X})$, recover β^*.
  - **Cryptographic:** $\beta^* \in \{0,1\}^n$ **plaintext**, $Y = \langle \mathbf{X}, \beta^* \rangle$ **ciphertext**, and $\mathbf{X}$ **public information**.

Preliminaries

Subset-Sum Problem

- **Given:** $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{Z}^n$ and $Y \in \mathbb{Z}$; find an $S \subseteq [n]$: $\sum_{i \in S} X_i = Y$.
- Interpretations:
  - **Regression:** $\beta^* \in \{0,1\}^n$, $Y = \langle \mathbf{X}, \beta^* \rangle$. Given $(Y, \mathbf{X})$, recover β^*.
  - **Cryptographic:** $\beta^* \in \{0,1\}^n$ **plaintext**, $Y = \langle \mathbf{X}, \beta^* \rangle$ **ciphertext**, and **X public information**.
- **NP-hard** in worst-case.

Preliminaries

Subset-Sum Problem

- **Given:** $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{Z}^n$ and $Y \in \mathbb{Z}$; find an $S \subseteq [n]$: $\sum_{i \in S} X_i = Y$.
- Interpretations:
  - **Regression:** $\beta^* \in \{0,1\}^n$, $Y = \langle \mathbf{X}, \beta^* \rangle$. Given $(Y, \mathbf{X})$, recover β^*.
  - **Cryptographic:** $\beta^* \in \{0,1\}^n$ **plaintext**, $Y = \langle \mathbf{X}, \beta^* \rangle$ **ciphertext**, and $\mathbf{X}$ **public information**.
- **NP-hard** in worst-case.
- **Average-Case Complexity** ([Lagarias & Odlyzko '85] and [Frieze '86]):

Subset-Sum Problem

- **Given:** $\mathbf{X} = (X_1, \dots, X_n) \in \mathbb{Z}^n$ and $Y \in \mathbb{Z}$; find an $S \subseteq [n]$: $\sum_{i \in S} X_i = Y$.
- Interpretations:
  - **Regression:** $\beta^* \in \{0,1\}^n$, $Y = \langle \mathbf{X}, \beta^* \rangle$. Given $(Y, \mathbf{X})$, recover β^*.
  - **Cryptographic:** $\beta^* \in \{0,1\}^n$ **plaintext**, $Y = \langle \mathbf{X}, \beta^* \rangle$ **ciphertext**, and **X public information**.
- **NP-hard** in worst-case.
- **Average-Case Complexity** ([Lagarias & Odlyzko '85] and [Frieze '86]):
  - Let $X_i \sim \mathrm{Unif}\{1, 2, \dots, 2^{cn^2}\}$ iid with $c > 1/2$.

Preliminaries

Subset-Sum Problem

- **Given:** $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{Z}^n$ and $Y \in \mathbb{Z}$; find an $S \subseteq [n]$: $\sum_{i \in S} X_i = Y$.
- Interpretations:
  - **Regression:** $\beta^* \in \{0,1\}^n$, $Y = \langle \mathbf{X}, \beta^* \rangle$. Given $(Y, \mathbf{X})$, recover β^*.
  - **Cryptographic:** $\beta^* \in \{0,1\}^n$ **plaintext**, $Y = \langle \mathbf{X}, \beta^* \rangle$ **ciphertext**, and $\mathbf{X}$ **public information**.
- **NP-hard** in worst-case.
- **Average-Case Complexity** ([Lagarias & Odlyzko '85] and [Frieze '86]):
  - Let $X_i \sim \mathrm{Unif}\{1, 2, \ldots, 2^{cn^2}\}$ iid with $c > 1/2$.
  - LLL algorithm [Lenstra et al. '82] recovers β^* whp as $n \to +\infty$ in $\mathrm{poly}(n)$ time.

Overview

Main Result (Discrete)

Theorem (Gamarnik & K., 2019)

Let $Y = \mathbf{X}\beta^* \in \mathbb{R}$ *with:*

- $\mathbf{X} \in \mathbb{Z}^{1\times p}$ *with iid entries;* $\exists N \in \mathbb{Z}^+$, *such that :* $\mathbb{E}[|X_1|] \leqslant O(2^N)$ *and* $\mathbb{P}(X_i = x) \leqslant O(2^{-N})$, *for every* $x \in \mathbb{Z}$.
- $\beta_i^* \in \mathcal{S} = \{a_1, \dots, a_R\}$, *rationally independent, known to learner,* $R = \mathrm{poly}(p)$.

There exists an algorithm, recovering β^* *whp (as* $p \to +\infty$*) in* $\mathrm{poly}(p, N, R)$ *time, provided* $N \geqslant (\frac{1}{2} + \epsilon)p^2$ *for any* $\epsilon > 0$.

Main Result (Discrete)

Theorem (Gamarnik & K., 2019)

Let $Y = \mathbf{X}\beta^* \in \mathbb{R}$ *with:*

- $\mathbf{X} \in \mathbb{Z}^{1\times p}$ *with iid entries;* $\exists N \in \mathbb{Z}^+$*, such that :* $\mathbb{E}[|X_1|] \leqslant O(2^N)$ *and* $\mathbb{P}(X_i = x) \leqslant O(2^{-N})$*, for every* $x \in \mathbb{Z}$*.*
- $\beta_i^* \in \mathcal{S} = \{a_1, \dots, a_R\}$*, rationally independent, known to learner,* $R = \mathrm{poly}(p)$*.*

There exists an algorithm, recovering β^* *whp (as* $p \to +\infty$*) in* $\mathrm{poly}(p, N, R)$ *time, provided* $N \geqslant (\frac{1}{2} + \epsilon)p^2$ *for any* $\epsilon > 0$*.*

- Single sample ($n = 1$), efficient recovery ($\mathrm{poly}(p, N, R)$ time).

Main Result (Discrete)

Theorem (Gamarnik & K., 2019)

Let $Y = \mathbf{X}\beta^* \in \mathbb{R}$ *with:*

- $\mathbf{X} \in \mathbb{Z}^{1\times p}$ *with iid entries;* $\exists N \in \mathbb{Z}^+$, *such that :* $\mathbb{E}[|X_1|] \leqslant O(2^N)$ *and* $\mathbb{P}(X_i = x) \leqslant O(2^{-N})$, *for every* $x \in \mathbb{Z}$.
- $\beta_i^* \in \mathcal{S} = \{a_1, \dots, a_R\}$, *rationally independent, known to learner,* $R = \mathrm{poly}(p)$.

There exists an algorithm, recovering β^* *whp (as* $p \to +\infty$*) in* $\mathrm{poly}(p, N, R)$ *time, provided* $N \geqslant (\frac{1}{2} + \epsilon)p^2$ *for any* $\epsilon > 0$.

- Single sample ($n = 1$), efficient recovery ($\mathrm{poly}(p, N, R)$ time).
- Works, provided (iid) entries of $\mathbf{X}$ are from a large integer support.

Main Result (Discrete)

Theorem (Gamarnik & K., 2019)

Let $Y = \mathbf{X}\beta^* \in \mathbb{R}$ with:

- $\mathbf{X} \in \mathbb{Z}^{1\times p}$ *with iid entries;* $\exists N \in \mathbb{Z}^+$, *such that :* $\mathbb{E}[|X_1|] \leqslant O(2^N)$ *and* $\mathbb{P}(X_i = x) \leqslant O(2^{-N})$, *for every* $x \in \mathbb{Z}$.
- $\beta_i^* \in \mathcal{S} = \{a_1, \dots, a_R\}$, *rationally independent, known to learner,* $R = \mathrm{poly}(p)$.

There exists an algorithm, recovering β^* *whp (as* $p \to +\infty$*) in* $\mathrm{poly}(p, N, R)$ *time, provided* $N \geqslant (\frac{1}{2} + \epsilon)p^2$ *for any* $\epsilon > 0$.

- Single sample ($n = 1$), efficient recovery ($\mathrm{poly}(p, N, R)$ time).
- Works, provided (iid) entries of $\mathbf{X}$ are from a large integer support.
- Uses PSLQ (integer relation) + LLL (lattice reduction) oracles.

Main Result (Continuous)

Theorem (Gamarnik & K., 2019)

Let $Y = \mathbf{X}\beta^* \in \mathbb{R}$ *with:*

- $\mathbf{X} \in \mathbb{R}^{1\times p}$, *jointly continuous.*
- $\beta_i^* \in \mathcal{S} = \{a_1, \dots, a_R\}$, *rationally independent, known to learner,* $R = \mathrm{poly}(p)$.

There exists an algorithm, recovering β^* *almost surely in* $\mathrm{poly}(p, R)$ *time.*

Main Result (Continuous)

Theorem (Gamarnik & K., 2019)

Let $Y = \mathbf{X}\beta^* \in \mathbb{R}$ *with:*

- $\mathbf{X} \in \mathbb{R}^{1\times p}$*, jointly continuous.*
- $\beta_i^* \in \mathcal{S} = \{a_1, \dots, a_R\}$*, rationally independent, known to learner,* $R = \mathrm{poly}(p)$*.*

There exists an algorithm, recovering β^* *almost surely in* $\mathrm{poly}(p, R)$ *time.*

- Single measurement ($n = 1$), efficient recovery ($\mathrm{poly}(p, R)$ time).

Main Result (Continuous)

Theorem (Gamarnik & K., 2019)

Let $Y = \mathbf{X}\beta^* \in \mathbb{R}$ *with:*

- $\mathbf{X} \in \mathbb{R}^{1\times p}$, *jointly continuous.*
- $\beta_i^* \in \mathcal{S} = \{a_1, \dots, a_R\}$, *rationally independent, known to learner,* $R = \mathrm{poly}(p)$.

There exists an algorithm, recovering β^* *almost surely in* $\mathrm{poly}(p, R)$ *time.*

- Single measurement ($n = 1$), efficient recovery ($\mathrm{poly}(p, R)$ time).
- Only joint continuity of $\mathbf{X}$ is required.

Main Result (Continuous)

Theorem (Gamarnik & K., 2019)

Let $Y = \mathbf{X}\beta^* \in \mathbb{R}$ *with:*

- $\mathbf{X} \in \mathbb{R}^{1\times p}$, *jointly continuous.*
- $\beta_i^* \in \mathcal{S} = \{a_1, \dots, a_R\}$, *rationally independent, known to learner,* $R = \mathrm{poly}(p)$.

There exists an algorithm, recovering β^* *almost surely in* $\mathrm{poly}(p, R)$ *time.*

- Single measurement ($n = 1$), efficient recovery ($\mathrm{poly}(p, R)$ time).
- Only joint continuity of $\mathbf{X}$ is required.
- Needs only PSLQ (integer relation) oracle.

Existence of Information

Information exists even with one sample ($n = 1$)!

Existence of Information

Information exists even with one sample ($n = 1$)!

Lemma

The following holds almost surely: For every $\beta, \beta^* \in \mathcal{S}^p$*, and jointly continuous random vector* $\mathbf{X} \in \mathbb{R}^p$*;* $\mathbf{X}\beta$ *and* $\mathbf{X}\beta^*$ *are distinct. Thus,* **brute-force search** *works.*

Existence of Information

Information exists even with one sample ($n = 1$)!

Lemma

The following holds almost surely: For every $\beta, \beta^* \in \mathcal{S}^p$, and jointly continuous random vector $\mathbf{X} \in \mathbb{R}^p$; $\mathbf{X}\beta$ and $\mathbf{X}\beta^*$ are distinct. Thus, **brute-force search** *works.*

Proof.

$\mathbb{P}(\text{Lemma}^c) = \mathbb{P}\left(\exists \beta \neq \beta^* : \mathbf{X}\beta = \mathbf{X}\beta^*\right) \leqslant R^{2p}\mathbb{P}\left(\mathbf{X}(\beta - \beta^*) = 0\right) = 0.$ □

Existence of Information

Information exists even with one sample ($n = 1$)!

Lemma

The following holds almost surely: For every $\beta, \beta^* \in \mathcal{S}^p$, and jointly continuous random vector $\mathbf{X} \in \mathbb{R}^p$; $\mathbf{X}\beta$ and $\mathbf{X}\beta^*$ are distinct. Thus, **brute-force search** *works.*

Proof.

$\mathbb{P}(\text{Lemma}^c) = \mathbb{P}\left(\exists \beta \neq \beta^* : \mathbf{X}\beta = \mathbf{X}\beta^*\right) \leqslant R^{2p}\mathbb{P}\left(\mathbf{X}(\beta - \beta^*) = 0\right) = 0.$ □

- Similar argument applies also to measurement matrix with discrete entries.

Existence of Information

Information exists even with one sample ($n = 1$)!

Lemma

The following holds almost surely: For every $\beta, \beta^* \in \mathcal{S}^p$, and jointly continuous random vector $\mathbf{X} \in \mathbb{R}^p$; $\mathbf{X}\beta$ and $\mathbf{X}\beta^*$ are distinct. Thus, **brute-force search** *works.*

Proof.

$$\mathbb{P}(\text{Lemma}^c) = \mathbb{P}\left(\exists \beta \neq \beta^* : \mathbf{X}\beta = \mathbf{X}\beta^*\right) \leqslant R^{2p}\mathbb{P}\left(\mathbf{X}(\beta - \beta^*) = 0\right) = 0.$$

□

- Similar argument applies also to measurement matrix with discrete entries.
- Brute force takes $O(R^p)$ time (exponential in p).

Existence of Information

Information exists even with one sample ($n = 1$)!

Lemma

The following holds almost surely: For every $\beta, \beta^* \in \mathcal{S}^p$, and jointly continuous random vector $\mathbf{X} \in \mathbb{R}^p$; $\mathbf{X}\beta$ and $\mathbf{X}\beta^*$ are distinct. Thus, **brute-force search** *works.*

Proof.

$\mathbb{P}(\text{Lemma}^c) = \mathbb{P}\left(\exists \beta \neq \beta^* : \mathbf{X}\beta = \mathbf{X}\beta^*\right) \leqslant R^{2p}\mathbb{P}\left(\mathbf{X}(\beta - \beta^*) = 0\right) = 0.$ □

- Similar argument applies also to measurement matrix with discrete entries.
- Brute force takes $O(R^p)$ time (exponential in p).
- Our result: Polynomial-time decoding.

Existence of Information

Information exists even with one sample ($n = 1$)!

Lemma

The following holds almost surely: For every $\beta, \beta^* \in \mathcal{S}^p$, and jointly continuous random vector $\mathbf{X} \in \mathbb{R}^p$; $\mathbf{X}\beta$ and $\mathbf{X}\beta^*$ are distinct. Thus, **brute-force search** *works.*

Proof.

$$\mathbb{P}(\text{Lemma}^c) = \mathbb{P}\left(\exists \beta \neq \beta^* : \mathbf{X}\beta = \mathbf{X}\beta^*\right) \leqslant R^{2p}\mathbb{P}\left(\mathbf{X}(\beta - \beta^*) = 0\right) = 0. \qquad \square$$

- Similar argument applies also to measurement matrix with discrete entries.
- Brute force takes $O(R^p)$ time (exponential in p).
- Our result: Polynomial-time decoding.
- **No statistical-computational gap,** when β^* supported on **rationally independent** $\mathcal{S}$.

Proof Idea (Discrete)

- **Recall:** $Y = X\beta^* \in \mathbb{R}$, $X \in \mathbb{Z}^p$ iid. $\beta_i^* \in \mathcal{S} = \{a_1, \ldots, a_R\}$ rat. independent.

Proof Idea (Discrete)

- **Recall:** $Y = X\beta^* \in \mathbb{R}$, $X \in \mathbb{Z}^p$ iid. $\beta_i^* \in \mathcal{S} = \{a_1, \dots, a_R\}$ rat. independent.
- Y is an integer combination of $\mathcal{S}$:

$$Y = \sum_{k=1}^{R} \theta_k^* a_k \quad \text{where} \quad \theta_k^* = \sum_{j:\beta_j^* = a_k} X_j \in \mathbb{Z}.$$

Proof Idea (Discrete)

- **Recall:** $Y = X\beta^* \in \mathbb{R}$, $X \in \mathbb{Z}^p$ iid. $\beta_i^* \in \mathcal{S} = \{a_1, \ldots, a_R\}$ rat. independent.
- Y is an integer combination of $\mathcal{S}$:

$$Y = \sum_{k=1}^{R} \theta_k^* a_k \quad \text{where} \quad \theta_k^* = \sum_{j:\beta_j^* = a_k} X_j \in \mathbb{Z}.$$

- Thus, $\exists$ an integer relation $\mathbf{m} \in \mathbb{Z}^{R+1} \setminus \{\mathbf{0}\}$ for the vector $\mathcal{A} = (Y, a_i : i \in [R]) \in \mathbb{R}^{R+1}$.

Proof Idea (Discrete)

- **Recall:** $Y = X\beta^* \in \mathbb{R}$, $X \in \mathbb{Z}^p$ iid. $\beta_i^* \in \mathcal{S} = \{a_1, \ldots, a_R\}$ rat. independent.
- Y is an integer combination of $\mathcal{S}$:

$$Y = \sum_{k=1}^{R} \theta_k^* a_k \quad \text{where} \quad \theta_k^* = \sum_{j:\beta_j^* = a_k} X_j \in \mathbb{Z}.$$

- Thus, $\exists$ an integer relation $\mathbf{m} \in \mathbb{Z}^{R+1} \setminus \{\mathbf{0}\}$ for the vector $\mathcal{A} = (Y, a_i : i \in [R]) \in \mathbb{R}^{R+1}$.
- PSLQ with input $\mathbf{Y}$ and $\{a_1, \ldots, a_R\}$ recovers an $\mathbf{m}$ in $\mathrm{poly}(p, R, N)$ time.

Proof Idea (Discrete)

- **Recall:** $Y = X\beta^* \in \mathbb{R}$, $X \in \mathbb{Z}^p$ iid. $\beta_i^* \in \mathcal{S} = \{a_1, \ldots, a_R\}$ rat. independent.
- Y is an integer combination of $\mathcal{S}$:

$$Y = \sum_{k=1}^{R} \theta_k^* a_k \quad \text{where} \quad \theta_k^* = \sum_{j:\beta_j^* = a_k} X_j \in \mathbb{Z}.$$

- Thus, $\exists$ an integer relation $\mathbf{m} \in \mathbb{Z}^{R+1} \setminus \{\mathbf{0}\}$ for the vector $\mathcal{A} = (Y, a_i : i \in [R]) \in \mathbb{R}^{R+1}$.
- PSLQ with input $\mathbf{Y}$ and $\{a_1, \ldots, a_R\}$ recovers an $\mathbf{m}$ in $\mathrm{poly}(p, R, N)$ time.
- $\mathbf{m}$ is essentially **unique** up to a constant multiple:

$$\{a_1, \ldots, a_R\} \text{ rat. indep. } \Rightarrow \mathbf{m} = -m_0(-1, \theta_1^*, \ldots, \theta_R^*)$$

Proof Idea (Discrete)

- **Recall:** $Y = X\beta^* \in \mathbb{R}$, $X \in \mathbb{Z}^p$ iid. $\beta_i^* \in \mathcal{S} = \{a_1, \ldots, a_R\}$ rat. independent.
- Y is an integer combination of $\mathcal{S}$:

$$Y = \sum_{k=1}^{R} \theta_k^* a_k \quad \text{where} \quad \theta_k^* = \sum_{j:\beta_j^* = a_k} X_j \in \mathbb{Z}.$$

- Thus, $\exists$ an integer relation $\mathbf{m} \in \mathbb{Z}^{R+1} \setminus \{\mathbf{0}\}$ for the vector $\mathcal{A} = (Y, a_i : i \in [R]) \in \mathbb{R}^{R+1}$.
- PSLQ with input $\mathbf{Y}$ and $\{a_1, \ldots, a_R\}$ recovers an $\mathbf{m}$ in $\mathrm{poly}(p, R, N)$ time.
- $\mathbf{m}$ is essentially **unique** up to a constant multiple:

$$\{a_1, \ldots, a_R\} \text{ rat. indep. } \Rightarrow \mathbf{m} = -m_0(-1, \theta_1^*, \ldots, \theta_R^*)$$

- $(\theta_1^*, \ldots, \theta_R^*)$ can be obtained in $\mathrm{poly}(p, N, R)$ time.

Proof Idea (Discrete)

- Recall:

$$Y = \sum_{k=1}^{R} \theta_k^* a_k \quad \text{where} \quad \theta_k^* = \sum_{j:\beta_j^* = a_k} X_j \in \mathbb{Z}, \forall k.$$

$(\theta_1^*, \ldots, \theta_R^*)$ recovered in $\mathrm{poly}(p, N, R)$ time.

Proof Idea (Discrete)

- Recall:

$$Y = \sum_{k=1}^{R} \theta_k^* a_k \quad \text{where} \quad \theta_k^* = \sum_{j:\beta_j^*=a_k} X_j \in \mathbb{Z}, \forall k.$$

  $(\theta_1^*, \ldots, \theta_R^*)$ recovered in $\mathrm{poly}(p, N, R)$ time.
- Define $S_k = \{j \in [p] : \beta_j^* = a_k\} \subseteq \{1, \ldots, p\}$. We have $\theta_k^* = \sum_{j \in S_k} X_j$.

Proof Idea (Discrete)

- Recall:

$$Y = \sum_{k=1}^{R} \theta_k^* a_k \quad \text{where} \quad \theta_k^* = \sum_{j:\beta_j^* = a_k} X_j \in \mathbb{Z}, \forall k.$$

  $(\theta_1^*, \ldots, \theta_R^*)$ recovered in $\mathrm{poly}(p, N, R)$ time.
- Define $S_k = \{j \in [p] : \beta_j^* = a_k\} \subseteq \{1, \ldots, p\}$. We have $\theta_k^* = \sum_{j \in S_k} X_j$.
- **Problem:** Given $X_1, \ldots, X_p \in \mathbb{Z}$, and $\theta_k^* = \sum_{j \in S_k} X_j \in \mathbb{Z}$, obtain $S_k \subseteq [p]$.

Proof Idea (Discrete)

- Recall:

$$Y = \sum_{k=1}^{R} \theta_k^* a_k \quad \text{where} \quad \theta_k^* = \sum_{j:\beta_j^* = a_k} X_j \in \mathbb{Z}, \forall k.$$

  $(\theta_1^*, \ldots, \theta_R^*)$ recovered in $\mathrm{poly}(p, N, R)$ time.
- Define $S_k = \{j \in [p] : \beta_j^* = a_k\} \subseteq \{1, \ldots, p\}$. We have $\theta_k^* = \sum_{j \in S_k} X_j$.
- **Problem:** Given $X_1, \ldots, X_p \in \mathbb{Z}$, and $\theta_k^* = \sum_{j \in S_k} X_j \in \mathbb{Z}$, obtain $S_k \subseteq [p]$.
- This is precisely the **subset-sum problem**!

Proof Idea (Discrete)

- Recall:

$$Y = \sum_{k=1}^{R} \theta_k^* a_k \quad \text{where} \quad \theta_k^* = \sum_{j:\beta_j^* = a_k} X_j \in \mathbb{Z}, \forall k.$$

  $(\theta_1^*, \dots, \theta_R^*)$ recovered in $\mathrm{poly}(p, N, R)$ time.
- Define $S_k = \{j \in [p] : \beta_j^* = a_k\} \subseteq \{1, \dots, p\}$. We have $\theta_k^* = \sum_{j \in S_k} X_j$.
- **Problem:** Given $X_1, \dots, X_p \in \mathbb{Z}$, and $\theta_k^* = \sum_{j \in S_k} X_j \in \mathbb{Z}$, obtain $S_k \subseteq [p]$.
- This is precisely the **subset-sum problem**!
- Apply LLL algorithm (à la Frieze) to conclude.

Proof Idea (Continuous)

- **Recall:** $Y = X\beta^* \in \mathbb{R}$, $X \in \mathbb{R}^p$ jointly continuous, $\beta_i^* \in \mathcal{S} = \{a_1, \dots, a_R\}$, $\mathcal{S}$ rationally independent, available to learner.

Proof Idea (Continuous)

- **Recall:** $Y = X\beta^* \in \mathbb{R}$, $X \in \mathbb{R}^p$ jointly continuous, $\beta_i^* \in \mathcal{S} = \{a_1, \ldots, a_R\}$, $\mathcal{S}$ rationally independent, available to learner.
- Let $\mathcal{L} = \{X_i a_j : 1 \leqslant i \leqslant p, 1 \leqslant j \leqslant R\}$.

Proof Idea (Continuous)

- **Recall:** $Y = X\beta^* \in \mathbb{R}$, $X \in \mathbb{R}^p$ jointly continuous, $\beta_i^* \in \mathcal{S} = \{a_1, \dots, a_R\}$, $\mathcal{S}$ rationally independent, available to learner.
- Let $\mathcal{L} = \{X_i a_j : 1 \leqslant i \leqslant p, 1 \leqslant j \leqslant R\}$.

Lemma

$\mathbb{P}(\mathcal{L}$ *is rationally independent*$) = 1$

Proof Idea (Continuous)

- **Recall:** $Y = X\beta^* \in \mathbb{R}$, $X \in \mathbb{R}^p$ jointly continuous, $\beta_i^* \in \mathcal{S} = \{a_1, \dots, a_R\}$, $\mathcal{S}$ rationally independent, available to learner.
- Let $\mathcal{L} = \{X_i a_j : 1 \leqslant i \leqslant p, 1 \leqslant j \leqslant R\}$.

Lemma

$\mathbb{P}(\mathcal{L}$ *is rationally independent*$) = 1$

- Y is an integer combination of $\mathcal{L}$: $Y = \sum_{i=1}^{p} \sum_{j=1}^{R} X_i a_j \xi_{ij}^*$ where $\xi_{ij}^* \in \{0, 1\}$.

Proof Idea (Continuous)

- **Recall:** $Y = X\beta^* \in \mathbb{R}$, $X \in \mathbb{R}^p$ jointly continuous, $\beta_i^* \in \mathcal{S} = \{a_1, \ldots, a_R\}$, $\mathcal{S}$ rationally independent, available to learner.
- Let $\mathcal{L} = \{X_i a_j : 1 \leqslant i \leqslant p, 1 \leqslant j \leqslant R\}$.

Lemma

$\mathbb{P}(\mathcal{L} \textit{ is rationally independent}) = 1$

- Y is an integer combination of $\mathcal{L}$: $Y = \sum_{i=1}^p \sum_{j=1}^R X_i a_j \xi_{ij}^*$ where $\xi_{ij}^* \in \{0, 1\}$.
- $\exists$ an integer relation $\mathbf{m}$ for vector $\mathcal{A}' = (Y, X_i a_j : i \in [p], j \in [R]) \in \mathbb{R}^{pR+1}$.

Proof Idea (Continuous)

- **Recall:** $Y = X\beta^* \in \mathbb{R}$, $X \in \mathbb{R}^p$ jointly continuous, $\beta_i^* \in \mathcal{S} = \{a_1, \ldots, a_R\}$, $\mathcal{S}$ rationally independent, available to learner.
- Let $\mathcal{L} = \{X_i a_j : 1 \leqslant i \leqslant p, 1 \leqslant j \leqslant R\}$.

Lemma

$\mathbb{P}(\mathcal{L}$ *is rationally independent*$) = 1$

- Y is an integer combination of $\mathcal{L}$: $Y = \sum_{i=1}^{p} \sum_{j=1}^{R} X_i a_j \xi_{ij}^*$ where $\xi_{ij}^* \in \{0,1\}$.
- $\exists$ an integer relation $\mathbf{m}$ for vector $\mathcal{A}' = (Y, X_i a_j : i \in [p], j \in [R]) \in \mathbb{R}^{pR+1}$.
- $\mathcal{L}$ rationally independent $\Rightarrow$ $\mathbf{m}$ is of form $\mathbf{m} = k(-1, \xi_{ij}^* : i \in [p], j \in [R])$, $k \in \mathbb{Z} \setminus \{0\}$.

Proof Idea (Continuous)

- **Recall:** $Y = X\beta^* \in \mathbb{R}$, $X \in \mathbb{R}^p$ jointly continuous, $\beta_i^* \in \mathcal{S} = \{a_1, \dots, a_R\}$, $\mathcal{S}$ rationally independent, available to learner.
- Let $\mathcal{L} = \{X_i a_j : 1 \leqslant i \leqslant p, 1 \leqslant j \leqslant R\}$.

Lemma

$\mathbb{P}(\mathcal{L}$ *is rationally independent*$) = 1$

- Y is an integer combination of $\mathcal{L}$: $Y = \sum_{i=1}^{p} \sum_{j=1}^{R} X_i a_j \xi_{ij}^*$ where $\xi_{ij}^* \in \{0, 1\}$.
- $\exists$ an integer relation $\mathbf{m}$ for vector $\mathcal{A}' = (Y, X_i a_j : i \in [p], j \in [R]) \in \mathbb{R}^{pR+1}$.
- $\mathcal{L}$ rationally independent $\Rightarrow$ $\mathbf{m}$ is of form $\mathbf{m} = k(-1, \xi_{ij}^* : i \in [p], j \in [R])$, $k \in \mathbb{Z} \setminus \{0\}$.
- PSLQ recovers an $\mathbf{m}$ in $\mathrm{poly}(p, R)$ time, from which $\{\xi_{ij}^* : i \in [p], j \in [R]\}$is obtained.

Overview

Phase Retrieval

Problem (Phase Retrieval)

Problem (Phase Retrieval)

- Let $\beta^* \in \mathbb{C}^p$. Learner sees n measurements $Y_i = |\langle X_i, \beta^* \rangle|$, $i \in [n]$.

Phase Retrieval

Problem (Phase Retrieval)

- Let $\beta^* \in \mathbb{C}^p$. Learner sees n measurements $Y_i = |\langle X_i, \beta^* \rangle|$, $i \in [n]$.
- **Goal:** Recover β^* efficiently and accurately, with smallest possible number n of samples.

Phase Retrieval

Problem (Phase Retrieval)

- Let $\beta^* \in \mathbb{C}^p$. Learner sees n measurements $Y_i = |\langle X_i, \beta^* \rangle|$, $i \in [n]$.
- **Goal:** Recover β^* efficiently and accurately, with smallest possible number n of samples.

Theorem (Gamarnik & K., 2019 (Informal))

Let $Y = |\langle X, \beta^* \rangle| \in \mathbb{R}$ *with:*

- $X \in \mathbb{Z}_+^p$ *with iid entries over a large support, or* $X \in \mathbb{R}^p$ *with iid continuous entries*
- $\beta_i^* \in \mathcal{S} = \{a_1, \ldots, a_R\} \subset \mathbb{C}$, *known to learner;* $\mathcal{S}' = \{a_i^H a_j + a_i a_j^H : i, j \in [R]\}$ *rationally independent.*

Then, there exists an algorithm recovering β^* *whp,* $\mathrm{poly}(p, R, \cdot)$ *time.*

Phase Retrieval: Subset-Sum with Dependent Inputs

- High-level proof idea is similar.

Phase Retrieval: Subset-Sum with Dependent Inputs

- High-level proof idea is similar.
- Y^2 integral combination of $\mathcal{S}' = \{a_i^H a_j + a_i a_j^H : i, j \in [R]\}$.

Phase Retrieval: Subset-Sum with Dependent Inputs

- High-level proof idea is similar.
- Y^2 integral combination of $\mathcal{S}' = \{a_i^H a_j + a_i a_j^H : i, j \in [R]\}$.
- (Integer) Relation coefficients are more involved.

Phase Retrieval: Subset-Sum with Dependent Inputs

- High-level proof idea is similar.
- Y^2 integral combination of $\mathcal{S}' = \{a_i^H a_j + a_i a_j^H : i, j \in [R]\}$.
- (Integer) Relation coefficients are more involved.
- Need to solve a subset sum problem with dependent inputs of form: Given $\theta^* \in \mathbb{Z}$ and $(X_i)_{i=1}^p \subset \mathbb{Z}$ iid, recover $\xi_{ij} \in \{0, 1\}$, where

$$\theta^* = \sum_{1 \leqslant i < j \leqslant p} X_i X_j \xi_{ij}.$$

Phase Retrieval: Subset-Sum with Dependent Inputs

- High-level proof idea is similar.
- Y^2 integral combination of $\mathcal{S}' = \{a_i^H a_j + a_i a_j^H : i,j \in [R]\}$.
- (Integer) Relation coefficients are more involved.
- Need to solve a subset sum problem with dependent inputs of form: Given $\theta^* \in \mathbb{Z}$ and $(X_i)_{i=1}^p \subset \mathbb{Z}$ iid, recover $\xi_{ij} \in \{0,1\}$, where

$$\theta^* = \sum_{1 \leqslant i < j \leqslant p} X_i X_j \xi_{ij}.$$

Theorem (Gamarnik & K., 2019)

- *Let* $\mathbf{X} = (X_i)_{i=1}^p$ *iid,* $\exists N \in \mathbb{Z}^+$ *such that* $\mathbb{P}(X_i = x) \leqslant O(2^{-N})$ *and* $\mathbb{E}[X_i] \leqslant O(2^N)$.
- $\theta^* = \sum_{i<j} X_i X_j \xi_{ij}$ *with* $\xi_{ij} \in \{0,1\}$.

Then, there exists an algorithm, which takes $(\theta^*, \mathbf{X})$ *as input and recovers* ξ_{ij} *whp in* $\mathrm{poly}(p,N)$ *time, provided* $N \geqslant (1/8+\epsilon)p^4$ *for any* $\epsilon > 0$.

Overview

Contributions

Contributions

- New **efficient** algorithm for high-dimensional linear regression problem, when $\beta^* \in \mathbb{R}^p$ supported on a **rationally independent** set of $\mathrm{poly}(p)$ size.

Contributions

- New **efficient** algorithm for high-dimensional linear regression problem, when $\beta^* \in \mathbb{R}^p$ supported on a **rationally independent** set of $\mathrm{poly}(p)$ size.
- Algorithm provably recovers $\beta^* \in \mathbb{R}^p$ w.h.p. as $p \to +\infty$, even with **one linear measurement** $Y = X\beta^* \in \mathbb{R}$, for a large class of distributions for entries of X.
    - In this regime, **sparsity-based methods** are known to **fail**!

Contributions

- New **efficient** algorithm for high-dimensional linear regression problem, when $\beta^* \in \mathbb{R}^p$ supported on a **rationally independent** set of $\mathrm{poly}(p)$ size.
- Algorithm provably recovers $\beta^* \in \mathbb{R}^p$ w.h.p. as $p \to +\infty$, even with **one linear measurement** $Y = X\beta^* \in \mathbb{R}$, for a large class of distributions for entries of X.
  - In this regime, **sparsity-based methods** are known to **fail**!
- Side product: LLL algorithm works for subset-sum problem with **dependent** inputs.

Contributions

- New **efficient** algorithm for high-dimensional linear regression problem, when $\beta^* \in \mathbb{R}^p$ supported on a **rationally independent** set of $\mathrm{poly}(p)$ size.
- Algorithm provably recovers $\beta^* \in \mathbb{R}^p$ w.h.p. as $p \to +\infty$, even with **one linear measurement** $Y = X\beta^* \in \mathbb{R}$, for a large class of distributions for entries of X.
  - In this regime, **sparsity-based methods** are known to **fail**!
- Side product: LLL algorithm works for subset-sum problem with **dependent** inputs.
- Algorithmic connection to certain discrete problems: integer-relation detection, subset-sum, approximate short vector.

Contributions

- New **efficient** algorithm for high-dimensional linear regression problem, when $\beta^* \in \mathbb{R}^p$ supported on a **rationally independent** set of $\mathrm{poly}(p)$ size.
- Algorithm provably recovers $\beta^* \in \mathbb{R}^p$ w.h.p. as $p \to +\infty$, even with **one linear measurement** $Y = X\beta^* \in \mathbb{R}$, for a large class of distributions for entries of X.
  - In this regime, **sparsity-based methods** are known to **fail**!
- Side product: LLL algorithm works for subset-sum problem with **dependent** inputs.
- Algorithmic connection to certain discrete problems: integer-relation detection, subset-sum, approximate short vector.
- No statistical-computational gap under our assumptions.

Thank you!