The All-or-Nothing Phenomenon in Sparse Linear Regression

Ilias Zadik[1,2],
joint work with Galen Reeves[3], Jiaming Xu[3]

[1]Massachusetts Institute of Technology → [2] New York University,
[3]Duke University

32th Conference on Learning Theory (COLT), 2019

Model: Sparse Linear Regression

For

- **(unknown) vector** $\beta \in \mathbb{R}^p$, *with* $\beta \sim \mathrm{Unif}\{v \in \{0,1\}^p : \|v\|_0 = k\}$
- **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. $\mathcal{N}\left(0, \sigma^2\right)$ entries

observe n noisy linear samples of β,

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

Goal: Minimum $n = n(p, k, \sigma^2)$ so that β 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* $\beta \sim \mathrm{Unif}\{v \in \{0,1\}^p : \|v\|_0 = k\}$
- **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. $\mathcal{N}\left(0,\sigma^2\right)$ entries

observe n noisy linear samples of β,

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

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

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

[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* $\beta \sim \mathrm{Unif}\{v \in \{0,1\}^p : \|v\|_0 = k\}$
- **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. $\mathcal{N}\left(0,\sigma^2\right)$ entries

observe n noisy linear samples of β,

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

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

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

Weak Recovery: $\limsup_{p \to +\infty} \mathrm{MMSE} < 1$. For which n?
Strong Recovery: $\lim_{p \to +\infty} \mathrm{MMSE} = 0$. For which n?

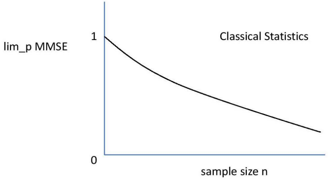
[GV'02], [AS+'10], [RP' 16], [BD+ '16], [SC' 17], [GZ' 17]...

Contribution: All-or-Nothing Phase Transition

For **sublinear sparsity** $k \leq \sqrt{p}$ and **high SNR** k/σ^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**!

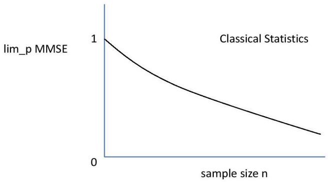
Contribution: All-or-Nothing Phase Transition

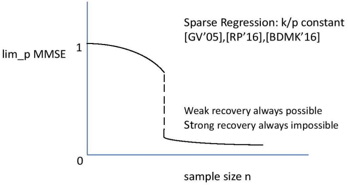
For **sublinear sparsity** $k \leq \sqrt{p}$ and **high SNR** k/σ^2,
we **identify** a *critical sample size* $n^* = n^*(p, k, \sigma^2)$ for which:
$n < n^*$ *weak recovery* is **impossible**, $n > n^*$ *strong recovery* is **possible**!



Contribution: All-or-Nothing Phase Transition

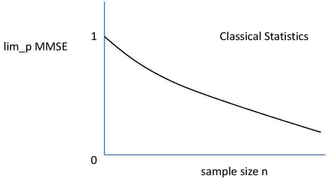
For **sublinear sparsity** $k \leq \sqrt{p}$ and **high SNR** k/σ^2,
we **identify** a *critical sample size* $n^* = n^*(p, k, \sigma^2)$ for which:
$n < n^*$ *weak recovery* is **impossible**, $n > n^*$ *strong recovery* is **possible**!

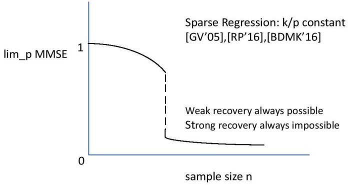


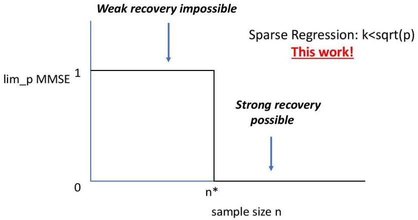


Contribution: All-or-Nothing Phase Transition

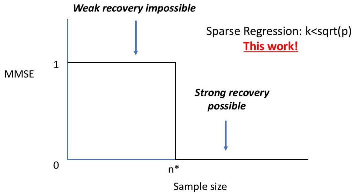
For **sublinear sparsity** $k \leq \sqrt{p}$ and **high SNR** k/σ^2,
we **identify** a *critical sample size* $n^* = n^*(p, k, \sigma^2)$ for which:
$n < n^*$ *weak recovery* is **impossible**, $n > n^*$ *strong recovery* is **possible**!







All-or-Nothing: Theorem



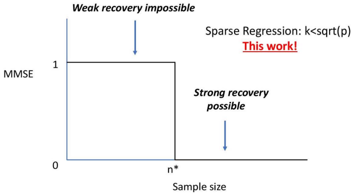
$$n^* = 2k \log(p/k) \,/\, \log\left(k/\sigma^2 + 1\right)$$

Theorem (All-or-Nothing Phenomenon)

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

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

All-or-Nothing: Theorem



$$n^* = 2k \log\left(p/k\right) / \log\left(k/\sigma^2 + 1\right)$$

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*

- $n > (1+\epsilon)\, n^*$, $\lim_p \mathrm{MMSE} = 0$. *(strong recovery possible!)*
- $n < (1-\epsilon)\, n^*$, $\lim_p \mathrm{MMSE} = 1$. *(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 $k < e^{\sqrt{\log p}}$ [G**Z**'17].

All or Nothing Theorem - Proof Sketch

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

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

All or Nothing Theorem - Proof Sketch

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

- *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 $\lambda = \sqrt{k/\sigma^2 + 1}$.

All or Nothing Theorem - Proof Sketch

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

- *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.

All or Nothing Theorem - Proof Sketch

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

- *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 $n \leq (1-\epsilon)n^*$: $\lim_p \mathrm{MMSE} = 1$.

- *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 $\beta : \|\beta\|_2 = k$):

  $$1 - \mathrm{MMSE} \leq 2\left(1 + \sigma^2/k\right) D_{KL}(P||Q).$$

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*

$$n^* \approx \underbrace{\log \binom{p}{k}}_{\text{entropy of } \beta} \Big/ \underbrace{0.5 \log \left(k/\sigma^2 + 1 \right)}_{\text{Gaussian Channel Capacity}} .$$

- **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*

$$n^* \approx \underbrace{\log \binom{p}{k}}_{\text{entropy of } \beta} \;/\; \underbrace{0.5 \log \left(k/\sigma^2 + 1\right)}_{\text{Gaussian Channel Capacity}} .$$

- **Intuition** from *replica-symmetric results* in the regime $k = \Theta(p)$.
- **Proof ideas** (*conditional second moment method* and *area theorem*)

Thank you!!