NIPS 2010

A Dirty Model for Multitask Learning

Ali Jalali, Pradeep Ravikumar, Sujay Sanghavi and Chao Ruan

UT Austin

Motivation

- Modern Settings: High-Dimensional Problems
  - number of observations $n \ll$ number of variables p
  - Biology, Vision, Nanotechnology, Financial Analysis, ...

- Low-Dimensional Structure only hope for consistency?
  - Sparsity, Block Sparsity, Low-Rank, Graphical model Structure

- What if parameters do not have such clean structure?

This talk:

- Superposition of structures: still low-dimensional but surprisingly useful for "dirty" data

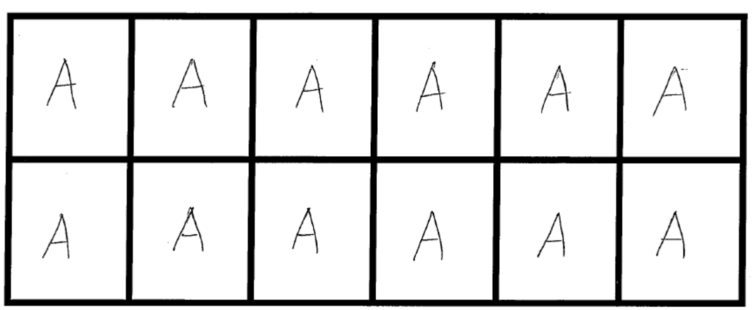
Multitask Learning

- Multiple tasks with some “shared” structure

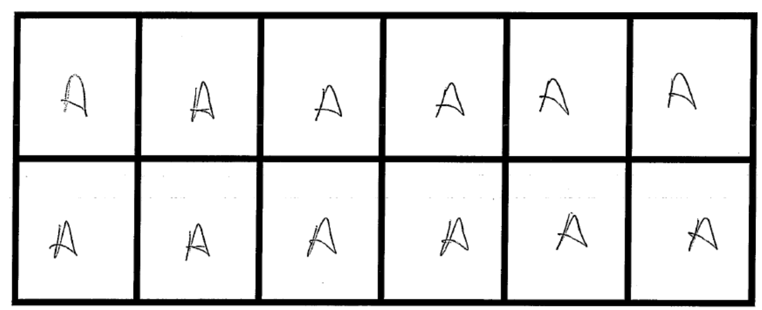
Problem:

- Learn tasks jointly (as opposed to separately)
  - e.g. Optical Character Recognition (OCR)

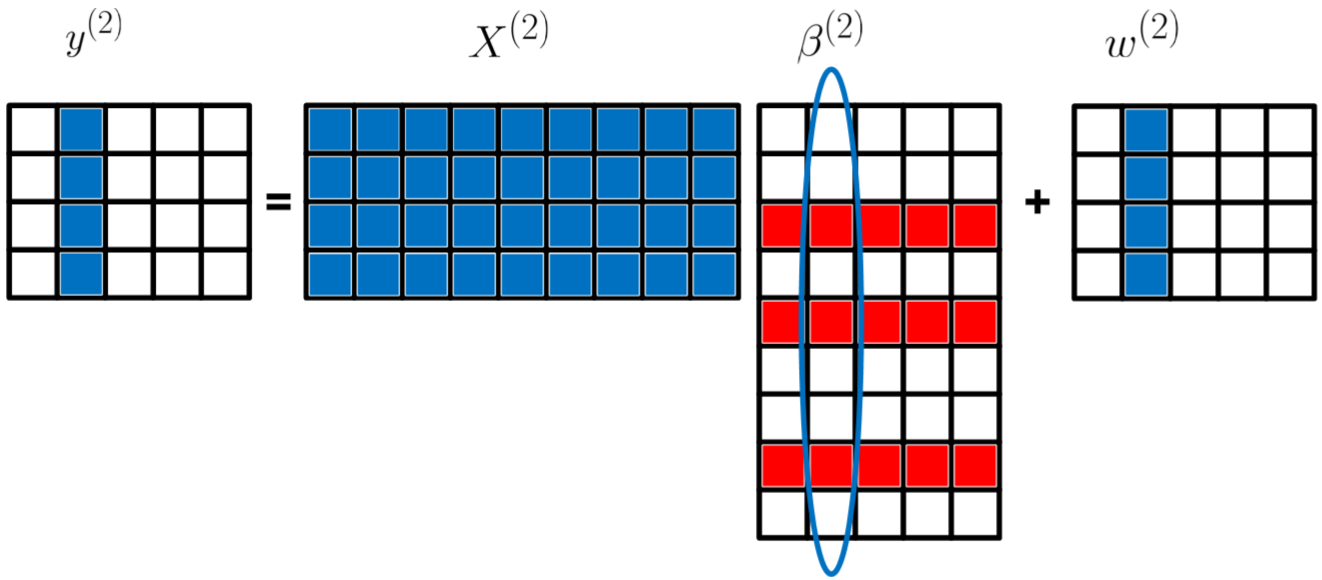
Writer 1



Writer 2

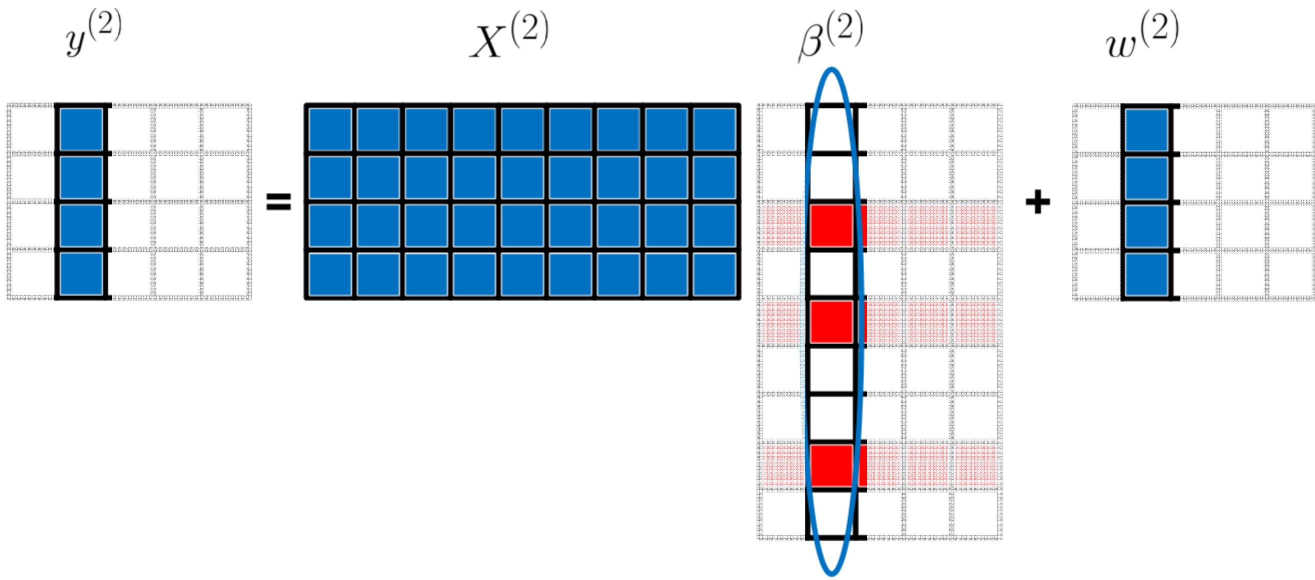


Multiple Linear Regressions



- Linear Model: $y^{(k)} = X^{(k)}\beta^{(k)} + w^{(k)}$ for all tasks $1 \leq k \leq r$
- **Problem:** Estimate β given n_k samples of $X_i^{(k)}$, $y_i^{(k)}$

Clean Model I: Sparsity

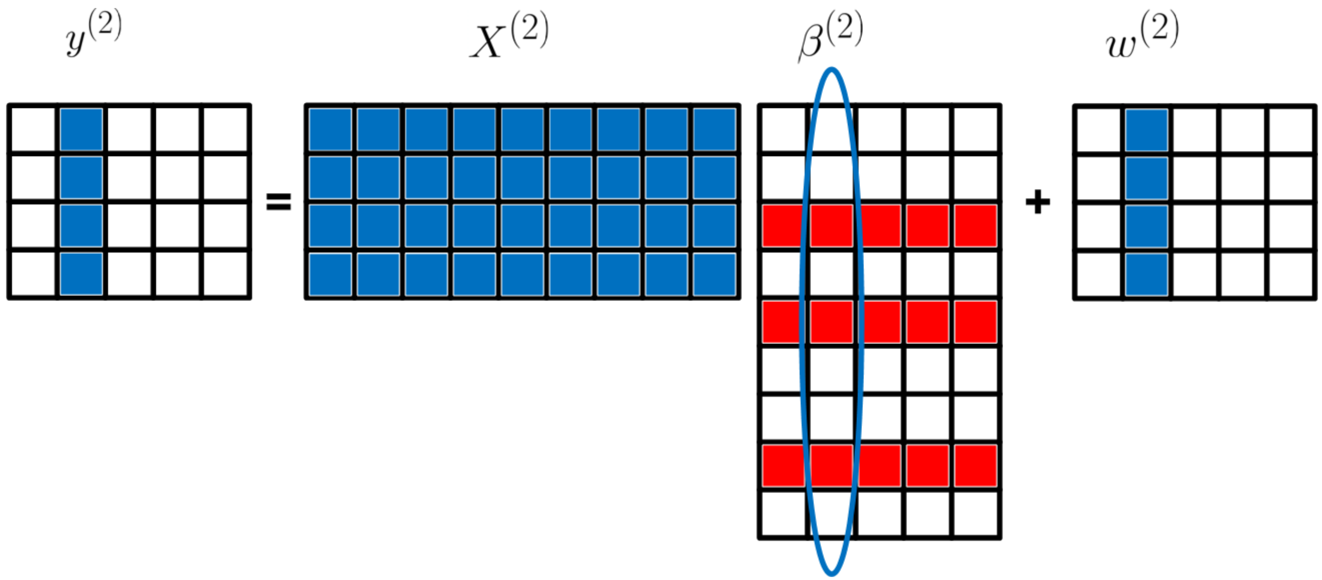


- Sparsity of each task modeled **independently.**

- LASSO [Tibshirani '96]

$$\min_{\beta^{(k)}} \frac{1}{n_k} \sum_{i=1}^{n_k} \left\| y_i^{(k)} - X_i^{(k)} \beta^{(k)} \right\|_2^2 + \lambda_k \left\| \beta^{(k)} \right\|_1$$

Clean Model II: Block-sparsity



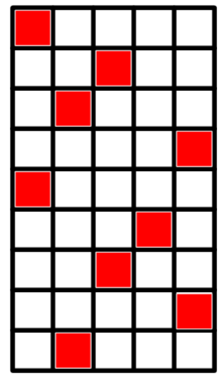
- Block-sparse structure: shared sparsity
- Group LASSO [Obozinski et al; Negahban et al; Huang et al]

$$\min_{\beta} \sum_{k=1}^{r} \frac{1}{n_k} \sum_{i=1}^{n_k} \left\| y_i^{(k)} - X_i^{(k)} \beta^{(k)} \right\|_2^2 + \lambda \left\| \beta \right\|_{1,\infty}$$

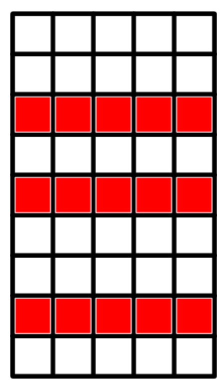
where $\|\beta\|_{1,\infty} = \sum_j \max_k \left| \beta_j^{(k)} \right|$ (sum of maximum of rows)

Existing Methods Performance

- LASSO
  - Does not model shared sparsity

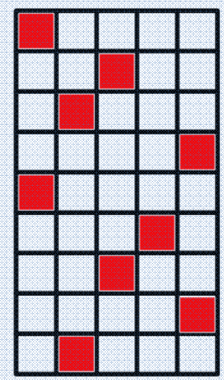


- Group LASSO
  - Does not model individual sparsity

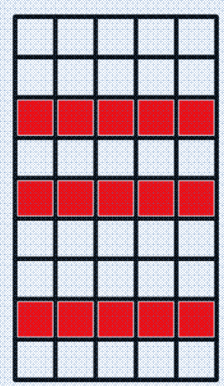


Existing Methods Performance

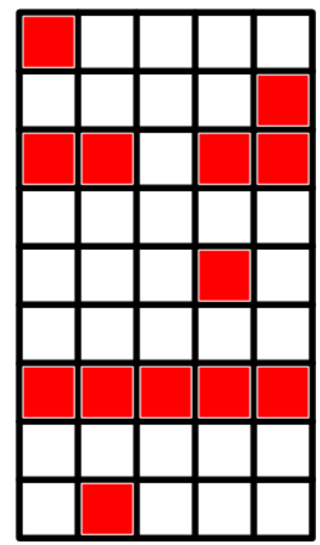
- LASSO
  - Does not model shared sparsity



- Group LASSO
  - Does not model individual sparsity



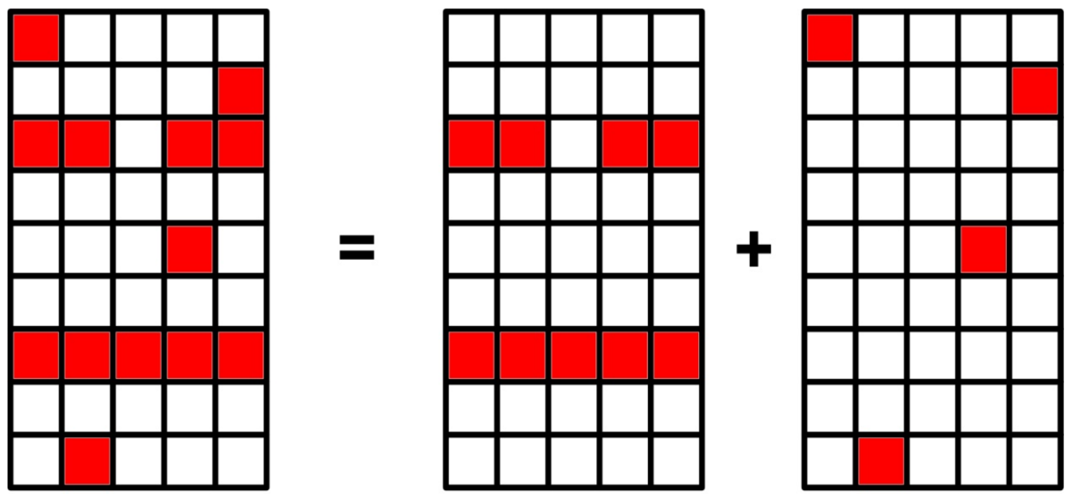
Realistic Data



β

Dirty Statistical Model

- Superposition of parameters with diff. structures



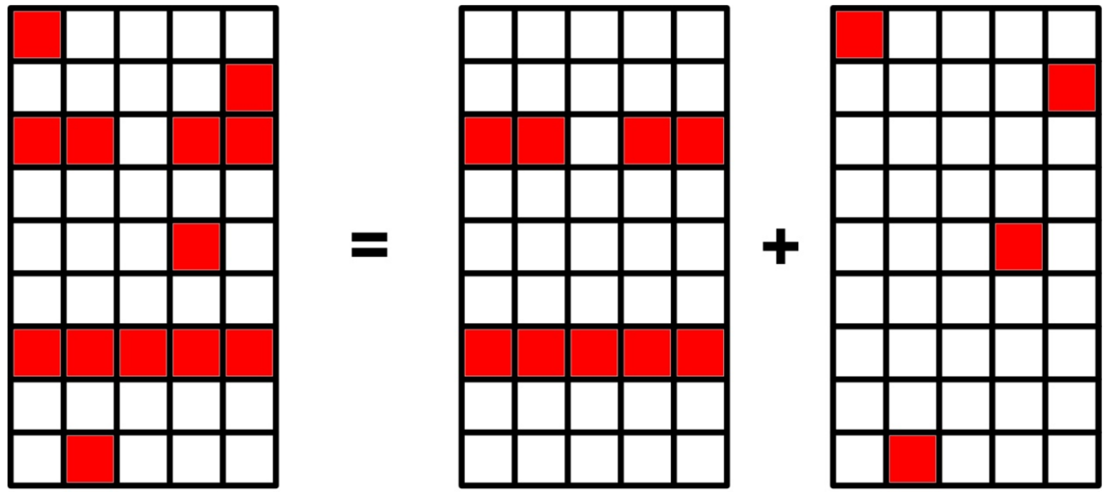
$$\beta = \mathbf{B} + \mathbf{S}$$

$$\|B\|_{1,\infty} \qquad \|S\|_{1,1}$$

shared features

non-shared features

Dirty Statistical Model



$$\beta = \mathbf{B} + \mathbf{S}$$

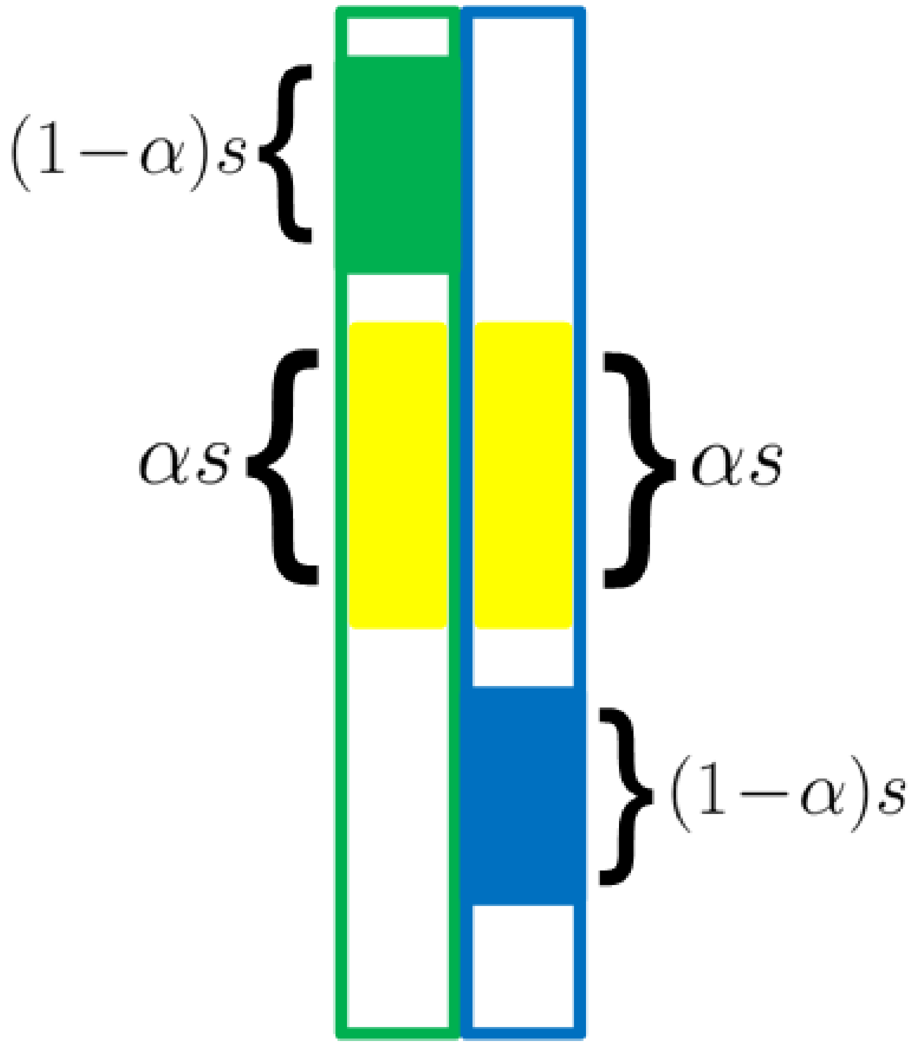
Algorithm:

$$\min_{B,S} \sum_{k=1}^{r} \frac{1}{n_k} \sum_{i=1}^{n_k} \left\| y_i^{(k)} - X_i^{(k)} \left(B^{(k)} + S^{(k)} \right) \right\|_2^2 + \lambda_b \|B\|_{1,\infty} + \lambda_s \|S\|_{1,1}$$

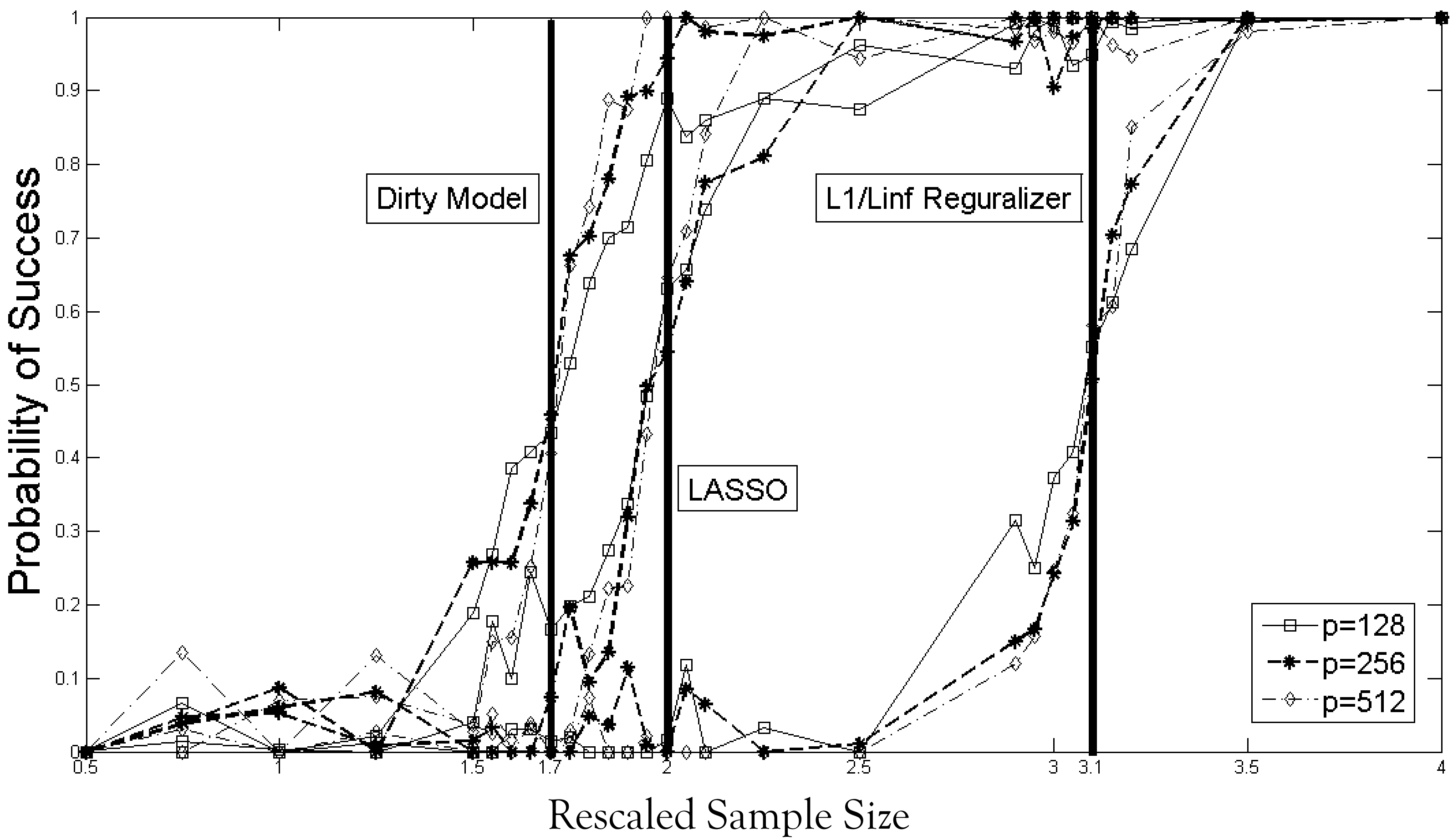
output $\hat{\beta} = \hat{B} + \hat{S}$

Two Tasks Case

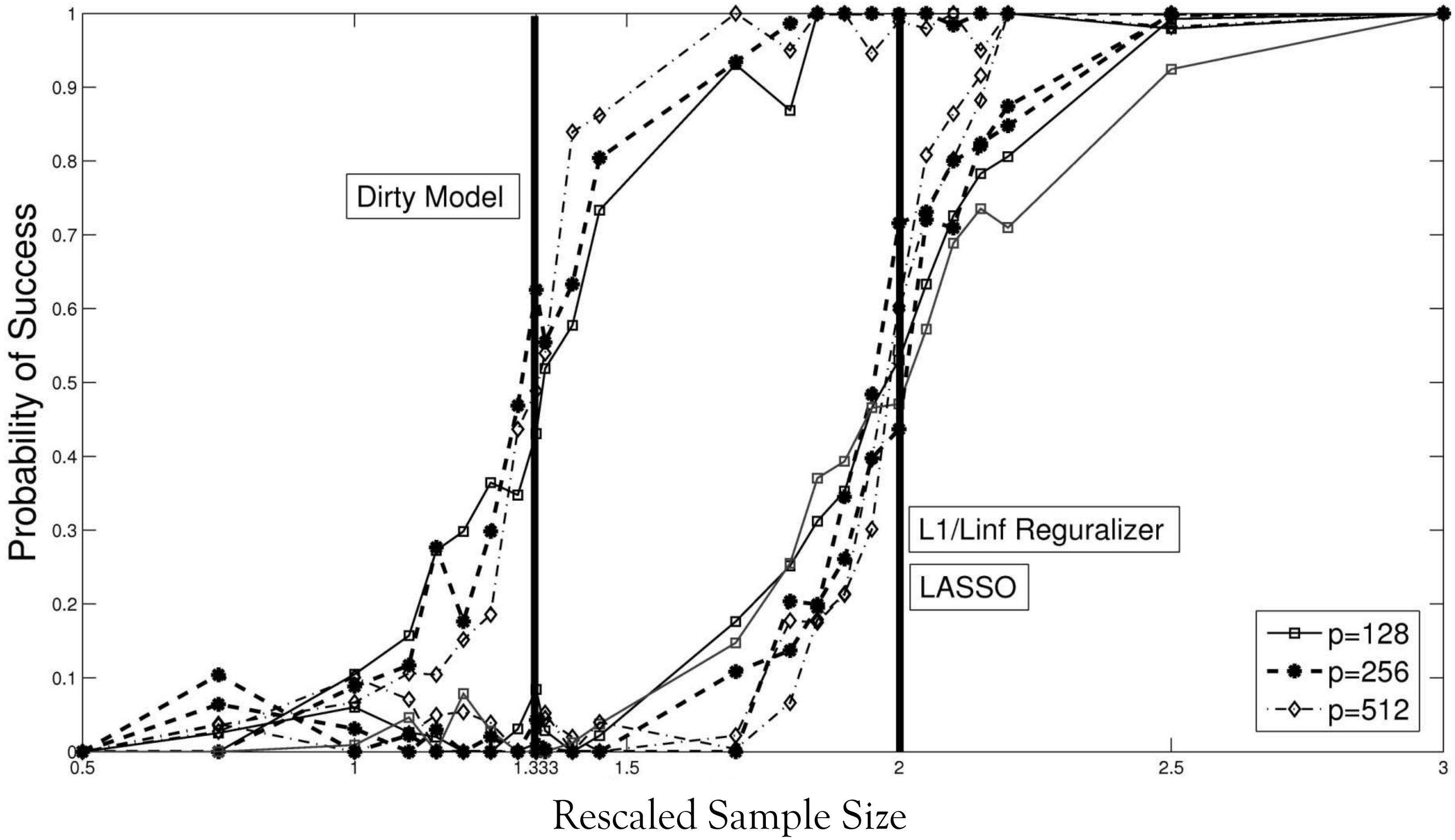
- Each task depends on "s" features
- α-portion of the features overlaps



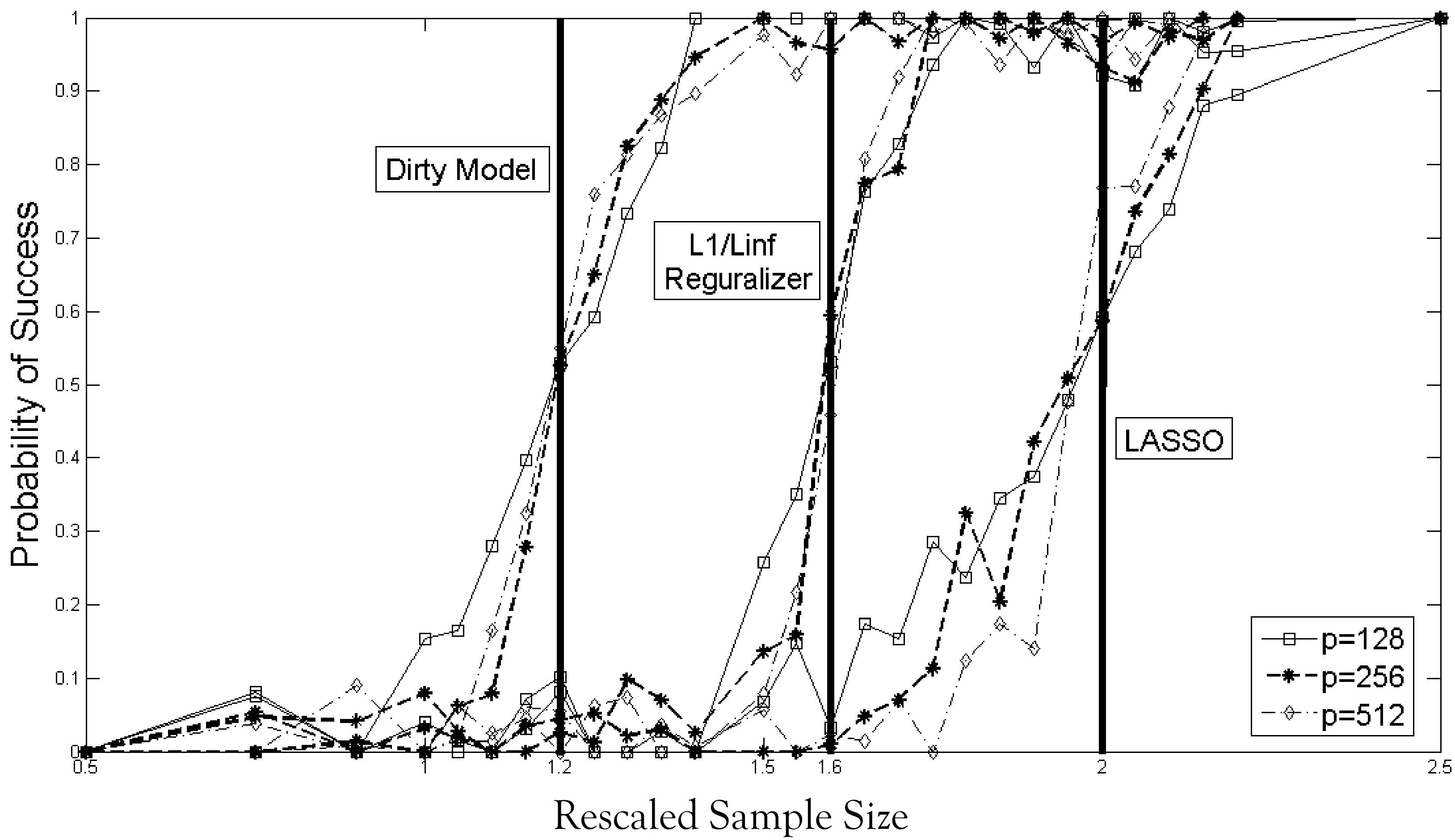
Little overlap: $\alpha = 0.3$



Medium overlap: $\alpha = 2/3$



High overlap: $\alpha = 0.8$



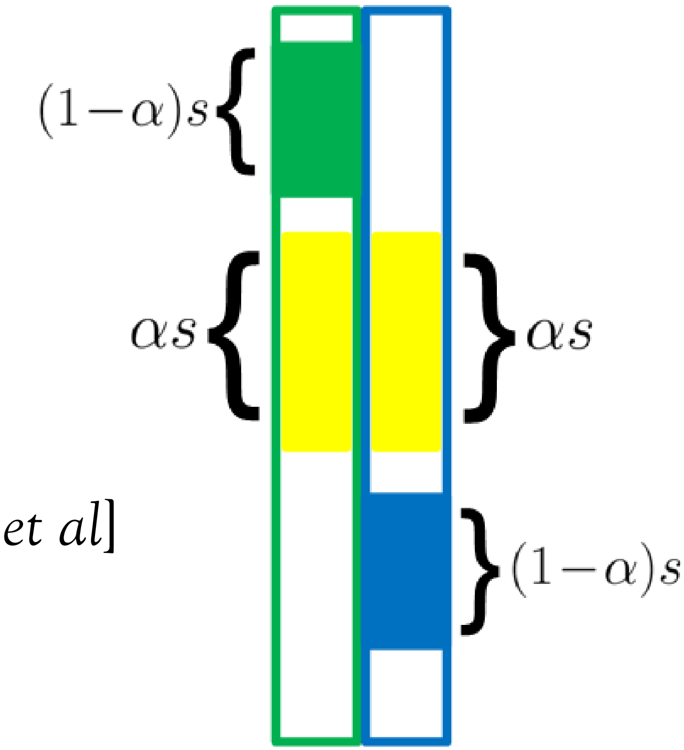
Phase Transition for Two Tasks

– LASSO:

$$\frac{n}{s\log(p)} \approx 2$$ [Wainwright]

– Group LASSO:

$$\frac{n}{s\log(p)} \approx 4-3\alpha$$ [Negahban *et al*]



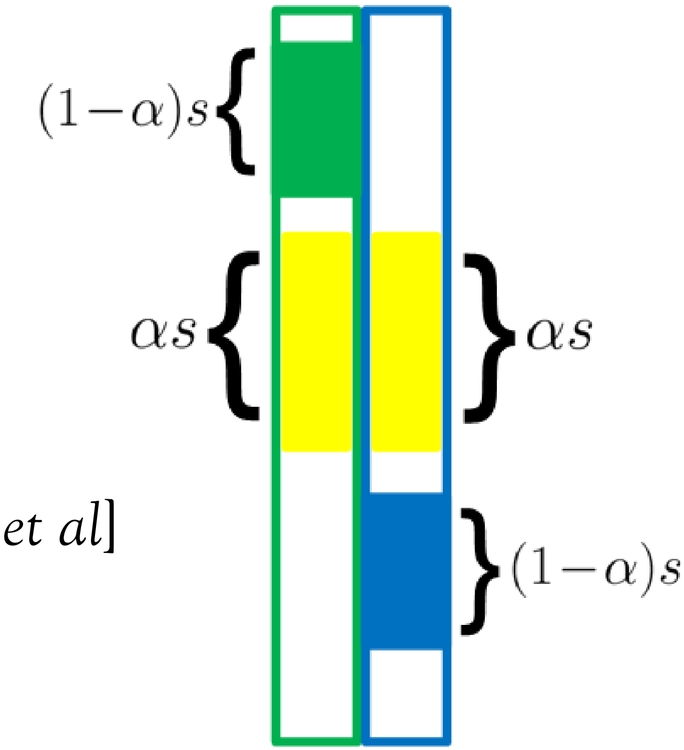
Phase Transition for Two Tasks

– LASSO:

$$\frac{n}{s\log(p)} \approx 2 \qquad \text{[Wainwright]}$$

– Group LASSO:

$$\frac{n}{s\log(p)} \approx 4-3\alpha \quad \text{[Negahban } et\ al\text{]}$$



Consequences:

$\alpha < 2/3$ :: Lasso is better

$\alpha > 2/3$ :: Group-Lasso is better

Phase Transition for Two Tasks

– LASSO:

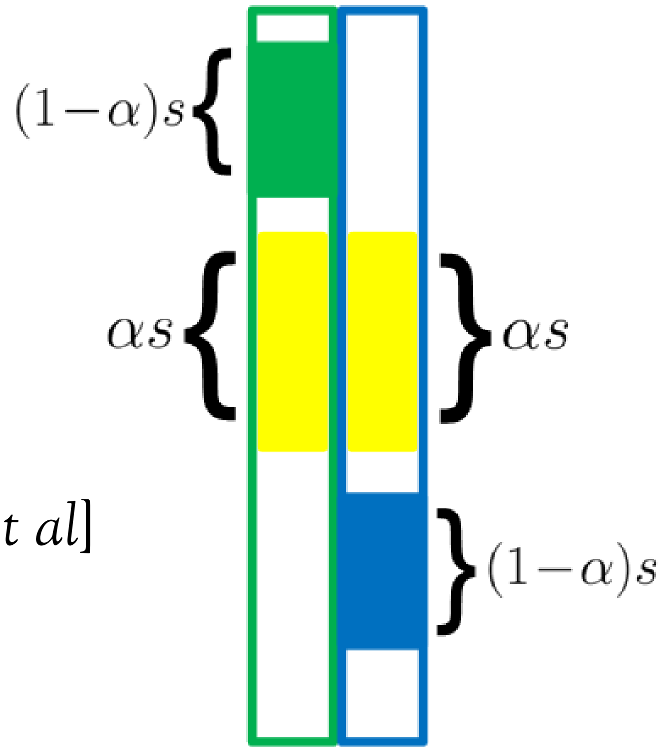
$$\frac{n}{s \log(p)} \approx 2$$ [Wainwright]

– Group LASSO:

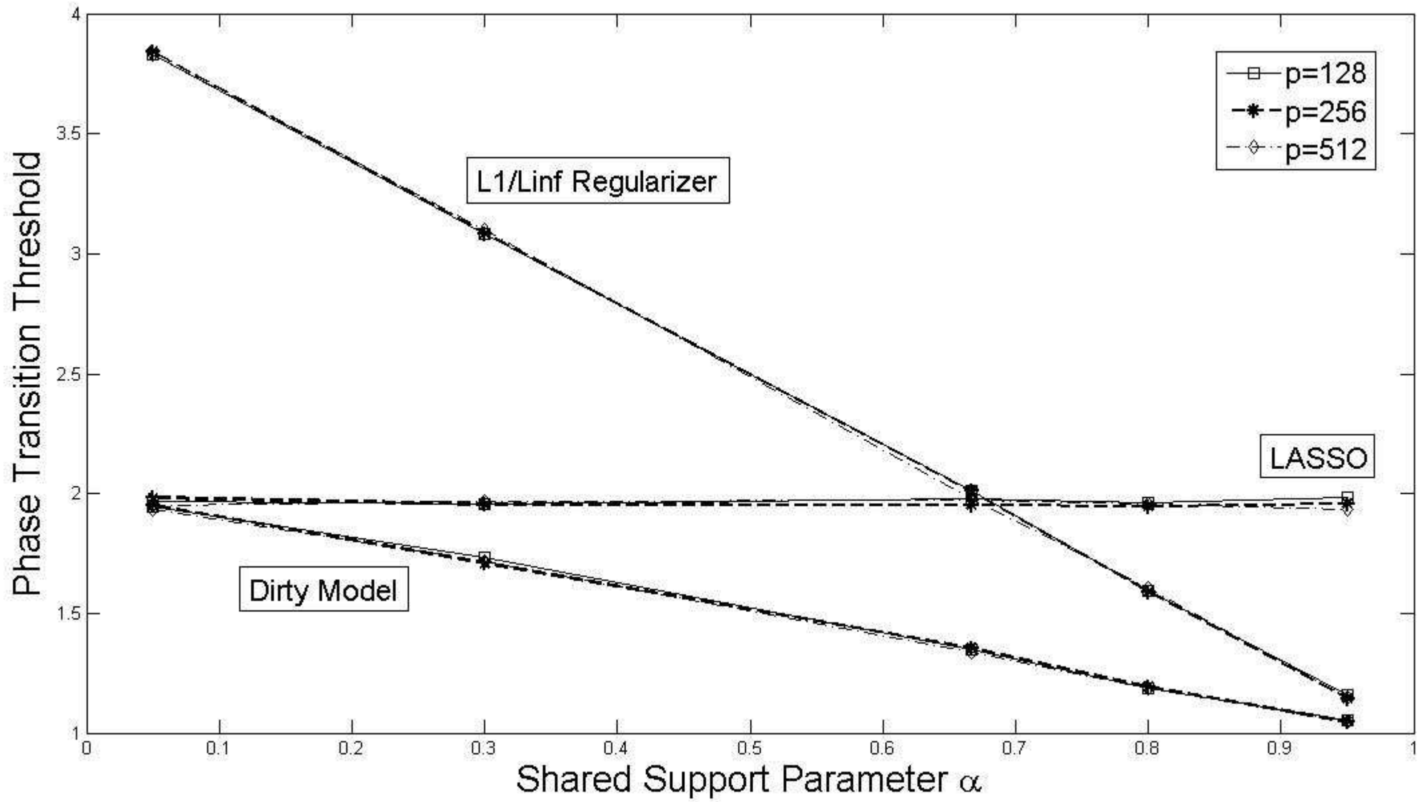
$$\frac{n}{s \log(p)} \approx 4 - 3\alpha$$ [Negahban *et al*]

– **Dirty Model:**

$$\frac{n}{s \log(p)} \approx 2 - \alpha$$



Comparison



Theorem. $$\frac{n}{s\log(p)} \approx 2-\alpha$$

Standard Assumptions

- $\mathcal{U}_k$: Support of task k
- $\mathcal{U} = \bigcup_k \mathcal{U}_k$
- $X \sim N(0, \Sigma)$
- $s = \max_k |\mathcal{U}_k|$

- Incoherence Condition

$$\max_{j \notin \mathcal{U}} \sum_{k=1}^{r} \left\| \Sigma_{j\mathcal{U}_k} \Sigma_{\mathcal{U}_k \mathcal{U}_k}^{-1} \right\|_1 < 1$$

- Eigenvalue Condition

$$\min_{1 \le k \le r} \lambda_{\min} \left(\Sigma_{\mathcal{U}_k \mathcal{U}_k} \right) > 0$$

General r-Task Case

Theorem (Gaussian Design): Under previous assumptions and

$$\lambda_b \asymp \sqrt{\frac{r\log(p)}{n}} \qquad \lambda_s \asymp \sqrt{\frac{\log(pr)}{n}}$$

If $n \geq K\, sr\log(p)$ then with probability at least $1 - C_1 \exp\left(-C_2 \log(p)\right)$:

(a) There is no false exclusion:

$$\mathrm{Supp}(\beta) \subseteq \mathrm{Supp}(\hat{\beta}) \qquad \left\|\hat{\beta} - \beta\right\|_{\infty,\infty} = \mathcal{O}\left(\sqrt{\frac{s\log(pr)}{n}}\right)$$

(b) If $\beta_{\min} = \Omega\left(\sqrt{\frac{s\log(pr)}{n}}\right)$, there is no false inclusion:

$$\mathrm{Supp}(\beta) = \mathrm{Supp}(\hat{\beta})$$

Two-Task Case

Theorem (Phase Transition): Under previous assumptions, for two tasks with α-sharedness in the support, then with high probability:

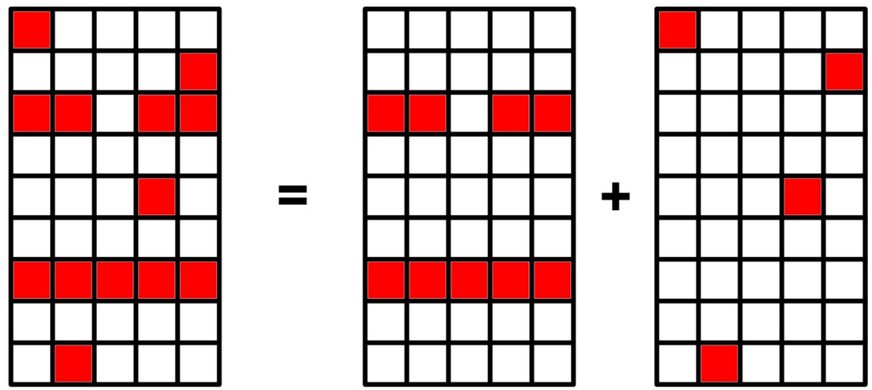
(Success) The algorithm finds the true signed support of β provided

$$\frac{n}{s\log(p-(2-\alpha)s)} > 2-\alpha$$

(Failure) The algorithm will NOT find the true signed support of β if

$$\frac{n}{s\log(p-(2-\alpha)s)} < 2-\alpha$$

General Dirty Models



$$\beta = \mathbf{B} + \mathbf{S}$$

- **Dirty Models**: "Superposition of Simple Structures"
  - – Sparse + Low-Rank
    - Latent Variables, Graph Clustering, PCA with corruptions, etc
  - – Block-Sparse + Low-Rank
    - Collaborative Filtering, PCA with Outliers, etc
  - – More Details in NIPS '10 Wkshp: Robust Statistical Learning

Summary

- Multi-task learning is challenging when there is partial overlap across tasks
  - relevant structure is neither sparsity nor block-sparsity

- A superposition of simple structures, i.e., dirty model surprisingly useful for modeling such "dirty" structure.

- For the multi-task learning problem, dirty model outperforms solo-structured lasso and group-lasso.