Algorithms and Algorithmic Intractability in High Dimensional Linear Regression

Ilias Zadik

Massachusetts Institute of Technology (MIT)

NYU MIC Seminar
2/6/19
Introduction - Big Data Challenges

Over the recent years, the number and magnitude of available datasets have been growing enormously.
Introduction- Big Data Challenges

Over the recent years, the **number** and **magnitude** of available **datasets** have been growing **enormously**.

Big impact across science: 
From **artificial intelligence** to **economics** to **medicine** and many others.
Over the recent years, the number and magnitude of available datasets have been growing enormously.

Big impact across science:
From artificial intelligence to economics to medicine and many others.

Required heavy statistical and computational tools on dealing with issues such as high dimensionality, large noise, missing entries.
Over the recent years, the **number** and **magnitude** of available **datasets** have been growing **enormously**.

Big impact across science:
From **artificial intelligence** to **economics** to **medicine** and many others.

Required heavy **statistical and computational tools** on dealing with issues such as high dimensionality, large noise, missing entries.

Still **many open problems**
even for **simple high dimensional statistical models**!
Overview

This talk

**Algorithms** and **algorithmic barriers**
for high dimensional linear regression.

- Improve **information-theory upper bounds**
  through **tight analysis of MLE**. ("All or Nothing Property")

- Explain **computational-statistical gap**, 
  through **statistical-physics** based methods. ("Overlap Gap Property")

- Offer **new polynomial time algorithm** for noiseless case
  using **lattice basis reduction** ("One Sample Suffices")

*Papers:*

(Gamarnik, Z. *COLT* '17)
(Gamarnik, Z. *Annals of Stats* (major revision) '17+)
(Gamarnik, Z. *NeurIPS* '18)
Outline of the Talk

(1) Introduction
(2) Background in High Dimensional Linear Regression
(3) Information Theory Limits: MLE performance
(4) Computational-Statistical Gap: a statistical-physics perspective
(5) The Noiseless Case: A lattice basis reduction approach
(6) Conclusion
(1) Introduction

(2) **Background in High Dimensional Linear Regression**

(3) Information Theory Limits: MLE performance

(4) Computational-Statistical Gap: a statistical-physics perspective

(5) The Noiseless Case: A lattice basis reduction approach

(6) Conclusion
Linear Regression

Let (unknown) $\beta^* \in \mathbb{R}^p$. $p$ number of features. For data matrix $X \in \mathbb{R}^{n \times p}$, and noise $W \in \mathbb{R}^n$, observe $n$ noisy linear samples of $\beta^*$, $Y = X\beta^* + W$.

Goal: Given $(Y, X)$, recover $\beta^*$. 
Linear Regression

Let (unknown) $\beta^* \in \mathbb{R}^p$. $p$ number of features.

For data matrix $X \in \mathbb{R}^{n \times p}$, and noise $W \in \mathbb{R}^n$,

observe $n$ noisy linear samples of $\beta^*$, $Y = X\beta^* + W$.

Goal: Given $(Y, X)$, recover $\beta^*$.

Simplifying assumption between dependent $Y$ and independent $X$. 
Main Question

Setting: \( Y = X\beta^* + W, \ X \in \mathbb{R}^{n \times p}, \ W \in \mathbb{R}^n. \)

Main Question: Sample Complexity

What is the **minimum** \( n \) so that \( \beta^* \) is (efficiently) recoverable?

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. \( n \geq p \).
Main Question

Setting: \( Y = X\beta^* + W, \; X \in \mathbb{R}^{n \times p}, \; W \in \mathbb{R}^n. \)

Main Question: Sample Complexity

What is the **minimum** \( n \) so that \( \beta^* \) is (efficiently) recoverable?

An immediate answer under full generality: *at least* \( p \).
Main Question

Setting: \( Y = X\beta^* + W, \ X \in \mathbb{R}^{n \times p}, W \in \mathbb{R}^n. \)

Main Question: Sample Complexity

What is the **minimum** \( n \) so that \( \beta^* \) is (efficiently) recoverable?

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. \( n \geq p \).
In many **real-life applications** of Linear Regression (e.g. *computer vision, digital economy, computational biology*) we observe **more** features than samples (i.e. \( n \ll p, p \to +\infty \)).
In many real-life applications of Linear Regression (e.g. computer vision, digital economy, computational biology) we observe more features than samples (i.e. \( n \ll p, p \to +\infty \)).

To be well-posed, need additional assumptions.
Structural Assumptions on $\beta^*$

Assumptions:

(1) $\beta^*$ is \textbf{k-sparse}: k non-zero coordinates, $k = o(p)$.
    (A lot of research, e.g. \textit{Compressed Sensing}.)

(2) $\beta^*$ is \textbf{binary valued}: $\beta^* \in \{0, 1\}^p$. (†)

(†) (non-trivial) \textit{simplification} of \textbf{well-studied} $\beta^*_{\min} := \min_{\beta_i^* \neq 0} |\beta_i^*| = \Theta(1)$
and \textit{support recovery task}.
Structural Assumptions on $\beta^*$

**Assumptions:**

1. $\beta^*$ is **k-sparse**: $k$ non-zero coordinates, $k = o(p)$.
   
   (A lot of research, e.g. *Compressed Sensing*.)

2. $\beta^*$ is **binary valued**: $\beta^* \in \{0, 1\}^p$. (†)

**Main Question: Sample Complexity**

What is the **minimum** $n$ so that $\beta^*$ is (efficiently) recoverable under these assumptions?

(†) (non-trivial) *simplification of well-studied* $\beta_{\min}^* := \min_{\beta_i^* \neq 0} |\beta_i^*| = \Theta(1)$ and support recovery task.
Structural Assumptions on $\beta^*$

Assumptions:

(1) $\beta^*$ is k-sparse: k non-zero coordinates, $k = o(p)$.
   (A lot of research, e.g. Compressed Sensing.)

(2) $\beta^*$ is binary valued: $\beta^* \in \{0, 1\}^p$. (†)

Main Question: Sample Complexity

What is the minimum $n$ so that $\beta^*$ is (efficiently) recoverable under these assumptions?

Assume: $X$ iid $\mathcal{N}(0, 1)$ entries, $W$ iid $\mathcal{N}(0, \sigma^2)$ entries.

(†) (non-trivial) simplification of well-studied $\beta_{\min}^* := \min_{\beta_i^* \neq 0} |\beta_i^*| = \Theta(1)$ and support recovery task.
The Model

Setup

Let \( \beta^* \in \{0, 1\}^p \) be a **binary** \( k \)-sparse vector, \( k = o(p) \). For

- \( X \in \mathbb{R}^{n \times p} \) consisting of i.i.d. \( \mathcal{N}(0, 1) \) entries
- \( W \in \mathbb{R}^n \) consisting of i.i.d. \( \mathcal{N}(0, \sigma^2) \) entries

we get \( n \) **noisy linear samples** of \( \beta^* \), \( Y \in \mathbb{R}^n \), given by,

\[
Y := X\beta^* + W.
\]
The Model

Setup

Let $\beta^* \in \{0, 1\}^p$ be a binary $k$-sparse vector, $k = o(p)$. For

- $X \in \mathbb{R}^{n \times p}$ consisting of i.i.d. $\mathcal{N}(0, 1)$ entries
- $W \in \mathbb{R}^n$ consisting of i.i.d. $\mathcal{N}(0, \sigma^2)$ entries

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

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

Goal

**Minimum** $n$ so that given $(Y, X)$, $\beta^*$ is (efficiently) recoverable with probability tending to 1 as $n, k, p \to +\infty$ (w.h.p.).
Algorithmic Results ([Wainwright ’09],[Fletcher et al ’11])

Set $n_{\text{alg}} = 2k \log p$. Assume $\text{SNR} = \frac{k}{\sigma^2} \to +\infty$.

If

$$n > (1 + \epsilon)n_{\text{alg}}$$

LASSO (convex relaxation) and OMP (greedy algorithm) succeed w.h.p.
**Algorithmic Results** ([Wainwright '09], [Fletcher et al '11])

Set $n_{\text{alg}} = 2k \log p$. Assume $\text{SNR} = \frac{k}{\sigma^2} \to +\infty$.

If

$$n > (1 + \epsilon)n_{\text{alg}}$$

LASSO (*convex relaxation*) and OMP (*greedy algorithm*) succeed w.h.p.

**Information-Theoretic Bounds**

Let $n^* := 2k \log \frac{p}{k} / \log \left( \frac{k}{\sigma^2} + 1 \right)$. Assume $\text{SNR} = \frac{k}{\sigma^2} \to +\infty$. 
A Computational-Statistical Gap

Algorithmic Results ([Wainwright ’09],[Fletcher et al ’11])

Set $n_{\text{alg}} = 2k \log p$. Assume $\text{SNR} = \frac{k}{\sigma^2} \to +\infty$.

If

$$n > (1 + \epsilon)n_{\text{alg}}$$

LASSO (convex relaxation) and OMP (greedy algorithm) succeed w.h.p.

Information-Theoretic Bounds

Let $n^* := 2k \log \frac{p}{k} / \log \left( \frac{k}{\sigma^2} + 1 \right)$. Assume $\text{SNR} = \frac{k}{\sigma^2} \to +\infty$.

- If $n < (1 - \epsilon)n^*$ no algorithm can succeed w.h.p. [Wang et al ’10]
Algorithmic Results ([Wainwright ’09],[Fletcher et al ’11])

Set $n_{\text{alg}} = 2k \log p$. Assume $\text{SNR} = \frac{k}{\sigma^2} \to +\infty$.

If

$$n > (1 + \epsilon)n_{\text{alg}}$$

LASSO (convex relaxation) and OMP (greedy algorithm) succeed w.h.p.

Information-Theoretic Bounds

Let $n^* := 2k \log \frac{p}{k} / \log \left( \frac{k}{\sigma^2} + 1 \right)$. Assume $\text{SNR} = \frac{k}{\sigma^2} \to +\infty$.

- If $n < (1 - \epsilon)n^*$ no algorithm can succeed w.h.p. [Wang et al ’10]
- For some large $C > 0$, if $n \geq Cn^*$, MLE succeeds [Rad’ 11].
Algorithmic Results ([Wainwright '09],[Fletcher et al '11])

Set $n_{\text{alg}} = 2k \log p$. Assume $\text{SNR} = \frac{k}{\sigma^2} \to +\infty$.

If

$$n > (1 + \epsilon)n_{\text{alg}}$$

LASSO (convex relaxation) and OMP (greedy algorithm) succeed w.h.p.

Information-Theoretic Bounds

Let $n^* := 2k \log \left( \frac{p}{k} / \log \left( \frac{k}{\sigma^2} + 1 \right) \right)$. Assume $\text{SNR} = \frac{k}{\sigma^2} \to +\infty$.

- If $n < (1 - \epsilon)n^*$ no algorithm can succeed w.h.p. [Wang et al '10]
- For some large $C > 0$, if $n \geq Cn^*$, MLE succeeds [Rad’ 11].
Questions
(1) Can we find the exact information theoretic bound of the problem?
Contribution: $n^*, \text{in an (asymptotic) strong sense.}$

(2) Is there some fundamental explanation for the apparent computational-statistical gap?
Contributions: Stat physics-based evidence for (landscape) hardness.

If $\sigma = 0$, $\beta^*$ truly binary: gap closes using lattice basis reduction.

Figure: Computational-Statistical Gap
Figure: Computational-Statistical Gap

Questions

(1) Can we find the **exact information theoretic bound** of the problem?

(2) Is there some **fundamental** explanation for the apparent **computational-statistical gap**?
Figure: Computational-Statistical Gap

Questions/Contributions

(1) Can we find the exact information theoretic bound of the problem?  
Contribution: $n^*$, in an (asymptotic) strong sense.

(2) Is there some fundamental explanation for the apparent computational-statistical gap?  
Contributions: Stat physics-based evidence for (landscape) hardness. If $\sigma = 0$, $\beta^*$ truly binary: gap closes using lattice basis reduction.
(1) Introduction
(2) Background in High Dimensional Linear Regression
(3) **Information Theory Limits: MLE performance**
(4) Computational-Statistical Gap: a statistical-physics perspective
(5) The Noiseless Case: A lattice basis reduction approach
(6) Conclusion
Maximum Likelihood Estimator (MLE)

\[ Y = X\beta^* + W \] with \( W \) iid \( N(0, \sigma^2) \) entries.

The MLE

\[ \hat{\beta}_{\text{MLE}} \] is the optimal solution of least-squares

\[
(\text{LS}) : \min_{\beta \in \{0,1\}^p, \|\beta\|_0 = k} \|Y - X\beta\|_2
\]

[Rad '11]: success with \( Cn^* \) samples.
“All or Nothing” Theorem [Gamarnik, Z. ’17]

Definition

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

\[
\text{overlap}(\beta) := |\text{Support}(\beta^*) \cap \text{Support}(\beta)|.
\]
“All or Nothing” Theorem [Gamarnik, Z. ’17]

**Definition**

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

\[
\text{overlap}(\beta) := |\text{Support}(\beta^*) \cap \text{Support}(\beta)|.
\]

**Theorem (“MLE: All or Nothing” (Gamarnik, Z. COLT ’17))**

Let \( \epsilon > 0 \) be arbitrary.

- If \( n > (1 + \epsilon) n^* \), then \( \frac{1}{k} \text{overlap}(\hat{\beta}_{\text{MLE}}) \to 1 \) whp.
- If \( n < (1 - \epsilon) n^* \), \( (\dagger) \) then \( \frac{1}{k} \text{overlap}(\hat{\beta}_{\text{MLE}}) \to 0 \) whp.

\( (\dagger) \ k \leq \exp(\sqrt{\log p}) \)
An “All or Nothing” phase transition!

• With \( n = (1 + \epsilon)n^* \), MLE recovers all but \( o(1) \)-fraction of the support.

• With \( n = (1 - \epsilon)n^* \), MLE recovers at most \( o(1) \)-fraction of the support.

Delicate argument: novel conditional second moment method for the existence of “low overlap” \( \beta \) with “small” \( \| Y - X \beta \|_2 \).

For \( Z = |\{"low-overlap\" \beta : "small\" \| Y - X \beta \|_2\}|\),

\[
P[Z \geq 1] \geq \frac{E[Z]}{E[Z^2]} \text{ (standard 2nd MM)}
\]

We use for \( Y = X \beta^* + W \)

\[
P[Z \geq 1] = \frac{E[W]}{E[Z]} \geq \frac{E[W]}{E[Z^2]} \text{ (conditional 2nd MM)}
\]
An “All or Nothing” phase transition!

- With $n = (1 + \epsilon)n^*$, MLE recovers all but $o(1)$-fraction of the support.
- With $n = (1 - \epsilon)n^*$, MLE recovers at most $o(1)$-fraction of the support.
An “All or Nothing” phase transition!

- With $n = (1 + \epsilon)n^*$, MLE recovers all but $o(1)$-fraction of the support.
- With $n = (1 - \epsilon)n^*$, MLE recovers at most $o(1)$-fraction of the support.
- Delicate argument: novel conditional second moment method for the existence of “low overlap” $\beta$ with “small” $\|Y - X\beta\|_2$.
“All or Nothing Theorem” - Comments

An “All or Nothing” phase transition!

• With $n = (1 + \epsilon)n^*$, MLE recovers all but $o(1)$-fraction of the support.
• With $n = (1 - \epsilon)n^*$, MLE recovers at most $o(1)$-fraction of the support.
• Delicate argument: novel conditional second moment method for the existence of “low overlap” $\beta$ with “small” $\|Y - X\beta\|_2$.

For $Z = |\{“low-overlap” \beta : “small” \|Y - X\beta\|_2\}|$, 

$$\mathbb{P}[Z \geq 1] \geq \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]} \quad (standard \ 2nd \ MM)$$
An “All or Nothing” phase transition!

- With \( n = (1 + \epsilon)n^* \), MLE recovers \textbf{all but} \( o(1) \)-fraction of the support.
- With \( n = (1 - \epsilon)n^* \), MLE recovers \textbf{at most} \( o(1) \)-fraction of the support.
- Delicate argument: \textbf{novel conditional second moment method} for the existence of “low overlap” \( \beta \) with “small” \( \|Y - X\beta\|_2 \).

For \( Z = |\{ \text{“low-overlap” } \beta : \text{“small” } \|Y - X\beta\|_2 \}| \),

\[
P[Z \geq 1] \geq \frac{E[Z]^2}{E[Z^2]} \quad (\text{standard 2nd MM})
\]

We use for \( Y = X\beta^* + W \)

\[
P[Z \geq 1] = E_W[P[Z \geq 1|W]] \geq E_W[\frac{E[Z|W]^2}{E[Z^2|W]}] \quad (\text{conditional 2nd MM})
\]
Sharp Information-Theoretic Limit $n^*$

$(1 + \epsilon)n^*$ samples MLE (asymptotically) succeeds.

$(1 - \epsilon)n^*$ samples MLE strongly fails.
(1) Introduction
(2) Background in High Dimensional Linear Regression
(3) Information Theory Limits: MLE performance
(4) Computational-Statistical Gap: a statistical-physics perspective
(5) The Noiseless Case: A lattice basis reduction approach
(6) Conclusion
Question 2

Is there some fundamental explanation for the apparent computational-statistical gap?
Question 2

Is there some **fundamental** explanation for the apparent **computational-statistical gap**?

**Contribution through Landscape Analysis**

$n_{\text{alg}}$ is a **phase transition point** for certain Overlap Gap Property (OGP) on the space of binary $k$-sparse vectors (origin in *spin glass theory*). **Conjecture computational hardness!**
Computational gaps appear frequently in random environments

(1) *randoms CSPs*,
    such as random-k-SAT (e.g. [MMZ ’05], [ACORT ’11])

(2) *average-case combinatorial opt problems*
    such as max-independent set in ER graphs (e.g. [GS ’17], [RV ’17])
Computational gaps appear frequently in random environments

(1) *randoms CSPs*,
    such as random-k-SAT (e.g. [MMZ '05], [ACORT '11])

(2) *average-case combinatorial opt problems*
    such as max-independent set in ER graphs (e.g. [GS '17], [RV '17])

Between easy and hard regime there is an “abrupt change in the geometry of the space of (near-optimal) solutions” [ACO '08].
Computational gaps appear frequently in random environments

(1) random CSPs, such as random-k-SAT (e.g. [MMZ ’05], [ACORT ’11])

(2) average-case combinatorial opt problems such as max-independent set in ER graphs (e.g. [GS ’17], [RV ’17])

Between easy and hard regime there is an “abrupt change in the geometry of the space of (near-optimal) solutions” [ACO ’08].

(Vague) Strategy of Studying the Geometry

Study realizable overlap sizes between “near-optimal” solutions. Algorithms appear to work as long as there are no gaps in the overlaps.
Computational gaps appear frequently in random environments

(1) *randoms CSPs,*
    such as random-k-SAT (e.g. [MMZ '05], [ACORT '11])

(2) *average-case combinatorial opt problems*
    such as max-independent set in ER graphs (e.g. [GS '17], [RV '17])

Between easy and hard regime there is an “abrupt change in the geometry of the space of (near-optimal) solutions” [ACO '08].

(Vague) Strategy of Studying the Geometry

Study **realizable overlap sizes** between “near-optimal” solutions. Algorithms appear to work as long as there are **no gaps** in the overlaps.

**Overlap Gap Property**, Shattering, Clustering, Free Energy Wells etc
The Overlap Gap Property (OGP) for Linear Regression

“Near-optimal solutions” \( \{ \beta \in \{0, 1\}^p : \| \beta \|_0 = k, \text{ “small” } \| Y - X\beta \|_2 \} \).

Idea:
Study overlaps between \( \beta \) and \( \beta^\ast \).

\[ \text{overlap}(\beta) = |\text{Support}(\beta) \cap \text{Support}(\beta^\ast)|. \]

The OGP (informally)
The set of \( \beta' \)s with “small” \( \| Y - X\beta \|_2 \) partitions in one group where \( \beta \) have low overlap with the ground truth \( \beta^\ast \) and the other group where \( \beta \) have high overlap with the ground truth \( \beta^\ast \).

Ilias Zadik (MIT)
The Overlap Gap Property (OGP) for Linear Regression

“Near-optimal solutions” \( \{ \beta \in \{0, 1\}^p : \|\beta\|_0 = k, \text{ “small” } \|Y - X\beta\|_2 \} \).

Idea: Study overlaps between \( \beta \) and \( \beta^* \).

\[ \text{overlap}(\beta) = |\text{Support}(\beta) \cap \text{Support}(\beta^*)| \].

The OGP (informally)

The set of \( \beta' \)'s with “small” \( \|Y - X\beta\|_2 \) partitions in one group where \( \beta \) have low overlap with the ground truth \( \beta^* \) and the other group where \( \beta \) have high overlap with the ground truth \( \beta^* \).
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)**

The linear regression problem satisfies OGP if there exists $r > 0$ and $0 < \zeta_1 < \zeta_2 < 1$ such that

(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.
OGP Phase Transition at $\Theta(n_{\text{alg}})$

**Theorem (Gamarnik, Z COLT ’17a), (Gamarnik, Z ’17b)**

Suppose $k \leq \exp(\sqrt{\log p})$. There exists $C > 1 > c > 0$ such that,

- If $n < cn_{\text{alg}}$ then w.h.p. OGP holds.
- If $n > Cn_{\text{alg}}$ then w.h.p. OGP does not hold.

**Figure: $n < cn_{\text{alg}}$**

**Figure: $n > Cn_{\text{alg}}$**
OGP Phase Transition at $\Theta(n_{\text{alg}})$

**Theorem (Gamarnik, Z COLT ’17a), (Gamarnik, Z ’17b)**

Suppose $k \leq \exp(\sqrt{\log p})$. There exists $C > 1 > c > 0$ such that,

- If $n < cn_{\text{alg}}$ then w.h.p. OGP holds.
- If $n > Cn_{\text{alg}}$ then w.h.p. OGP does not hold.

OGP coincides with the failure of **convex relaxation** and **compressed sensing** methods!

Figure: $n < cn_{\text{alg}}$

Figure: $n > Cn_{\text{alg}}$
OGP and Local Search

Local Step: $\beta \rightarrow \beta'$ if $d_H(\beta, \beta') = 2$. E.g. \[
\begin{bmatrix}
* \\
0 \\
1 \\
*
\end{bmatrix}
\rightarrow
\begin{bmatrix}
* \\
1 \\
0 \\
*
\end{bmatrix}
\]
OGP and Local Search

Local Step: $\beta \rightarrow \beta'$ if $d_H(\beta, \beta') = 2$. E.g.

$$
\begin{bmatrix}
* \\
0 \\
1 \\
*
\end{bmatrix} \rightarrow
\begin{bmatrix}
* \\
1 \\
0 \\
*
\end{bmatrix}
$$

(LS): $\min_{\beta \in \{0,1\}^p, \|\beta\|_0 = k} \|Y - X\beta\|_2$. 

Under OGP, there are low-overlap local minima in (LS).

If $n < cn_{\text{alg}}$, greedy local-search algorithm fails (worst-case) w.h.p.
OGP and Local Search

Local Step: $\beta \rightarrow \beta'$ if $d_H(\beta, \beta') = 2$. E.g. $\begin{bmatrix} * \\ 0 \\ 1 \\ * \end{bmatrix} \rightarrow \begin{bmatrix} * \\ 1 \\ 0 \\ * \end{bmatrix}$

(LS): $\min_{\beta \in \{0,1\}^p, \|\beta\|_0=k} \|Y - X\beta\|_2$.

Local Search Barrier

Under OGP, there are low-overlap local minima in (LS).
If $n < cn_{\text{alg}}$, greedy local-search algorithm fails (worst-case) w.h.p.
Theorem (Gamarnik, Z '17b)

If \( n > C_{n_{\text{alg}}} \), the only local minimum in \((LS)\) is \( \beta^* \) whp and greedy local search algorithm succeeds in \( O(k/\sigma^2) \) iterations whp.

\[X\beta\]

with medium-overlap \( \beta \)

\[X\beta\]

with high-overlap \( \beta \)

\[X\beta\]

with low-overlap \( \beta \)
Summary of Contribution

Sharp Information-Theoretic Limit $n^*$

$(1 + \epsilon)n^*$ samples MLE (asymptotically) succeeds.

$(1 - \epsilon)n^*$ samples MLE strongly fails.

OGP Phase Transition at $n_{\text{alg}}$

$n < cn_{\text{alg}}$ OGP holds and $n > Cn_{\text{alg}}$ OGP does not hold.

Computational Hardness conjectured!
Outline of the Talk

(1) Introduction
(2) Background in High Dimensional Linear Regression
(3) Information Theory Limits: MLE performance
(4) Computational-Statistical Gap: a statistical-physics perspective
(5) The Noiseless Case: A lattice basis reduction approach
(6) Conclusion
Fact

Under $X \in \mathbb{R}^{n \times p}$ iid $\mathcal{N}(0, 1)$, one samples suffices for $\sigma = 0$. ($n^* = 1$)
Fact

Under $X \in \mathbb{R}^{n \times p}$ iid $\mathcal{N}(0, 1)$, one samples suffices for $\sigma = 0$. ($n^* = 1$)

**Reason:** Recall $y_1 = \langle X_1, \beta^* \rangle$ and no other binary $\beta$ satisfies $y_1 = \langle X_1, \beta \rangle$

For any $\beta \neq \beta^*$ $\mathbb{P}[y_1 = \langle X_1, \beta \rangle] = 0$ (*no sparsity needed.*)
Noiseless Case: One Sample Suffices

Fact

Under $X \in \mathbb{R}^{n \times p}$ iid $\mathcal{N}(0, 1)$, one samples suffices for $\sigma = 0$. ($n^* = 1$)

Reason: Recall $y_1 = \langle X_1, \beta^* \rangle$ and no other binary $\beta$ satisfies $y_1 = \langle X_1, \beta \rangle$

For any $\beta \neq \beta^*$ $\mathbb{P}[y_1 = \langle X_1, \beta \rangle] = 0$ (no sparsity needed.)

Question

Can we make brute-force search efficient?
Fact

Under $X \in \mathbb{R}^{n \times p}$ iid $\mathcal{N}(0, 1)$, one samples suffices for $\sigma = 0$. ($n^* = 1$)

Reason: Recall $y_1 = \langle X_1, \beta^* \rangle$ and no other binary $\beta$ satisfies $y_1 = \langle X_1, \beta \rangle$
For any $\beta \neq \beta^*$ $\mathbb{P} [y_1 = \langle X_1, \beta \rangle] = 0$ (no sparsity needed.)

Question

Can we make brute-force search efficient?

$n_{\text{alg}} = 2k \log p$ and OGP for $n < n_{\text{alg}}$. 
Fact

Under $X \in \mathbb{R}^{n \times p}$ iid $\mathcal{N}(0, 1)$, one sample suffices for $\sigma = 0$. ($n^* = 1$)

**Reason:** Recall $y_1 = \langle X_1, \beta^* \rangle$ and no other binary $\beta$ satisfies $y_1 = \langle X_1, \beta \rangle$

For any $\beta \neq \beta^*$ $\mathbb{P}[y_1 = \langle X_1, \beta \rangle] = 0$ (*no sparsity needed.*)

Question

Can we make brute-force search efficient?

$n_{\text{alg}} = 2k \log p$ and OGP for $n < n_{\text{alg}}$.

Contribution: Beyond the sparsity constraint

Offer an **efficient algorithm**

which recovers any **rational-valued** $\beta^*$ (no-sparsity)

from $n = 1$ **noiseless sample** $y_1 = \langle X_1, \beta^* \rangle$ and $p \rightarrow +\infty$.

*Generalizes to higher $n$ and tolerates small noise.*
Suppose $\beta^*$ has Q-rational entries: $\beta_i^* \in \frac{1}{Q}\mathbb{Z}$.

Theorem ("One Sample Suffices", (Gamarnik, Z. NeurIPS '18))

Assume any $n = o(p)$ samples and $\sigma \leq e^{-p \max\{p, \log Q\}/n}$.

Then there exists a polynomial-in-$n$, $p$, $\log Q$ time algorithm with input $(Y, X)$ outputs $\beta^*$ w.h.p. as $p \to +\infty$. 

Works for any iid (bounded mean) continuous entries on $X$.

The Algorithm: Lattice-Based Method

Reduces to Shortest Vector Problem on a lattice and uses lattice basis reduction technique.

Based on pioneering work [Lagarias, Odlyzko '83], [Frieze '86] on randomly generated subset-sum problems.
Suppose $\beta^*$ has $Q$-rational entries: $\beta_i^* \in \frac{1}{Q}\mathbb{Z}$.

**Theorem ("One Sample Suffices", (Gamarnik, Z. NeurIPS '18))**

Assume any $n = o(p)$ samples and $\sigma \leq e^{-p \max\{p, \log Q\}/n}$.

Then there exists a **polynomial-in**-n, p, log Q **time algorithm** with input $(Y, X)$ outputs $\beta^*$ w.h.p. as $p \to +\infty$.

Works for any iid (bounded mean) continuous entries on $X$. 
Suppose $\beta^*$ has $Q$-rational entries: $\beta^*_i \in \frac{1}{Q}\mathbb{Z}$.

**Theorem ("One Sample Suffices", (Gamarnik, Z. NeurIPS ’18))**

Assume any $n = o(p)$ samples and $\sigma \leq e^{-p \max\{p, \log Q\}/n}$.

Then there exists a polynomial-in-$n, p, \log Q$ time algorithm with input $(Y, X)$ outputs $\beta^*$ w.h.p. as $p \to +\infty$.

Works for any iid (bounded mean) continuous entries on $X$.

**The Algorithm: Lattice-Based Method**

Reduces to **Shortest Vector Problem** on a lattice and uses **lattice basis reduction** technique.

Based on pioneering work [Lagarias, Odlyzko ’83], [Frieze ’86] on *randomly generated subset-sum problems*. 
Lattice produced by matrix $A \in \mathbb{Z}^{d \times d}$: $\mathcal{L} = \{Aw : w \in \mathbb{Z}^d\}$.
Lattices

- **Lattice** produced by matrix $A \in \mathbb{Z}^{d \times d}$: $\mathcal{L} = \{Aw : w \in \mathbb{Z}^d\}$.

- **Shortest Vector Problem**: $\min \|z\|_2 : z \in \mathcal{L} \setminus \{0\}$, say optimum $z_{SV}$.

**Shortest Vector Problem (SVP):** given a lattice, find a shortest (nonzero) vector.
Lattices

- **Lattice** produced by matrix $A \in \mathbb{Z}^{d \times d}$: $\mathcal{L} = \{Aw : w \in \mathbb{Z}^d\}$.

- **Shortest Vector Problem**: $\min \|z\|_2 : z \in \mathcal{L} \setminus \{0\}$, say optimum $z_{SV}$.

- **NP-hard**, but **Lenstra-Lenstra-Lovász** efficiently approximates it, outputs $\hat{z} \in \mathcal{L} \setminus \{0\}$ with $\|\hat{z}\|_2 \leq 2^{d/2}\|z_{SV}\|_2$.

**Shortest Vector Problem (SVP):** given a lattice, find a shortest (nonzero) vector.
**Main Idea (High Level)**

Construct lattice $\mathcal{L}(Y, X)$ where

- **shortest vector** is $\beta^*$
- **all “approximately” short** vectors are multiples of $\beta^*$.

Use **Lenstra-Lenstra-Lovász** to recover $\beta^*$.
Outline of the Talk

(1) Introduction
(2) Background in High Dimensional Linear Regression
(3) Information Theory Limits: MLE performance
(4) Computational-Statistical Gap: a statistical-physics perspective
(5) The Noiseless Case: A lattice basis reduction approach
(6) Conclusion
Conclusion - Overview

This talk

**Algorithms and algorithmic barriers**
for *high dimensional linear regression*.

- Improve **information-theory upper bounds**
  through tight analysis of MLE. ("All or Nothing Property")

- Explain **computational-statistical gap**, 
  through *statistical-physics* based methods. ("Overlap Gap Property")

- Offer new **polynomial time algorithm** for noiseless case 
  using *lattice basis reduction* ("One Sample Suffices")

*Papers:*
(Gamarnik, Z. *COLT* '17)
(Gamarnik, Z. *Annals of Stats* (major revision) '17+)
(Gamarnik, Z. *NeurlIPS* '18)
(1) How fundamental is the “All-or-Nothing” Property?
Does it appear in other settings?
Ongoing work with Jiaming Xu and Galen Reeves.
Conclusion - Future Directions

(1) How fundamental is the “All-or-Nothing” Property? Does it appear in other settings? Ongoing work with Jiaming Xu and Galen Reeves.

(2) OGP framework for computational-statistical hardness. Does it work for e.g. planted clique? Relation to SOS hierarchy/average-case reductions?
(1) How fundamental is the “All-or-Nothing” Property? Does it appear in other settings? Ongoing work with Jiaming Xu and Galen Reeves.

(2) OGP framework for computational-statistical hardness. Does it work for e.g. planted clique? Relation to SOS hierarchy/average-case reductions?

(3) Study power of lattice-based methods for regression (instead of convex relaxation, message passing, greedy methods)? Can they generalize to real coefficients/higher noise level?
(1) How fundamental is the "All-or-Nothing" Property? Does it appear in other settings? Ongoing work with Jiaming Xu and Galen Reeves.

(2) OGP framework for computational-statistical hardness. Does it work for e.g. planted clique? Relation to SOS hierarchy/average-case reductions?

(3) Study power of \textbf{lattice-based methods} for regression (instead of convex relaxation, message passing, greedy methods)? Can they generalize to real coefficients/higher noise level?

Thank you!!
The Algorithm (special case, [F ’84])

Assume

• $n = 1$, $\sigma = 0$, $\beta^*$ binary: $y = \langle X_1, \beta^* \rangle$.
• $X_1 \in \mathbb{Z}^p$ with iid uniform in $[2^N]$ entries for large $N$ (say $N = p^2$).
The Algorithm (special case, [F ’84])

Assume

- \( n = 1, \sigma = 0, \beta^* \) binary: \( y = \langle X_1, \beta^* \rangle \).
- \( X_1 \in \mathbb{Z}^p \) with iid uniform in \([2^N]\) entries for large \( N \) (say \( N = p^2 \)).

(1) For \( M \) sufficiently large enough set \( \mathcal{L}_M(y_1, X_1) \) produced by the columns of

\[
A_M := \begin{bmatrix}
MX_1 & -My_1 \\
I_{p \times p} & 0
\end{bmatrix}
\]
The Algorithm (special case, [F '84])

Assume

- \( n = 1, \sigma = 0, \beta^* \) binary: \( y = \langle X_1, \beta^* \rangle \).
- \( X_1 \in \mathbb{Z}^p \) with iid **uniform in** \([2^N]\) **entries** for large \( N \) (say \( N = p^2 \)).

(1) For \( M \) sufficiently large enough set \( \mathcal{L}_M(y_1, X_1) \) produced by the columns of

\[
A_M := \begin{bmatrix} MX_1 & -My_1 \\ I_{p \times p} & 0 \end{bmatrix}
\]

**Lemma:** Each \( z \in \mathcal{L}_M, \|z\|_2 < M \) is a multiple of \( \begin{bmatrix} 0 \\ \beta^* \end{bmatrix} \), w.h.p. \((N \text{ large})\)
The Algorithm (special case, [F ’84])

Assume

• $n = 1$, $\sigma = 0$, $\beta^*$ binary: $y = \langle X_1, \beta^* \rangle$.
• $X_1 \in \mathbb{Z}^p$ with iid uniform in $[2^N]$ entries for large $N$ (say $N = p^2$).

(1) For $M$ sufficiently large enough set $\mathcal{L}_M(y_1, X_1)$ produced by the columns of

$$A_M := \begin{bmatrix} MX_1 & -My_1 \\ l_{p \times p} & 0 \end{bmatrix}$$

Lemma: Each $z \in \mathcal{L}_M$, $\|z\|_2 < M$ is a multiple of $\begin{bmatrix} 0 \\ \beta^* \end{bmatrix}$, w.h.p. (N large)

Intuition:

$$z = A_M \begin{bmatrix} \beta \\ \lambda \end{bmatrix} = M\langle X_1, \beta \rangle - M\lambda y_1 = M\langle X_1, \beta - \lambda \beta^* \rangle$$

$$\mathbb{P}(\text{Lemma is false}) \leq \mathbb{P}(\exists \beta \neq \lambda \beta^* : \|\beta\|_2 < M, \langle X_1, \beta - \lambda \beta^* \rangle = 0) \rightarrow 0.$$
“All or Nothing” Theorem [Gamarnik, Z. ’17]

Definition

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

$$\text{overlap}(\beta) := |\text{Support}(\beta^*) \cap \text{Support}(\beta)|.$$
“All or Nothing” Theorem [Gamarnik, Z. ’17]

Definition

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

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

Theorem ("All or Nothing" (Gamarnik, Z. COLT ’17))

Let $\epsilon > 0$ be arbitrary.

- If $n > (1 + \epsilon)n^*$, then $\frac{1}{k}\text{overlap}(\hat{\beta}_{\text{MLE}}) \to 1$ whp.
- If $n < (1 - \epsilon)n^*$, (†) then $\frac{1}{k}\text{overlap}(\hat{\beta}_{\text{MLE}}) \to 0$ whp.

(†) $k \leq \exp(\sqrt{\log p})$
• Set $\text{OPT} = \min_{\beta \in \{0,1\}^p, \|\beta\|_0 = k} (\|Y - X\beta\|_2)$. 

• For any $\ell \in \{0,1,\ldots,k\}$ set $T_\ell = \{\beta \in \{0,1\}^p \mid \|\beta\|_0 = k, \text{overlap} (\beta) = \ell\}$. 

• Set $\text{OPT}_\ell = \min_{\beta \in T_\ell} (\|Y - X\beta\|_2)$. Then $\text{OPT} = \min_{\ell = 0,1,\ldots,k} \text{OPT}_\ell$. 
Proof Ideas-1

- Set \( \text{OPT} = \min_{\beta \in \{0,1\}^p, \|\beta\|_0 = k} (\|Y - X\beta\|_2) \).

- For any \( \ell \in \{0, 1, \ldots, k\} \) set

\[
\mathcal{T}_\ell = \{ \beta \in \{0, 1\}^p \big| \|\beta\|_0 = k, \text{overlap}(\beta) = \ell \}.
\]
Proof Ideas-1

- Set \( \text{OPT} = \min_{\beta \in \{0,1\}^p, \|\beta\|_0 = k} (\|Y - X\beta\|_2) \).

- For any \( \ell \in \{0, 1, \ldots, k\} \) set

  \[ \mathcal{T}_\ell = \{ \beta \in \{0, 1\}^p \|\beta\|_0 = k, \text{overlap}(\beta) = \ell \} \].

- Set \( \text{OPT}_\ell = \min_{\beta \in \mathcal{T}_\ell} (\|Y - X\beta\|_2) \). Then \( \text{OPT} = \min_{\ell = 0, 1, \ldots, k} \text{OPT}_\ell \).
• We show that w.h.p. for all $\ell = 0, 1, \ldots, k$,

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

(requires novel conditional second moment method)
Proof Ideas-2

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

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

(requires novel conditional second moment method)

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

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

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

Proof Ideas-3

- So w.h.p. for $\alpha = 1 - \frac{\ell}{k}$ (false detection rate - FDR),

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

- $f(\alpha) := \sqrt{2\alpha k + \sigma^2} \exp(-\alpha k \log p_n)$ is strictly log-concave.

- $\text{OPT} \sim \min(f(0), f(1))$. But $f(0) > f(1) \iff \sqrt{\sigma^2} > \sqrt{2k + \sigma^2 \exp(-k \log p_n)} \iff n^* := 2k \log \left( \frac{2k \sigma^2 + 1}{\sigma^2} \right) > n.$

- "All or Nothing Phase Transition": $n < n^*$ full FDR or zero overlap but $n > n^*$ zero FDR or full overlap.
Proof Ideas-3

• So w.h.p. for $\alpha = 1 - \frac{\ell}{k}$ (false detection rate - FDR),

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

• $f(\alpha) := \sqrt{2\alpha k + \sigma^2 \exp \left( -\alpha \frac{k \log p}{n} \right)}$ is strictly log-concave.

• "All or Nothing Phase Transition": $n < n^{\ast}$ full FDR or zero overlap but $n > n^{\ast}$ zero FDR or full overlap.
Proof Ideas-3

• So w.h.p. for $\alpha = 1 - \frac{\ell}{k}$ (false detection rate - FDR)

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

• $f(\alpha) := \sqrt{2\alpha k + \sigma^2} \exp \left( -\alpha \frac{k \log p}{n} \right)$ is strictly log-concave.

• OPT $\sim \min (f(0), f(1))$. But

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

$$\iff n^* := \frac{2k}{\log \left( \frac{2k}{\sigma^2} + 1 \right)} \log p > n.$$
Proof Ideas-3

• So w.h.p. for $\alpha = 1 - \frac{\ell}{k}$ (false detection rate - FDR),

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

• $f (\alpha) := \sqrt{2\alpha k + \sigma^2} \exp \left( -\alpha \frac{k \log p}{n} \right)$ is strictly log-concave.

• $\text{OPT } \sim \min (f(0), f(1))$. But

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

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

• “All or Nothing Phase Transition”:

$n < n^*$ full FDR or zero overlap
but $n > n^*$ zero FDR or full overlap.
OGP curve

Figure: OGP

Figure: no-OGP

Figure: OGP

Figure: no-OGP