The All-or-Nothing Phenomenon in Sparse Linear Regression

Ilias Zadik\textsuperscript{1,2},
joint work with Galen Reeves\textsuperscript{3}, Jiaming Xu\textsuperscript{3}

\textsuperscript{1}Massachusetts Institute of Technology \rightarrow \textsuperscript{2} New York University,
\textsuperscript{3}Duke University

32th Conference on Learning Theory (COLT), 2019
Model: Sparse Linear Regression

For

- *(unknown) vector* $\beta \in \mathbb{R}^p$, *with* $\beta \sim \text{Unif}\{v \in \{0, 1\}^p : \|v\|_0 = k\}$
- *data matrix* $X \in \mathbb{R}^{n \times p}$ *with* i.i.d. $\mathcal{N}(0, 1)$ entries
- *noise* $W \in \mathbb{R}^n$ *with* i.i.d. $\mathcal{N}(0, \sigma^2)$ entries

observe $n$ noisy linear samples of $\beta$,

$$Y = X\beta + W.$$ 

**Goal:** Minimum $n = n(p, k, \sigma^2)$ so that $\beta$ can be recovered by $(Y, X)$. 

[GV’02], [AS+’10], [RP’16], [BD+’16], [SC’17], [GZ’17]...
Model: Sparse Linear Regression

For

- (unknown) vector $\beta \in \mathbb{R}^p$, with $\beta \sim \text{Unif}\{v \in \{0, 1\}^p : \|v\|_0 = k\}$
- data matrix $X \in \mathbb{R}^{n \times p}$ with i.i.d. $\mathcal{N}(0, 1)$ entries
- noise $W \in \mathbb{R}^n$ with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries

observe $n$ noisy linear samples of $\beta$,

$$Y = X\beta + W.$$ 

**Goal:** Minimum $n = n(p, k, \sigma^2)$ so that $\beta$ can be recovered by $(Y, X)$.

$$\text{MMSE} = \min_{\hat{\beta} = \hat{\beta}(Y, X)} \frac{1}{k} \mathbb{E} \left[ \|\hat{\beta} - \beta\|_2^2 \right] \in [0, 1]$$

[GV’02], [AS+’10], [RP’16], [BD+’16], [SC’17], [GZ’17]...
Model: Sparse Linear Regression

For

- **(unknown) vector** \( \beta \in \mathbb{R}^p \), with \( \beta \sim \text{Unif}\{v \in \{0, 1\}^p : \|v\|_0 = k\} \)
- **data matrix** \( X \in \mathbb{R}^{n \times p} \) with i.i.d. \( \mathcal{N}(0, 1) \) entries
- **noise** \( W \in \mathbb{R}^n \) with i.i.d. \( \mathcal{N}(0, \sigma^2) \) entries

**observe** \( n \) noisy linear samples of \( \beta \),

\[ Y = X\beta + W. \]

**Goal:** Minimum \( n = n(p, k, \sigma^2) \) so that \( \beta \) can be **recovered** by \( (Y, X) \).

**MMSE** = \( \min_{\hat{\beta} = \hat{\beta}(Y, X)} \frac{1}{k} \mathbb{E} \left[ \|\hat{\beta} - \beta\|_2^2 \right] \in [0, 1] \)

**Weak Recovery:** \( \limsup_{p \to +\infty} \text{MMSE} < 1 \). For which \( n \)?

**Strong Recovery:** \( \lim_{p \to +\infty} \text{MMSE} = 0 \). For which \( n \)?

[GV’02], [AS+’10], [RP’16], [BD+’16], [SC’17], [GZ’17]...
For sublinear sparsity $k \leq \sqrt{p}$ and high SNR $k/\sigma^2$, we identify a critical sample size $n^* = n^*(p, k, \sigma^2)$ for which:

- $n < n^*$ weak recovery is impossible,
- $n > n^*$ strong recovery is possible!
Contribution: All-or-Nothing Phase Transition

For sublinear sparsity \( k \leq \sqrt{p} \) and high SNR \( k/\sigma^2 \), we identify a critical sample size \( n^* = n^*(p, k, \sigma^2) \) for which:

- \( n < n^* \) weak recovery is impossible,
- \( n > n^* \) strong recovery is possible!
Contribution: All-or-Nothing Phase Transition

For **sublinear sparsity** $k \leq \sqrt{p}$ and **high SNR** $k/\sigma^2$, we **identify** a **critical sample size** $n^* = n^*(p, k, \sigma^2)$ for which: $n < n^*$ **weak recovery is impossible**, $n > n^*$ **strong recovery is possible**!
For sublinear sparsity $k \leq \sqrt{p}$ and high SNR $k/\sigma^2$, we identify a critical sample size $n^* = n^*(p, k, \sigma^2)$ for which: $n < n^*$ weak recovery is impossible, $n > n^*$ strong recovery is possible!
All-or-Nothing: Theorem

\[ n^* = \frac{2k \log (p/k)}{\log \left( \frac{k}{\sigma^2} + 1 \right)} \]

Theorem (All-or-Nothing Phenomenon)

For any \( \epsilon, \delta > 0 \) if \( k \leq p^{1/2-\delta} \) and \( k/\sigma^2 \geq C(\delta, \epsilon) > 0 \) then, if

- \( n > (1 + \epsilon) n^* \), \( \lim_p \text{MMSE} = 0 \). (strong recovery possible!)
- \( n < (1 - \epsilon) n^* \), \( \lim_p \text{MMSE} = 1 \). (weak recovery impossible!)
All-or-Nothing: Theorem

\[ n^* = \frac{2k \log (p/k)}{\log (k/\sigma^2 + 1)} \]

Theorem (All-or-Nothing Phenomenon)

For any \( \epsilon, \delta > 0 \) if \( k \leq p^{1/2 - \delta} \) and \( k/\sigma^2 \geq C(\delta, \epsilon) > 0 \) then, if

- \( n > (1 + \epsilon) n^* \), \( \lim_p \text{MMSE} = 0 \). (*strong recovery possible!*)
- \( n < (1 - \epsilon) n^* \), \( \lim_p \text{MMSE} = 1 \). (*weak recovery impossible!*)

Prior results for \( n \geq Cn^* \) [R’11] or \( n = o(n^*) \) [WW ’10, ASZ’10, SC’17].

All-or-nothing (MLE) if \( k < e^{\sqrt{\log p}} \) [GZ’17].
All or Nothing Theorem - Proof Sketch

Negative Result for $n \leq (1 - \epsilon)n^*$: $\lim_{p} \text{MMSE} = 1$.

- **Step 1:**
  “Impossibility of Testing”: Data Look Like Pure Noise.

  Let $P$ the law of $(Y = X\beta + W, X)$, and $Q$ the law of $(Y = \lambda W, X)$ for $\lambda = \sqrt{k/\sigma^2} + 1$.

  We show, $\lim_{p \to +\infty} D_{KL}(P||Q) = 0$.

  Requires conditional second moment method.

- **Step 2:**
  “Impossibility of Testing” implies “Impossibility of Estimation”.

  We show the general (any $n, p, k$ and any $\beta$: $\|\beta\|_2 = k$):

  $$1 - \text{MMSE} \leq 2 \left(1 + \frac{\sigma^2}{k}\right) D_{KL}(P||Q) .$$
All or Nothing Theorem - Proof Sketch

Negative Result for $n \leq (1 - \epsilon)n^*$: $\lim_{p} \text{MMSE} = 1$.

• **Step 1:**
  “Impossibility of Testing”: Data Look Like Pure Noise.
  Let $P$ the law of $(Y = X\beta + W, X)$, and $Q$ the law of $(Y = \lambda W, X)$ for $\lambda = \sqrt{k/\sigma^2 + 1}$.

  We show, $\lim_{p} D_{KL}(P || Q) = 0$.

  Requires conditional second moment method.

• **Step 2:**
  “Impossibility of Testing” implies “Impossibility of Estimation”.
  We show the general (any $n, p, k$ and any $\beta$): $1 - \text{MMSE} \leq 2 \left(1 + \frac{\sigma^2}{k}\right) D_{KL}(P || Q)$.
All or Nothing Theorem - Proof Sketch

Negative Result for \( n \leq (1 - \epsilon)n^* \): \( \lim_{p} \text{MMSE} = 1 \).

- **Step 1:**
  “Impossibility of Testing”: Data Look Like Pure Noise.
  Let \( P \) the law of \((Y = X_\beta + W, X)\),
  and \( Q \) the law of \((Y = \lambda W, X)\) for \( \lambda = \sqrt{k/\sigma^2} + 1 \).
  We show,
  \[
  \lim_{p \to +\infty} D_{\text{KL}}(P || Q) = 0.
  \]
  Requires *conditional* second moment method.

Reeves, Xu, Zadik (Duke, MIT, NYU)
All or Nothing Theorem - Proof Sketch

Negative Result for \( n \leq (1 - \epsilon)n^* \): \( \lim_{p} \text{MMSE} = 1 \).

• **Step 1:**
  “Impossibility of Testing”: Data Look Like Pure Noise.
  Let \( P \) the law of \( (Y = X\beta + W, X) \),
  and \( Q \) the law of \( (Y = \lambda W, X) \) for \( \lambda = \sqrt{k/\sigma^2 + 1} \).
  We show,
  \[
  \lim_{p \to +\infty} D_{KL} (P\|Q) = 0.
  \]
  Requires *conditional* second moment method.

• **Step 2:**
  “Impossibility of Testing” implies “Impossibility of Estimation”.
All or Nothing Theorem - Proof Sketch

Negative Result for \( n \leq (1 - \epsilon)n^* \): \( \lim_{p} \text{MMSE} = 1. \)

- **Step 1:**
  “Impossibility of Testing”: Data Look Like Pure Noise.
  Let \( P \) the law of \( (Y = X\beta + W, X) \),
  and \( Q \) the law of \( (Y = \lambda W, X) \) for \( \lambda = \sqrt{k/\sigma^2 + 1} \).
  We show,
  \[
  \lim_{p \to +\infty} D_{KL}(P || Q) = 0.
  \]
  Requires *conditional* second moment method.

- **Step 2:**
  “Impossibility of Testing” implies “Impossibility of Estimation”.
  We show the general (any \( n, p, k \) and any \( \beta : \|\beta\|_2 = k \)):
  \[
  1 - \text{MMSE} \leq 2 \left(1 + \sigma^2/k\right) D_{KL}(P || Q).
  \]
Conclusion

All-or-Nothing Phenomenon: $k < \sqrt{p}$, high SNR

- When $n > (1 + \epsilon) n^*$, strong recovery is possible!
- When $n < (1 - \epsilon) n^*$, weak recovery is impossible!

Come to the poster 166 for:

- **Interpretation** of $n^*$ with *Gaussian communication channel analogy*
  \[
  n^* \approx \log \binom{p}{k} / 0.5 \log \left( k/\sigma^2 + 1 \right) \]
  - entropy of $\beta$
  - Gaussian Channel Capacity

- **Intuition** from *replica-symmetric results* in the regime $k = \Theta(p)$.

- **Proof ideas** (*conditional second moment method* and *area theorem*)
Conclusion

All-or-Nothing Phenomenon: $k < \sqrt{p}$, high SNR

- When $n > (1 + \epsilon) n^*$, strong recovery is possible!
- When $n < (1 - \epsilon) n^*$, weak recovery is impossible!

Come to the poster 166 for:

- **Interpretation** of $n^*$ with Gaussian communication channel analogy
  \[
  n^* \approx \log \left( \frac{p}{k} \right) / 0.5 \log \left( \frac{k}{\sigma^2 + 1} \right).
  \]
  - entropy of $\beta$
  - Gaussian Channel Capacity

- **Intuition** from replica-symmetric results in the regime $k = \Theta(p)$.

- **Proof ideas** (*conditional second moment method* and *area theorem*)

Thank you!!

Reeves, Xu, Zadik (Duke, MIT, NYU)  All-or-Nothing Phenomenon in Regression