Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using LASSO

Presenter: Morteza Mardani

Csci 8980: Machine Learning at Large Scale and High Dimensions

February 12, 2014

Motivation

- Many practical signals are inherently parsimonious
- Sparse high-dimensional signal $\beta^* \in \mathbb{R}^p$

$$S(\beta^*) := \{i \; : \; {\beta^*}_i \neq 0\} \qquad k := |S(\beta^*)| \ll p$$

- Observations with random noise

$$y = X\beta^* + w$$

$$X \in \mathbb{R}^{n \times p} \qquad \text{Regression matrix}$$

$$\mathbb{E}[w] = 0 \qquad \text{i.i.d noise}$$

- Typically $p \gg n$ → seriously underdetermined
- **Objective**: Given *y* and *X* find the sparse β^*

Variable selection via LASSO

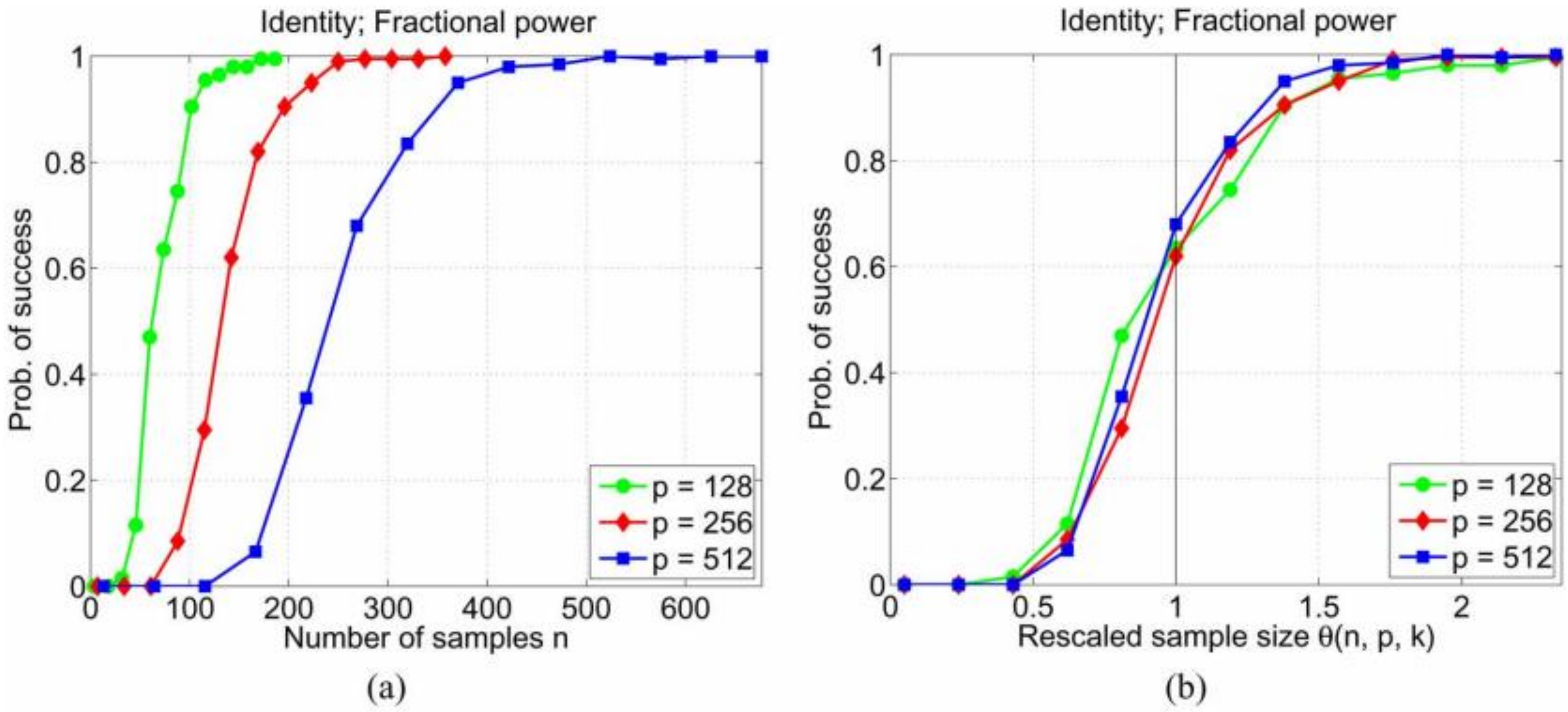
- Least Absolute Shrinkage and Selection Operator (LASSO)

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda_n \|\beta\|_1 \right\}$$

Q: What are the necessary and sufficient conditions on *(n,p,k)*

- Possible (or impossible) to ensure $S(\beta^*) = S(\hat{\beta})$,
- More ambitiously $\mathrm{sgn}(\beta^*) = \mathrm{sgn}(\hat{\beta})$

Observations



$$k = 0.4p^{0.75}$$

$$\theta(n, p, k) = n/(2k\log(p-k))$$

Road ahead

- Optimality and uniqueness conditions
- Guarantees for deterministic regression matrix
  - Achievability
  - Inachievability
- Guarntees for random (Gaussian) regression matrix
  - Achieveability
  - Inachievability

Optimality conditions

- Convex non-smooth objective function

$$f(\beta) := \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda_n\|\beta\|_1$$

- Subgradient $z \in \partial\|\beta\|_1$

$$z_i = \mathrm{sgn}(\beta_i), \quad \beta_i \neq 0; \qquad z_i \in [-1, 1], \; \beta_i = 0$$

Lemma 1: $\hat{\beta}$ is an optimal solution of LASSO iff $\exists\ \hat{z} \in \partial\|\hat{\beta}\|_1$ s.t.

$$\frac{1}{n}X^T X(\hat{\beta} - \beta^*) - \frac{1}{n}X^T w + \lambda_n \hat{z} = 0$$

Uniqueness

- $\hat{\beta}$ unique optimal $\Longleftrightarrow$ $f(\hat{\beta}) < f(\beta), \quad \forall \beta \in \mathbb{R}^p$

Lemma 2: If

1) $X^T_{S(\hat{\beta})} X_{S(\hat{\beta})}$ invertible, and
2) The subgradient $\hat{z}$ satisfies $|\hat{z}_j| < 1, \ \forall j \notin S(\hat{\beta})$, then $\hat{\beta}$ is the unique optimal solution of the LASSO problem.

- Proof …

Primal-dual witness construction

Q: If there exists a valid primal-dual witness (PDW) pair $(\check{\beta}, \check{z})$ s.t.

- $\check{\beta}$ is the unique optimal solution with $S(\check{\beta}) \subseteq S(\beta^*)$?

- Candidate primal variable

$$\check{\beta}_S = \arg\min_{\beta_S \in \mathbb{R}^k} \left\{ \frac{1}{2n} \|y - X_S \beta_S\|_2^2 + \lambda_n \|\beta_S\|_1 \right\}$$
$$\check{\beta}_{S^c} = 0$$

- Candidate dual variable

$$\check{z}_S \in \partial \|\check{\beta}_S\|_1 \qquad [\check{z}_S = \mathrm{sgn}({\beta^*}_S)]$$
$$\check{z}_{S^c} = X_{S^c}^T \left\{ X_S (X_s^T X_S)^{-1} \check{z}_S + \Pi_{X_{S^c}} \left(\frac{w}{n\lambda_n}\right) \right\}$$

PDW success

Lemma 3: Suppose that $X_S^T X_S$ invertible.

1) If $\|\check{z}_{S^c}\|_\infty < 1$, then $\check{\beta}$ unique optimal solution w/ $S(\check{\beta}) \subseteq S(\beta^*)$

2) If $\|\check{z}_{S^c}\|_\infty < 1$, $\check{z}_S = \mathrm{sgn}(\beta^*)$, then $\check{\beta}$ unique optimal w/ $\mathrm{sgn}(\beta^*) = \mathrm{sgn}(\hat{\beta})$

3) If either $\|\check{z}_{S^c}\|_\infty > 1$ or $\check{z}_S \neq \mathrm{sgn}(\beta^*)$, then LASSO fails.

- LASSO has a unique optimal solution with correct signed support if and only if PDW construction succeeds.

Existence of a valid PDW

$$Z_j := X_j^T \Big\{ X_S (X_S^T X_S)^{-1} \check{z}_S + \Pi_{X_{S^c}} \big(\tfrac{w}{n\lambda_n}\big) \Big\}, \quad j \in S^c$$

$$\Delta_i := e_i^T \left(\tfrac{1}{n} X_S^T X_S \right)^{-1} \left[\tfrac{1}{n} X_S^T w - \lambda_n \mathrm{sgn}({\beta^*}_S) \right], \qquad i \in S$$

Lemma 4: If $X_S^T X_S$ invertible,

a) $\|\check{z}_S^c\|_\infty < 1$ iff

$$|Z_j| < 1, \quad j \in S^c$$

b) $\check{z}_S = \mathrm{sgn}(\beta^*)$ iff

$$\mathrm{sgn}(\beta_i^* + \Delta_i) = \mathrm{sgn}(\beta_i^*), \quad i \in S$$

Deterministic *X*

- Incoherence conditions

A1) $|||X_{S^c}^T X_S (X_S^T X_S)^{-1}|||_\infty \leq 1 - \gamma, \quad \gamma \in (0, 1]$

A2) $\Lambda_{\min}\left(\frac{1}{n} X_S^T X_S\right) \geq C_{\min} > 0$

- $w \in \mathbb{R}^n$ i.i.d. sub-guassian with parameter σ^2

$$\mathrm{P}[|w_i| > \tau] \leq 2\exp(-\tau^2/2\sigma^2)$$

- Any r.v. with strongly log-concave density …

Achievability result

Theorem 1: Assume that A1) and A2) hold, and $\frac{1}{\sqrt{n}} \max_{j \in S^c} \|X_j\|_2 \leq 1$. If

$$\lambda_n > \frac{2}{\gamma}\sqrt{\frac{2\sigma^2 \log(p)}{n}}$$

Then, w.p. $\geq 1 - 4\exp(-c_1 n \lambda_n^2)$

a) $\hat{\beta}$ is the unique optimal sol. of LASSO w/ $S(\hat{\beta}) \subseteq S(\beta^*)$, and

$$\|\hat{\beta}_S - \beta^*{}_S\|_\infty \leq \lambda_n \left[|||(X_S^T X_S/n)^{-1}|||_\infty + \frac{4\sigma}{\sqrt{C_{\min}}} \right]$$

b) In addition, if $\beta_{\min} > g(\lambda_n)$, then $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^*)$

- Remark: $n = O(k \log(p))$

$$p = \mathcal{O}(\exp(n^{\delta_3})), k = \mathcal{O}(n^{\delta_1}),\ \beta_{\min}^2 > n^{\delta_2 - 1} \quad \text{w/} \quad 0 < \delta_1 + \delta_3 < \delta_2 < 1$$

Inachievability result

Theorem 2: Assume that A2 holds, and noise is symmetric.

a) If A1 is violated, i.e., $\max_{j \in S^c} |X_j^T X_S (X_S^t X_S)^{-1} \mathrm{sgn}({\beta^*}_S)| = 1 + \nu$, then

$$\mathrm{P}[\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^*)] \leq \frac{1}{2}, \qquad \forall n, \lambda_n > 0$$

b) If $|\beta_i^*| < \tilde{g}_i(\lambda_n)$ for some $i \in S$, where

$$\tilde{g}_i(\lambda_n) = \lambda_n \mathbf{e}_i^T (X_S^T X_S / n)^{-1} \mathrm{sgn}(\beta^*)$$

then, $\mathrm{P}[\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^*)] \leq \frac{1}{2}$.

- Mutual incoherence essential for signed support recovery

Proof sketch (Theorem 1)

- Strict dual-feasibility: $\max_{j \in S_c} |Z_j| < 1$

Step 1) Describe the r.v.

$$Z_j = \mu_j + \tilde{Z}_j$$

$$\mu_j = X_j^T X_S (X_S^T X_S)^{-1} \check{z}_S$$

$$|\mu_j| < 1 - \gamma$$

$$\tilde{Z}_j = X_j^T \Pi_{X_{S^c}} (w / \lambda_n n)$$

$$\mathrm{var}(\tilde{Z}_j) = \frac{\sigma^2}{\lambda_n^2 n^2}$$

$$\max_{j \in S^c} |Z_j| \leq (1 - \gamma) + \underbrace{\max_{j \in S^c} |\tilde{Z}_j|}_{:= \Theta}$$

Step 2) Tail bound + union bound

$$\mathrm{P}[\Theta > t] \leq 2(p - k) \exp\left(- \frac{\lambda_n^2 n t^2}{2\sigma^2} \right)$$

Cont'd

- bounded ℓ_∞- norm ($\max_{i \in S} \Delta_i := |\hat{\beta}_i - {\beta^*}_i|$)

Step 1) Describe the r.v.

$$\max_{i \in S} \Delta_i \leq \max_{i \in S} |\underbrace{\mathbf{e}_i^T \left(X_S^T X_S / n \right)^{-1} X_S^T \frac{w}{n}}_{:= V_i}| + \lambda_n ||| \left(X_S^T X_S / n \right)^{-1} |||_\infty$$

$$\mathrm{var}(V_i) \leq \frac{\sigma^2}{n C_{\min}}$$

Step 2) Tail bound + union bound

$$\mathrm{P}[\max_{i \in S} |V_i| > t] \leq 2k \exp\left(- \frac{t^2 C_{\min} n}{2\sigma^2} \right) \qquad t = 4\sigma \lambda_n / \sqrt{C_{\min}}$$

$$\max_{i \in S} \Delta_i \leq \lambda_n \left[\frac{4\sigma}{\sqrt{C_{\min}}} + ||| (X_S^T X_S / n)^{-1} |||_\infty \right]$$

Proof sketch (Theorem 2)

- **Part (a):** Assume that $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^*)$ → $\mathrm{P}[\max_{j \in S^c} |Z_j| > 1] \geq \frac{1}{2}$

$$\ell = \arg\max_{j \in S_c} \left| \mu_j := X_j^T X_S (X_S^T X_S)^{-1} \mathrm{sgn}({\beta^*}_S) \right|$$

$$Z_\ell = \mu_\ell + \tilde{Z}_\ell$$

$$|\mu_\ell| = 1 + \nu > 1 \quad \rightarrow \quad \mathrm{P}[|Z_\ell| > 1] \geq \frac{1}{2}$$

- **Part (b):** Assume that $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^*)$ → $\mathrm{P}[\mathrm{sgn}({\beta^*}_i + \Delta_i) \neq \mathrm{sgn}({\beta^*}_i)] \geq \frac{1}{2}$

  W.L.O.G. suppose ${\beta^*}_i \in (0, \tilde{g}_i(\lambda_n))$

$${\beta^*}_i + \Delta_i = \underbrace{{\beta^*}_i - \tilde{g}_i(\lambda_n)}_{:= D_i \leq 0} + \underbrace{\mathbf{e}_i^T (X_S^T X_S / n)^{-1} X_S^T w}_{:= \tilde{w}_i}$$

Random Gaussian *X*

- Observation model $y = X\beta^* + w$

  X w/ i.i.d rows $x_i \sim N(0, \Sigma)$

  $w \sim N(0, \sigma^2 I_n)$

- Incoherence conditions

  B1) $|||\Sigma_{S^c S}(\Sigma_{SS})^{-1}|||_\infty \leq 1 - \gamma, \quad \gamma \in (0, 1]$

  B2) $\Lambda_{\min}(\Sigma_{SS}) \geq C_{\min} > 0$

  B3) $\Lambda_{\max}(\Sigma_{SS}) \leq C_{\max} < \infty$

- $\Sigma = I_p \rightarrow \gamma = C_{\min} = C_{\max} = 1$

Achievability result

Theorem 3: If B1 and B2 hold, and (*n,p,k*) satisfy

$$\frac{n}{2k\log(p-k)} > (1+\delta)\theta_u(\Sigma)\Big(1+\frac{\sigma^2 C_{\min}}{\lambda_n^2 k}\Big)$$

$$\theta_u = \frac{\rho_u}{C_{\min}\gamma^2}$$

for some $\delta > 0$ then, w.h.p

a) $\hat{\beta}$ is the unique sol. of LASSO w/ $S(\hat{\beta}) \subseteq S(\beta^*)$

b) If $\beta_{\min} > g(\lambda_n)$, then $\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta^*)$, and $\|\hat{\beta}_S - \beta^*{}_S\|_\infty \leq g(\lambda_n)$

- For large λ_n, need $n \geq 2\theta_u(\Sigma)k\log(p-k)$

Inachievability result

Theorem 3: If B1-B3 hold, and (*n*,*p*,*k*) satisfy

$$\frac{n}{2k\log(p-k)} < (1-\delta)\theta_\ell(\Sigma)\Big(1+\frac{\sigma^2 C_{\max}}{\lambda_n^2 k}\Big)$$

$$\theta_\ell = \frac{\rho_\ell}{C_{\max}(2-\gamma)^2}$$

then, w.h.p. no solution of LASSO has the correct signed support.

- For large λ_n → LASSO fails for $n < 2\theta_\ell(\Sigma)k\log(p-k)$
- For small λ_n → noise dominates the signal → LASSO fails
- For uniform Gaussian $\Sigma = I_p$ → $\frac{n}{2k\log(p-k)} \gtrsim 1 + \frac{\sigma^2}{\lambda_n^2 k}$
- Noiseless case (BP): $n = \mathcal{O}(p)$ → $k = \mathcal{O}(p)$