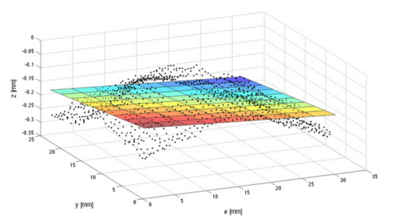
The All-or-Nothing Phenomenon in Sparse Linear Regression

Galen Reeves[1] & Jiaming Xu[1] & Ilias Zadik[2]

[1]Duke University & [2]Massachusetts Institute of Technology

Fitting linear models in high dimensional data has been the focus on many applications, from genomics to MRI to economics.



The Sparse Linear Regression Model

Setup: Let sample size $n \in \mathbb{N}$, feature size $p \in \mathbb{N}$, sparsity $k \in \mathbb{N}$ **with** $k = o(p)$ and $\sigma^2 > 0$. Assume:

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

We make n noisy linear observations of β:

$$Y = X\beta + W$$

Task: Recover β from data (Y, X).

Performance Metric:
Focus on $\hat{\beta} = \hat{\beta}(Y, X)$, with small Mean Squared Error (MSE):

$$\mathrm{MSE}\left(\hat{\beta}\right) := \mathbb{E}\left[\|\hat{\beta} - \beta\|_2^2\right].$$

Performance of Random Guess:
For $\hat{\beta} = \mathbb{E}[\beta] = \frac{k}{p}(1, 1, \ldots, 1)^\top$, $\mathrm{MSE}_0 = \mathbb{E}\left[\|\beta - \mathbb{E}[\beta]\|_2^2\right] = k\left(1 - \frac{k}{p}\right)$.

Definition 1 (Strong and weak recovery) *We say that $\hat{\beta} = \hat{\beta}(Y, X) \in \mathbb{R}^p$ achieves*

- *strong recovery if* $\limsup_{p \to \infty} \mathrm{MSE}\left(\hat{\beta}\right)/\mathrm{MSE}_0 = 0$*;*
- *weak recovery if* $\limsup_{p \to \infty} \mathrm{MSE}\left(\hat{\beta}\right)/\mathrm{MSE}_0 < 1$.

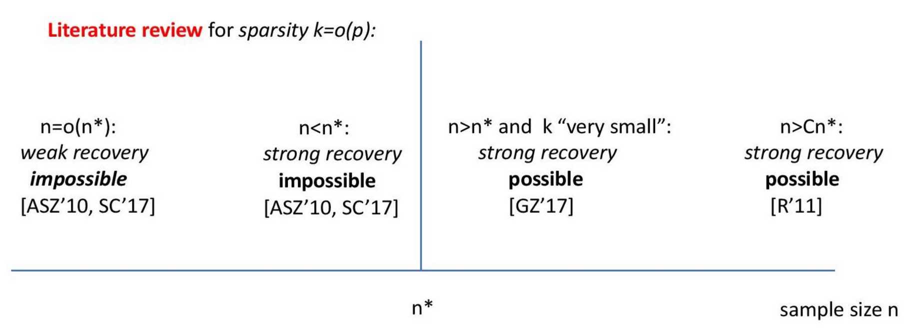
Asymptotics: $n = n_p, k = k_p, \sigma = \sigma_p$ and $p \to +\infty$. Also $k = o(p)$ and $\mathrm{SNR} = k/\sigma^2 = \Omega(1)$.

This Work

We identify a sample size $n^* = n^*(p, k, \sigma^2)$ for which:
if $n < n^*$ *weak recovery* is **impossible**, but if $n > n^*$ *strong recovery* is **possible**!
An All-or-Nothing phase transition!

Literature Review

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



Positive Results:

- (Rad '11): Under $\sigma^2 = \Theta(1)$, $k \to +\infty$,
  if $n \geq Cn^*$ for sufficiently large $C > 0$: *strong recovery possible* (MLE).
- (Gamarnik, Zadik '17): Under $k/\sigma^2 \to +\infty$ and $k \leq e^{\sqrt{\log p}}$,
  if $n \geq (1+\epsilon)n^*$ for any $\epsilon > 0$: *strong recovery possible* (MLE).

Negative Results: (Aeron, Saligrama, Zhao '10), (Scarlett, Cevher '17):

- If $n = o(n^*)$: *weak recovery impossible.*
- If $n \leq (1-\epsilon)n^*$ for any $\epsilon > 0$: *strong recovery impossible.*

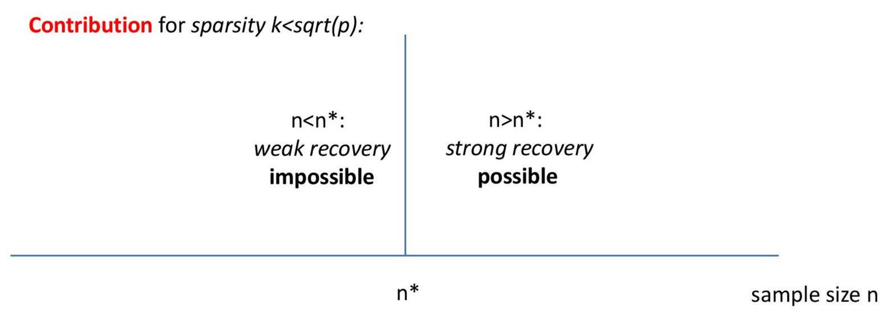
Information-theoretic Importance of n^*

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

- *Encoding-decoding* scheme of $\beta \in \{0,1\}^p$ with $\|\beta\|_0 = k$ from $Y = X\beta + W \in \mathbb{R}^n$.
- Capacity *achieved*: $\log\binom{p}{k}/n$.
  Capacity of *the Gaussian Channel:* $0.5\log(\mathrm{SNR} + 1) = 0.5\log\left(k/\sigma^2 + 1\right)$.

Hence, for *strong recovery of β from (Y, X):* $\log\binom{p}{k}/n \leq 0.5\log\left(k/\sigma^2 + 1\right)$ or $n^* \leq n$.

Main Result

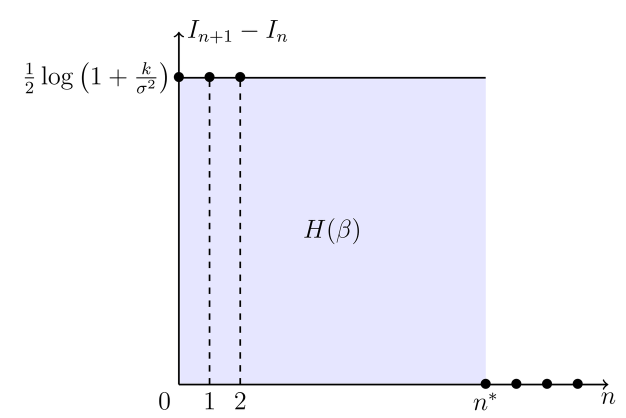


Theorem 1 (All-or-Nothing Phase Transition) *Let $\delta \in (0, \frac{1}{2})$ and $\epsilon \in (0,1)$ be two arbitrary but fixed constants. For constant $C(\delta, \epsilon) > 0$ if $k/\sigma^2 \geq C(\delta, \epsilon)$, then*

- *When $k \leq p^{\frac{1}{2} - \delta}$ and*
  $$n < (1-\epsilon)n^*,$$
  then for any $\hat{\beta} = \hat{\beta}(Y, X)$
  $$\lim_{p \to \infty} \mathrm{MSE}\left(\hat{\beta}\right)/\mathrm{MSE}_0 = 1.$$
  (weak recovery impossible!)
- *When $k = o(p)$ and*
  $$n > (1+\epsilon)n^*,$$
  then for $\hat{\beta}_{\mathrm{MLE}} = \arg\min_{v \in \{0,1\}^p, \|v\|_0 = k} \|Y - Xv\|_2$,
  $$\lim_{p \to \infty} \mathrm{MSE}\left(\hat{\beta}_{\mathrm{MLE}}\right)/\mathrm{MSE}_0 = 0.$$
  (strong recovery possible!)

Coding Theory Interpretation: "Area Theorem"

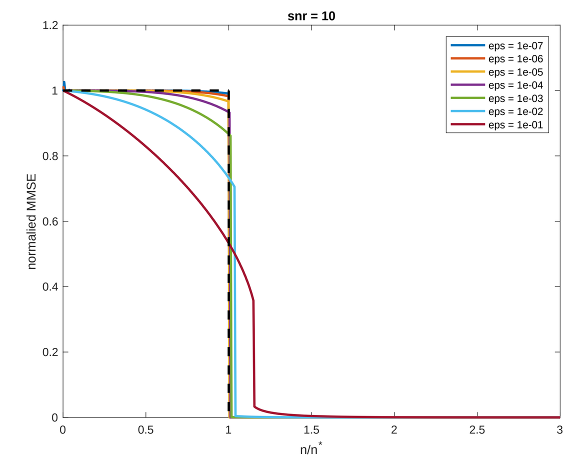
Strong recovery at $n = n^*$ implies weak recovery is impossible with $n < n^*$!



- $I_n := I(Y_1^n; X, \beta)$, the mutual information (MI) between β and $(Y_1^n; X)$.
- **Step 1:** *MI-MMSE inequality:* $I_{n+1} - I_n \leq 0.5\log(\mathrm{MMSE}_n/\sigma^2 + 1) \leq 0.5\log(k/\sigma^2 + 1)$,
  for MMSE_n the minimum MSE with n samples.
  Hence $\forall n$, $I_n \leq \frac{n}{2}\log(k/\sigma^2 + 1)$, *equality* if $\forall m < n$: $\mathrm{MMSE}_m = k$.
- **Step 2:** Strong recovery for *some* n: $I_n = H(\beta) - H(\beta|Y_1^n; X) \approx H(\beta) = \log\binom{p}{k}$.
- **Combining:** *Strong recovery* for $n = n^*$, $I_{n^*} \approx \log\binom{p}{k} \approx \frac{n^*}{2}\log(k/\sigma^2 + 1)$ (from (1))
  and therefore for $m < n^*$, $\mathrm{MMSE}_m \approx k$, i.e. *weak recovery impossible.*

Step Behavior Interpretation when k/p a small constant

(Guo, Verdu '05), (Reeves, Pfister '16), (Barbier et al '16):
Tight (replica-predicted) results on asymptotic normalised MMSE when k/p is a constant.
Plot MMSE vs n/n^* when $k/p = \epsilon \to 0$: a limiting step function jumping at $n/n^* = 1$!



Proof Ideas for the Impossibility of Weak Recovery

$$D_f(P\|Q) := \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right].$$

Cases: $\mathrm{TV}(P, Q)$: $f(x) = |x - 1|/2$, $D_{\mathrm{KL}}(P\|Q)$: $f(x) = x\log x$, $\chi^2(P\|Q)$: $f(x) = (x-1)^2$.

Step 1: Impossibility of Testing: Data Look Like Pure Noise

Let $P = P(Y, X)$ the distribution of $(Y = X\beta + W, X)$ of our data (planted distribution).
Let $Q = Q(Y, X)$ the distribution of $(Y = \lambda W, X)$ for $X \in \mathbb{R}^{n \times p}$ with i.i.d. $\mathcal{N}(0,1)$ entries, $W \in \mathbb{R}^{n \times 1}$ with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries and $\lambda = \sqrt{k/\sigma^2 + 1}$ (null model).

We show that for any $n \leq (1-\epsilon)n^*$,

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

Proof uses **conditional second moment method**: unconditional gives **wrong threshold!**

Proof Sketch for simpler $\lim_{p \to +\infty} TV(P, Q) = 0$. It holds

$$TV(P, Q) \leq \sqrt{2D_{\mathrm{KL}}(P||Q)} \leq \sqrt{\log(\chi^2(P||Q) + 1)}.$$

By straightforward calculations $\chi^2(P||Q) = \mathbb{E}_{S \sim \mathrm{Hyp}(p,k,k)}\left[\left(1 - \frac{S}{k + \sigma^2}\right)^{-n}\right] - 1$ and therefore

$$\lim_{p \to +\infty} \chi^2(P||Q) = \begin{cases} 0, & n < n^*/2 \\ +\infty, & n^*/2 < n \end{cases}$$

The case $n^*/2 < n < n^*$ is hard as **lottery effect** takes place:
Low probability events cause χ^2 to explode.

- We condition on an *appropriate* high probability event $\mathcal{E} = \mathcal{E}(Y, X)$ with $P(\mathcal{E}) = 1 - o(1)$.
- For the *conditional measure* $P_{\mathcal{E}}(\cdot) = P(\cdot \cap \mathcal{E})/P(\mathcal{E})$ we prove
  $$\lim_{p \to +\infty} \chi^2(P_{\mathcal{E}}||Q) = 0, \forall n \leq (1-\epsilon)n^*.$$

Hence $TV(P_{\mathcal{E}}, Q) = 0$ or $TV(P, Q) = 0$.

Step 2: Impossibility of Testing implies Impossibility of Estimation

We prove that for any $\hat{\beta} = \hat{\beta}(Y, X)$,

$$\mathrm{MSE}\left(\hat{\beta}\right)/k \geq 1 - 2\left(1 + \frac{\sigma^2}{k}\right)D_{\mathrm{KL}}(P||Q).$$

- A simple quantitative relation **connecting estimation and testing.**
- Possible of **independent interest**, applies for any n, p, k and β with ℓ_2 norm equal to k.
- Proof based on **the MI-MMSE inequality**.