Sparse High Dimensional Linear Regression: Estimating squared error and a Phase Transition

Ilias Zadik, joint work with D. Gamarnik

Massachusetts Institute of Technology (MIT)

30th Conference on Learning Theory (COLT) 2017

Introduction

The Linear Regression Problem:

Introduction

The Linear Regression Problem:
Setup: Let $\beta^* \in \mathbb{R}^p$. For some **measurement matrix** $X \in \mathbb{R}^{n \times p}$, and **noise vector** $W \in \mathbb{R}^n$, we observe n noisy linear samples of β^*, $Y \in \mathbb{R}^n$, given by

$$Y := X\beta^* + W.$$

.

Goal: Given (Y, X), **recover** β^*.

Introduction

The Linear Regression Problem:
Setup: Let $\beta^* \in \mathbb{R}^{\mathsf{p}}$. For some **measurement matrix** $\mathsf{X} \in \mathbb{R}^{\mathsf{n} \times \mathsf{p}}$, and **noise vector** $\mathsf{W} \in \mathbb{R}^{\mathsf{n}}$, we observe n noisy linear samples of β^*, $\mathsf{Y} \in \mathbb{R}^{\mathsf{n}}$, given by

$$\mathsf{Y} := \mathsf{X}\beta^* + \mathsf{W}.$$

.

Goal: Given (Y, X), **recover** β^*.

(Notation: We call p the number of **features** and n the number of **samples**.)

Main Question:

Question: "What is the **minimum** n (numbers of samples) we need to recover β^* in some general Linear Regression setting?"

Main Question:

Question: "What is the **minimum** n (numbers of samples) we need to recover β^* in some general Linear Regression setting?"

An immediate answer under full generality: at least p.

Main Question:

Question: "What is the **minimum** n (numbers of samples) we need to recover β^* in some general Linear Regression setting?"

An immediate answer under full generality: at least p.

Reason: Even if $W = 0$, we have $Y = X\beta^*$, a linear system with p unknowns and n equations!
To solve it, we need at least p equations, i.e. $n \geq p$.

Problem: A High Dimensional Reality

In many real-life applications (e.g. natural language processing, computational biology, computer vision, image processing etc) of Linear Regression we observe **much more** features than samples (i.e. $n \ll p$.)

Problem: A High Dimensional Reality

In many real-life applications (e.g. natural language processing, computational biology, computer vision, image processing etc) of Linear Regression we observe **much more** features than samples (i.e. $n \ll p$.)

Question: Are we doomed to not use all the features or can we handle such a situation?

Put Structural Assumptions on β^*

(1) Sparsity assumption; we assume β^*_i is zero for all $i \in \{1, 2, \ldots, p\}$ except a subset of the indices of cardinality $k \ll p$.

Put Structural Assumptions on β^*

(1) Sparsity assumption; we assume β^*_i is zero for all $i \in \{1, 2, \ldots, p\}$ except a subset of the indices of cardinality $k \ll p$.

Appears a lot

- in applications; e.g. in signal and image coding [Mallat and Zhang '93].
- in theory; e.g. in Compressed Sensing ([Candes, Tao '06], [Donoho '06]).

Put Structural Assumptions on β^*

(1) Sparsity assumption; we assume β_i^* is zero for all $i \in \{1, 2, \ldots, p\}$ except a subset of the indices of cardinality $k \ll p$.

Appears a lot

- in applications; e.g. in signal and image coding [Mallat and Zhang '93].
- in theory; e.g. in Compressed Sensing ([Candes, Tao '06], [Donoho '06]).

(2) We assume binary β_i^*'s, i.e. we assume $\beta^* \in \{0, 1\}^p$.

Put Structural Assumptions on β^*

(1) Sparsity assumption; we assume β_i^* is zero for all $i \in \{1, 2, \ldots, p\}$ except a subset of the indices of cardinality $k \ll p$.

Appears a lot

- in applications; e.g. in signal and image coding [Mallat and Zhang '93].
- in theory; e.g. in Compressed Sensing ([Candes, Tao '06], [Donoho '06]).

(2) We assume binary β_i^*'s, i.e. we assume $\beta^* \in \{0, 1\}^p$.

Less known in the literature, but

Put Structural Assumptions on β^*

(1) Sparsity assumption; we assume β^*_i is zero for all $i \in \{1, 2, \ldots, p\}$ except a subset of the indices of cardinality $k \ll p$.

Appears a lot

- in applications; e.g. in signal and image coding [Mallat and Zhang '93].
- in theory; e.g. in Compressed Sensing ([Candes, Tao '06], [Donoho '06]).

(2) We assume binary β^*_i's, i.e. we assume $\beta^* \in \{0, 1\}^p$.

Less known in the literature, but

- Discrete structure $\Rightarrow$ easier to analyze.

Put Structural Assumptions on β^*

(1) Sparsity assumption; we assume β^*_i is zero for all $i \in \{1, 2, \ldots, p\}$ except a subset of the indices of cardinality $k \ll p$.

Appears a lot

- in applications; e.g. in signal and image coding [Mallat and Zhang '93].
- in theory; e.g. in Compressed Sensing ([Candes, Tao '06], [Donoho '06]).

(2) We assume binary β^*_i's, i.e. we assume $\beta^* \in \{0, 1\}^p$.

Less known in the literature, but

- Discrete structure $\Rightarrow$ easier to analyze.
- Keeps the challenge of **support recovery** (a highly nontrivial task)

Put Structural Assumptions on β^*

(1) Sparsity assumption; we assume β^*_i is zero for all $i \in \{1, 2, \ldots, p\}$ except a subset of the indices of cardinality $k \ll p$.

Appears a lot

- in applications; e.g. in signal and image coding [Mallat and Zhang '93].
- in theory; e.g. in Compressed Sensing ([Candes, Tao '06], [Donoho '06]).

(2) We assume binary β^*_i's, i.e. we assume $\beta^* \in \{0, 1\}^p$.

Less known in the literature, but

- Discrete structure $\Rightarrow$ easier to analyze.
- Keeps the challenge of **support recovery** (a highly nontrivial task)
- Best known information theoretic lower bound is **much smaller** than the best known algorithmic upper bound.

Assumptions on X, W

We assume that

(1) $X_{i,j}$ is i.i.d. standard normal $N(0,1)$ for all $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$.

(2) W_i is i.i.d. normal $N(0, \sigma^2)$ for $i = 1, 2, \ldots, n$, where $\sigma^2 = o(k)$.

(3) X, W are independent.

Classic in literature ([Candes, Tao '06], [Donoho '06],[Wainwright '09])

The New Model

Setup: Let $\beta^* \in \{0,1\}^p$ be a **binary** k-**sparse** vector. For

- $X \in \mathbb{R}^{n \times p}$ consisting of entries i.i.d $N(0,1)$ **random variables**
- $W \in \mathbb{R}^n$ consisting of entries i.i.d. $N(0,\sigma^2)$ **random variables** with $\sigma^2 = o(k)$

we get n noisy linear samples of β^*, $Y \in \mathbb{R}^n$, given by,

$$Y := X\beta^* + W.$$

The New Model

Setup: Let $\beta^* \in \{0,1\}^p$ be a **binary** k-**sparse** vector. For

- $X \in \mathbb{R}^{n \times p}$ consisting of entries i.i.d $N(0,1)$ **random variables**
- $W \in \mathbb{R}^n$ consisting of entries i.i.d. $N(0,\sigma^2)$ **random variables** with $\sigma^2 = o(k)$

we get n noisy linear samples of β^*, $Y \in \mathbb{R}^n$, given by,

$$Y := X\beta^* + W.$$

Goal: Given (Y,X), recover β^* with the minimum number of samples. The recovery should happen with probability tending to 1 as the problem parameters tend to infinity **(w.h.p.)**.

(Very-Brief) Literature Review

- **Upper bounds** ([Candes, Tao '06],[Donoho '09],[Wainwright '09])
  If
  $$n > 2k \log p$$
  LASSO and other efficient algorithms recover β^* w.h.p. .

(Very-Brief) Literature Review

- **Upper bounds** ([Candes, Tao '06],[Donoho '09],[Wainwright '09]) If

$$n > 2k \log p$$

  LASSO and other efficient algorithms recover β^* w.h.p. .
- **Lower bounds** ([Wang et al '10])

(Very-Brief) Literature Review

- **Upper bounds** ([Candes, Tao '06],[Donoho '09],[Wainwright '09])
  If
  $$n > 2k \log p$$
  LASSO and other efficient algorithms recover β^* w.h.p. .
- **Lower bounds** ([Wang et al '10])

  If $n < n^* := \frac{2k}{\log\left(\frac{2k}{\sigma^2}+1\right)} \log p$, then there is **no recovery mechanism** of β^* which succeeds w.h.p.

The next natural question:
Is it **possible** to recover β^* for n with

$$n^* < n < 2k\log p?$$

If yes, is there an **efficient** way to make this recovery?

The next natural question:
Is it **possible** to recover β^* for n with

$$n^* < n < 2k \log p?$$

If yes, is there an **efficient** way to make this recovery?

Main Results: We answer **yes** to the first question, and conjecture (based on geometrical arguments) that the answer is **no** to the second.

Maximum Likelihood Estimator

It has a **simple-to-state form:** the MLE $\hat{\beta}$ is the optimal solution of

$$(\Phi_2) : \min_{\beta \in \{0,1\}^p, \sum_{i=1}^p \beta_i = k} ||Y - X\beta||_2.$$

Maximum Likelihood Estimator- "All or Nothing" Theorem

Definition

For $\beta \in \{0,1\}^p$, k-sparse we define

$$\text{Overlap}(\beta) := |\text{Support}(\beta^*) \cap \text{Support}(\beta)|.$$

Maximum Likelihood Estimator- "All or Nothing" Theorem

Definition

For $\beta \in \{0,1\}^{\mathrm{p}}$, k-sparse we define

$$\mathrm{Overlap}(\beta) := |\mathrm{Support}(\beta^*) \cap \mathrm{Support}(\beta)|.$$

Theorem ("All or nothing")

(Gamarnik,Z. 2016) Set $\mathrm{n}^* := \frac{2\mathrm{k}}{\log\left(\frac{2\mathrm{k}}{\sigma^2}+1\right)} \log \mathrm{p}$ *and let* $\epsilon > 0$ *be arbitrary.*

- *If* $\mathrm{n} < (1-\epsilon)\,\mathrm{n}^*$, *then w.h.p.* $\frac{1}{\mathrm{k}}\mathrm{Overlap}(\hat{\beta}) \to 0$, *as* $\mathrm{n}, \mathrm{p}, \mathrm{k} \to +\infty$.
- *If* $\mathrm{n} > (1+\epsilon)\,\mathrm{n}^*$, *then w.h.p.* $\frac{1}{\mathrm{k}}\mathrm{Overlap}(\hat{\beta}) \to 1$, *as* $\mathrm{n}, \mathrm{p}, \mathrm{k} \to +\infty$.

Maximum Likelihood Estimator

Comments:

(1) Information **exists** when $n > (1+\epsilon)n^*$!

Maximum Likelihood Estimator

Comments:

(1) Information **exists** when $n > (1+\epsilon)n^*$!

(2) A **sharp** phase transition!

Maximum Likelihood Estimator

Comments:

(1) Information **exists** when $n > (1+\epsilon)n^*$!

(2) A **sharp** phase transition!

(3) A challenging application of **the second moment method**.

Algorithmic Hardness (?)

Question: Why no efficient algorithm is known when $n^* < n < 2k \log p$ and many are when $n > 2k \log p$?

Algorithmic Hardness (?)

Question: Why no efficient algorithm is known when $n^* < n < 2k \log p$ and many are when $n > 2k \log p$?

A *usual* picture in the analysis of randoms CSPs. Theory of random CSPs suggests that a usual reason is an **"important change in the geometry of the space of solutions"** between the two regimes.[Achlioptas et al, 2008].

Algorithmic Hardness (?)

Question: Why no efficient algorithm is known when $n^* < n < 2k \log p$ and many are when $n > 2k \log p$?

A *usual* picture in the analysis of randoms CSPs. Theory of random CSPs suggests that a usual reason is an **"important change in the geometry of the space of solutions"** between the two regimes.[Achlioptas et al, 2008].

Usually when such a property holds no efficient algorithm exists and when it ceases, even "local" algorithms work (remember yesterday's talk).

Algorithmic Hardness (?)

Question: Why no efficient algorithm is known when $n^* < n < 2k \log p$ and many are when $n > 2k \log p$?

A *usual* picture in the analysis of randoms CSPs. Theory of random CSPs suggests that a usual reason is an **"important change in the geometry of the space of solutions"** between the two regimes.[Achlioptas et al, 2008].

Usually when such a property holds no efficient algorithm exists and when it ceases, even "local" algorithms work (remember yesterday's talk).

Various names: shattering property, overlap gap property.

The Overlap Gap Property (OGP) for Linear Regression

The OGP (informally): The set of β's with "small" $\|Y - X\beta\|_2$ **"shatters" in two components**, one where β have **low** overlap with the ground truth β^* and one where they have **high** overlap with β^*.

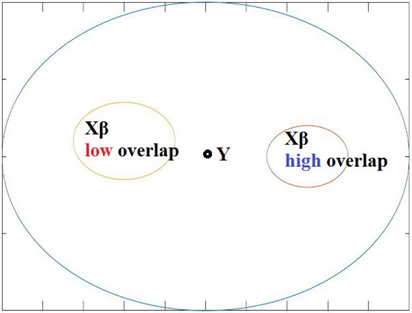

Figure: The OGP around Y

The Overlap Gap Property for Linear Regression-definition

For $r > 0$, set $S_r := \{\beta \in \{0,1\}^p : ||\beta||_0 = k, n^{-\frac{1}{2}}||Y - X\beta||_2 < r\}$.

Definition (The Overlap Gap Property)

Let $r > 0$ and $0 < \zeta_1 < \zeta_2 < 1$. We say that the high-dimensional linear regression problem defined by (X, W, β^*) satisfies the Overlap Gap Property with parameters (r, ζ_1, ζ_2) if the following holds.

(a) For every $\beta \in S_r$,

$$\frac{1}{k}\text{Overlap}(\beta) < \zeta_1 \text{ or } \frac{1}{k}\text{Overlap}(\beta) > \zeta_2.$$

(b) Both the sets

$$S_r \cap \{\beta : \frac{1}{k}\text{Overlap}(\beta) < \zeta_1\} \text{ and } S_r \cap \{\beta : \frac{1}{k}\text{Overlap}(\beta) > \zeta_2\}$$

are non-empty.

The Overlap Gap Property- The result

Theorem

There exists $C > c > 0$ *such that,*

- *If* $n^* < n < ck \log p$ *then w.h.p. OGP holds for some* $r = r_k$ *and* $0 < \zeta_1 < \zeta_2 < 1$.
- *If* $n > Ck \log p$ *then w.h.p. OGP does* **not** *hold for any choice of* $r = r_k$ *and* $0 < \zeta_1 < \zeta_2 < 1$*.(post-COLT)*

The Overlap Gap Property- The result

Theorem

There exists $C > c > 0$ *such that,*

- *If* $n^* < n < ck \log p$ *then w.h.p. OGP holds for some* $r = r_k$ *and* $0 < \zeta_1 < \zeta_2 < 1$.
- *If* $n > Ck \log p$ *then w.h.p. OGP does* **not** *hold for any choice of* $r = r_k$ *and* $0 < \zeta_1 < \zeta_2 < 1$*.(post-COLT)*

An easy **corollary**: if $n < ck \log p$ then any "local-greedy" algorithm will fail w.h.p.

The Overlap Gap Property- The result

Theorem

There exists $C > c > 0$ *such that,*

- *If* $n^* < n < ck \log p$ *then w.h.p. OGP holds for some* $r = r_k$ *and* $0 < \zeta_1 < \zeta_2 < 1$.
- *If* $n > Ck \log p$ *then w.h.p. OGP does* **not** *hold for any choice of* $r = r_k$ *and* $0 < \zeta_1 < \zeta_2 < 1$.*(post-COLT)*

An easy **corollary**: if $n < ck \log p$ then any "local-greedy" algorithm will fail w.h.p.

Also, if $n > Ck \log p$ then the simplest "local-greedy" works!(post-COLT)

Summary

(1) We show that when $\mathsf{n} > (1+\epsilon)\mathsf{n}^*$ for some $\epsilon > 0$, **information exists** to recover β^*.

Summary

(1) We show that when $n > (1+\epsilon)n^*$ for some $\epsilon > 0$, **information exists** to recover β^*.

(2) The performance of the optimal estimator M.L.E. **changes suddenly** w.h.p. when the number of samples crosses the value n^*.

Summary

(1) We show that when $n > (1+\epsilon)n^*$ for some $\epsilon > 0$, **information exists** to recover β^*.

(2) The performance of the optimal estimator M.L.E. **changes suddenly** w.h.p. when the number of samples crosses the value n^*.

(3) We conjecture that the regime $n^* < n < 2k \log p$ is algorithmically hard and we prove **a geometrical phase transition** to provide support for it.

Open Problems

- Can it be proven that assuming $n < (1-\epsilon)n^*$, there is no information to recover any fraction of the support of β^*?
- Can we prove/provide more support that $n^* < n < 2k \log p$ is algorithmically hard? For example, can we find a reduction from the planted clique like in sparse PCA [Berthet, Rigollet '13]?

Open Problems

- Can it be proven that assuming $n < (1-\epsilon)n^*$, there is no information to recover any fraction of the support of β^*?
- Can we prove/provide more support that $n^* < n < 2k \log p$ is algorithmically hard? For example, can we find a reduction from the planted clique like in sparse PCA [Berthet, Rigollet '13]?

Thank you!!

Proof Ideas-1

- Set $d = \min_{\beta \in \{0,1\}^p, \sum_{i=1}^p \beta_i = k} \left(||Y - X\beta||_2 \right)$.

Proof Ideas-1

- Set $d = \min_{\beta \in \{0,1\}^p, \sum_{i=1}^p \beta_i = k} \left(||Y - X\beta||_2 \right)$.
- For any $\ell \in \{0, 1, \ldots, k\}$ set

$$T_\ell = \{\beta \in \{0,1\}^p | \sum_{i=1}^{p} \beta_i = k, \text{Overlap}(\beta) = \ell\}.$$

Proof Ideas-1

- Set $d = \min_{\beta \in \{0,1\}^p, \sum_{i=1}^p \beta_i = k} \left(\|Y - X\beta\|_2 \right)$.
- For any $\ell \in \{0, 1, \ldots, k\}$ set

$$T_\ell = \{\beta \in \{0,1\}^p \big| \sum_{i=1}^p \beta_i = k, \text{Overlap}(\beta) = \ell\}.$$

- Set $d_\ell = \min_{\beta \in T_\ell} \left(\|Y - X\beta\|_2 \right)$. Then $d = \min_{\ell = 0,1,\ldots,k} d_\ell$.

Proof Ideas-2

- We show that w.h.p. for all $\ell = 0, 1, \ldots, k$,

$$d_\ell \sim \sqrt{2k(1-\frac{\ell}{k})+\sigma^2}\exp\left(-\frac{k(1-\frac{\ell}{k})\log p}{n}\right).$$

Proof Ideas-2

- We show that w.h.p. for all $\ell = 0, 1, \ldots, k$,

$$d_\ell \sim \sqrt{2k(1-\frac{\ell}{k}) + \sigma^2} \exp\left(-\frac{k(1-\frac{\ell}{k})\log p}{n}\right).$$

- So, w.h.p. for all $\ell = 0, 1, \ldots, k$,

$$d_\ell \sim f\left(1-\frac{\ell}{k}\right),$$

for $f(\alpha) := \sqrt{2\alpha k + \sigma^2} \exp\left(-\alpha \frac{k \log p}{n}\right), \alpha \in [0, 1]$

Proof Ideas-2

- We show that w.h.p. for all $\ell = 0, 1, \dots, k$,

$$d_\ell \sim \sqrt{2k(1-\frac{\ell}{k}) + \sigma^2} \exp\left(-\frac{k(1-\frac{\ell}{k}) \log p}{n}\right).$$

- So, w.h.p. for all $\ell = 0, 1, \dots, k$,

$$d_\ell \sim f\left(1 - \frac{\ell}{k}\right),$$

for $f(\alpha) := \sqrt{2\alpha k + \sigma^2} \exp\left(-\alpha \frac{k \log p}{n}\right), \alpha \in [0, 1]$

- So w.h.p.

$$d \sim \min_{\ell=0,1,\dots,k} f\left(1 - \frac{\ell}{k}\right) \sim \min_{\alpha \in [0,1]} f(\alpha).$$

Proof Ideas-3

- f is strictly log-concave, so $d \sim \min\left(f(0), f(1)\right)$.

Proof Ideas-3

- f is strictly log-concave, so $d \sim \min(f(0), f(1))$.
- But

$$f(0) > f(1) \Leftrightarrow \sqrt{\sigma^2} > \sqrt{2k + \sigma^2} \exp\left(-\frac{k \log p}{n}\right)$$

$$\Leftrightarrow n > \frac{2k}{\log\left(\frac{2k}{\sigma^2} + 1\right)} \log p.$$

Proof Ideas-3

- f is strictly log-concave, so $d \sim \min(f(0), f(1))$.
- But

$$f(0) > f(1) \Leftrightarrow \sqrt{\sigma^2} > \sqrt{2k + \sigma^2} \exp\left(-\frac{k \log p}{n}\right)$$

$$\Leftrightarrow n > \frac{2k}{\log\left(\frac{2k}{\sigma^2} + 1\right)} \log p.$$

- So the optimization problem changes behavior exactly at

$$n^* := \frac{2k}{\log\left(\frac{2k}{\sigma^2} + 1\right)} \log p.$$

Proof Ideas-3

- f is strictly log-concave, so $d \sim \min(f(0), f(1))$.
- But

$$f(0) > f(1) \Leftrightarrow \sqrt{\sigma^2} > \sqrt{2k+\sigma^2}\exp\left(-\frac{k\log p}{n}\right)$$

$$\Leftrightarrow n > \frac{2k}{\log\left(\frac{2k}{\sigma^2}+1\right)}\log p.$$

- So the optimization problem changes behavior exactly at

$$n^* := \frac{2k}{\log\left(\frac{2k}{\sigma^2}+1\right)}\log p.$$

- Therefore $n > n^*$ iff f is minimized at 1 iff d_ℓ being minimized at 0, which happens iff the optimal vector has full common support with β^*.

Proof Ideas-4

Two pictures behind the phase transition
$(p = 10^9, k = 10, \sigma^2 = 1, n^* = 136)$;

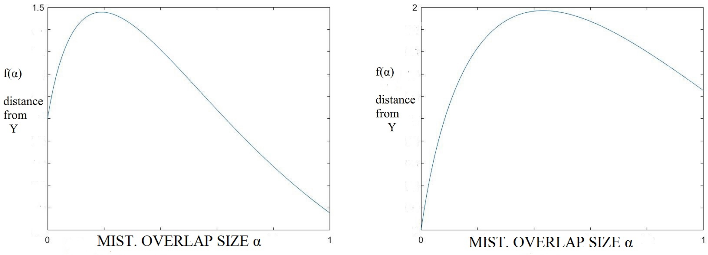

Figure: The behavior of f for $n = 40 < n^*$.

Figure: The behavior of f for $n^* < n = 150$.

Comment: $\alpha := 1 - \frac{\ell}{k}$, so $\alpha = 1$ means no recovery and $\alpha = 0$ full recovery.