Iterative algorithms

→ alternative to convex relaxation, with no shrinkage bias

→ computationally efficient.

① Iterative hard thresholding

. Coordinate sparse linear regression:

$$Y = X\beta^* + \varepsilon \quad \text{with} \quad |\beta^*|_0 \text{ small and } \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$$

. Reminder on Lasso and proximal optimisation:

. The lasso estimator

$$\hat{\beta}^{\text{Lasso}} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|Y - X\beta\|^2 + \lambda |\beta|_1 \right\} \qquad (6.2)$$

has been derived as a convex relaxation of

$$\hat{\beta}^{\text{ns}} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|Y - X\beta\|^2 + \lambda^2 |\beta|_0 \right\} \qquad (6.1)$$

. proximal recipe:

→ approximate $F(\beta) = \|Y - X\beta\|^2$ by $F(\beta^t) + \langle \nabla F(\beta^t), \beta - \beta^t \rangle + \frac{1}{2\eta} \|\beta - \beta^t\|^2$

→ iterate

$$\begin{aligned} \beta^{t+1} &\in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ F(\beta^t) + \langle \nabla F(\beta^t), \beta - \beta^t \rangle + \frac{1}{2\eta} \|\beta - \beta^t\|^2 + \lambda |\beta|_1 \right\} \\ &= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\beta - (\beta^t - \eta \nabla F(\beta^t))\|^2 + \eta\lambda |\beta|_1 \right\} \\ &= S_{\lambda\eta}(\beta^t - \eta \nabla F(\beta^t)) \end{aligned}$$

where $S_\lambda(\alpha) \in \mathbb{R}^p$

$$\begin{aligned} S_\lambda(\alpha) &= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\alpha - \beta\|^2 + \lambda |\beta|_1 \right\} \\ &= \begin{bmatrix} \alpha_1 (1 - \lambda/|\alpha_1|)_+ \\ \vdots \\ \alpha_p (1 - \lambda/|\alpha_p|)_+ \end{bmatrix} \in \mathbb{R}^p \end{aligned}$$

} Soft thresholding operator

Iterative Hard Thresholding:

Why not applying the proximal method on the original problem (6.1) ?

$$\beta^{t+1} \in \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ F(\beta^t) + \langle \nabla F(\beta^t), \beta-\beta^t\rangle + \frac{1}{2\eta}\|\beta-\beta^t\|^2 + \lambda^2|\beta|_0 \right\}$$

$$= \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\beta - (\beta^t - \eta\nabla F(\beta^t))\|^2 + 2\eta\lambda^2|\beta|_0 \right\}$$

$$= H_{\lambda\sqrt{2\eta}}\left(\beta^t - \eta\nabla F(\beta^t)\right)$$

where

$$H_\lambda(\alpha) = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\alpha-\beta\|^2 + \lambda^2|\beta|_0 \right\} \quad (\alpha \in \mathbb{R}^p)$$

$$\overset{\text{Exercise 2.8.1}}{=} \begin{bmatrix} \alpha_1 \mathbb{1}_{|\alpha_1|>\lambda} \\ \vdots \\ \alpha_p \mathbb{1}_{|\alpha_p|>\lambda} \end{bmatrix} \in \mathbb{R}^p$$

Hard Thresholding.

☺ no shrinkage

☹ (6.1) is non-convex so no convergence guarantee

- fix $\eta = 1/2$, $\hat\beta^0 = 0$
- choose λ_t decreasing from a high-value to the optimal level $\lambda_{\text{opt}} = c\,\sigma\sqrt{\log p}$ for (6.1)

(IHT)
- $\eta = 1/2$, $\lambda_t = a^{-t}A + B$, $\hat\beta^0 = 0$
- $\hat\beta^{t+1} = H_{\lambda_{t+1}}\left(\hat\beta^t - \underbrace{\tfrac{1}{2}\nabla F(\hat\beta^t)}_{=X^T(X\hat\beta^t - Y)}\right) = H_{\lambda_{t+1}}\left(\underbrace{(I-X^TX)}_{=:A}\hat\beta^t + X^TY\right)$

② Risk analysis for IHT

a/ Discussion

- β^t_{Lasso} = approximation $\hat{\beta}^{Lasso}$ after t steps of Soft-Thresholding iterations.

- classical approach:

$$\|\beta^t_{Lasso} - \beta^*\| \leq \underbrace{\|\beta^t_{Lasso} - \hat{\beta}^{Lasso}\|}_{\text{optimisation error}} + \underbrace{\|\hat{\beta}^{Lasso} - \beta^*\|}_{\text{statistical error}}$$

⇝ the two errors are analysed apart

- IHT:
  - $\hat{\beta}^t \xrightarrow{?} \cdots$ ⇒ cannot split into optimisation / statistical errors
  - recipe: $\Lambda = I - X^T X$

$$\hat{\beta}^{t+1} := H_{\lambda_{t+1}}(\Lambda \hat{\beta}^t + X^T Y) \overset{Y = X\beta^* + \varepsilon}{=\!=} H_{\lambda_{t+1}}\big(\beta^* + \underbrace{\Lambda(\hat{\beta}^t - \beta^*)}_{\text{contraction?}} + \underbrace{X^T\varepsilon}_{=Z}\big)$$

(β^*: target; $\Lambda(\hat{\beta}^t - \beta^*)$: contraction?; $X^T\varepsilon = Z$: statistical noise.)

With a suitable tuning of λ_t (to keep β^t sparse) $\hat{\beta}^t - \beta^*$ will be sparse and Λ will act as a contraction.

Hence for t large

$$\hat{\beta}^t \text{ "}\approx\text{" } \underbrace{H_{\lambda_\infty}(\beta^* + Z)}_{\text{optimal estimation of } \beta^* \text{ from } \beta^* + Z}$$

b/ Risk bound

Theorem 6.1 (with $c=1$) Deterministic bound

Assumptions: $\Lambda = I - X^T X$

- for some $0 < \delta < 1/3$:

$$\max_{\substack{|S| \le \tilde{k} \\ S \subset \{1,\dots,p\}}} |\Lambda_{SS}|_{op} \le \delta, \quad \text{with } \tilde{k} = 3|\beta^*|_0 \tag{6.10}$$

- $1 < a \le \frac{1}{3\delta}$, $A \ge \frac{\|\beta^*\|}{3|\beta^*|_0^{1/2}}$, $B > \frac{a}{a-1}|X^T\varepsilon|_\infty$ (6.11)

Then: writting $m^* = \mathrm{supp}(\beta^*)$

(i) $|\hat{\beta}^t_{\bar{m}^*}|_0 \le |\beta^*|_0$ where $\bar{m}^* = \{1,\dots,p\} \setminus m^*$

(ii) $\|\hat{\beta}^t - \beta^*\| \le 3\lambda_t \sqrt{|\beta^*|_0}$

Discussion:

① since $|X^T\varepsilon|_\infty \simeq \sigma\sqrt{2\log p}$, choosing $B \simeq c\sigma\sqrt{\log p}$,

we have $\lambda_t \gg c\sigma\sqrt{\log p}$ and $\|\hat{\beta}^t - \beta^*\|^2 \underset{t\to\infty}{\lesssim} c'\sigma^2|\beta^*|_0 \log p$

$\rightsquigarrow$ See Corollary 6.3 below

② The condition (6.10) ensures the contraction of $\Lambda(\hat{\beta}^t - \beta^*)$ for $\hat{\beta}^t$ sparse enough

③ (i) states that $\hat{\beta}^t$ remains sparse for $\lambda_t = a^{-t}A + B$.

④ The condition (6.10) ensures the restricted isometry property

$$(1-\delta)\|\beta\|^2 \le \|X\beta\|^2 \le (1+\delta)\|\beta\|^2 \quad \text{for all } \beta \text{ with } |\beta|_0 \le \tilde{k}$$

(Check it!)

Proof of Theorem 6.1:

notation: $Z = X^T\varepsilon$, $b^{t+1} = \underset{+Z}{\beta^* + \Lambda(\hat\beta^t - \beta^*)}$ so that $\hat\beta^{t+1} = H_{\lambda_{t+1}}(b^{t+1})$;
$k^* = |\beta^*|_0$.

Lemma 6.2: contraction property.

If (i) and (ii) hold at step t, then $\forall |S| \le k^*$

$$\|(b^{t+1} - \beta^*)_S\| \le \delta \|\beta^* - \hat\beta^t\| + \sqrt{|S|}\,|Z|_\infty$$

Proof Lemma 6.2:

- Set $\tilde S^t = m^* \cup \mathrm{supp}(\hat\beta^t_{\bar m^*})$ and $S' = \tilde S^t \cup S$.

  $|S'| \le |\tilde S^t| + |S| \le 3k^* = \bar k$ by (i) at step t.

- $\|(b^{t+1} - \beta^*)_S\| \le \|(\Lambda(\hat\beta^t - \beta^*))_S\| + \|Z_S\|$

  $S \subset S'$ $\le \|(\Lambda(\hat\beta^t - \beta^*))_{S'}\| + |S|^{1/2}|Z|_\infty$

  $\le \|\Lambda_{S'S'}(\hat\beta^t - \beta^*)_{S'}\| + |S|^{1/2}|Z|_\infty$

  (6.10) $\to \le \delta \|(\hat\beta^t - \beta^*)_{S'}\| + |S|^{1/2}|Z|_\infty$

  $= \delta \|\hat\beta^t - \beta^*\| + |S|^{1/2}|Z|_\infty$. □

- We analyse $\hat\beta^{t+1} - \beta^*$ on m^* and $\bar m^*$ apart.

- On m^*:

$$\|\hat\beta^{t+1}_{m^*} - \beta^*_{m^*}\| \overset{\Delta}{\le} \underbrace{\|b^{t+1}_{m^*} - H_{\lambda_{t+1}}(b^{t+1}_{m^*})\|}_{\text{definition of } H_{\lambda_{t+1}} \to \le \sqrt{|m^*|}\,\lambda_{t+1}} + \underbrace{\|b^{t+1}_{m^*} - \beta^*_{m^*}\|}_{\underset{\text{Lem. 6.2}}{\le} \delta\|\hat\beta^t - \beta^*\| + \sqrt{|m^*|}\,|Z|_\infty \overset{\text{(ii)}}{\le} 3\lambda_t\sqrt{|m^*|}}$$

$$\le \sqrt{k^*}\,(\lambda_{t+1} + \underbrace{3\delta\lambda_t + |Z|_\infty}) \le 2\lambda_{t+1}\sqrt{k^*} \quad (6.17)$$

$$(6.11) \to \le \tfrac{1}{a}\lambda_t + \tfrac{a-1}{a}B = \lambda_{t+1}$$

on $\bar{m}^*$:

· We first prove $|\hat{\beta}^{t+1}_{\bar{m}^*}|_0 < k^*$.

For any $S \subset \text{supp}(\hat{\beta}^{t+1}_{\bar{m}^*})$ with $|S| \le k^*$ we have

$$\lambda_{t+1}\sqrt{|S|} \overset{\text{def. of } H_{\lambda_{t+1}}}{\le} \|\hat{\beta}^{t+1}_S\| \overset{S \subset \text{supp}(\hat{\beta}^{t+1}_{\bar{m}^*})}{=} \|b^{t+1}_S\| \overset{S \subset \bar{m}^*}{=} \|b^{t+1}_S - \beta^*_S\|$$

$$\overset{\text{Lem. 6.2}}{\le} \sqrt{|S|}\,|Z|_\infty + \delta \underbrace{\|\beta^* - \hat{\beta}^t\|}_{\overset{(ii)}{\le} 3\lambda_t\sqrt{k^*}} \qquad (6.18)$$

$$\overset{\text{(6.11) and } |S|\le k^*}{\le} \sqrt{k^*}\Big(\frac{a-1}{a}B + \underbrace{3\delta\lambda_t}_{\le \frac{1}{a}\lambda_t}\Big) \qquad (6.18\text{ bis})$$

$$\le \lambda_{t+1} \quad \text{since } \lambda_{t+1} = \frac{1}{a}\lambda_t + \frac{a-1}{a}B$$

Hence, $|S| < k^*$ which means that we cannot find $S \subset \text{supp}(\hat{\beta}^{t+1}_{\bar{m}^*})$ with $|S| = k^*$.

We have proved $|\hat{\beta}^{t+1}_{\bar{m}^*}|_0 < k^*$, i.e. (i) for step $t+1$.

· upper-bound on $\|(\hat{\beta}^{t+1} - \beta^*)_{\bar{m}^*}\|$: (6.18 bis) with $S = \text{supp}(\hat{\beta}^{t+1}_{\bar{m}^*})$ gives

$$\|\hat{\beta}^{t+1}_{\bar{m}^*} - \beta^*_{\bar{m}^*}\| = \|\hat{\beta}^{t+1}_{\bar{m}^*}\| \overset{(6.18\text{ bis})}{\le} \lambda_{t+1}\sqrt{k^*}$$

conclusion: with (6.17)

$$\|\hat{\beta}^{t+1} - \beta^*\| \le \|\hat{\beta}^{t+1}_{m^*} - \beta^*_{m^*}\| + \|\hat{\beta}^{t+1}_{\bar{m}^*} - \beta^*_{\bar{m}^*}\|$$

$$\le 3\lambda_{t+1}\sqrt{k^*}$$

which is (ii) for step $t+1$.

□

Discussion: if we set $\hat{f} = \lceil \log_a (A/B) \rceil$, then

$\lambda_{\hat{\ell}} = a^{-\hat{f}} A + B \leq 2B.$

So $\| \hat{\beta}^{\hat{\ell}} - \beta^* \| \leq 6B \sqrt{|\beta^*|_0}$.

Corollary 6.3: error bound for IHT

- Assume that $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, $\|X_j\| = 1$, for $j = 1, \dots, p$ and (6.10) hold.
- Set for $K > 1$ and $1 < a \leq 1/(3\delta)$

$$A = \frac{\|X^T Y\| + \sigma |X|_{op} (\sqrt{2} + \sqrt{2K \log p})}{3(1-\delta)} \quad \text{and} \quad B = \frac{a\sigma}{a-1} \sqrt{2K \log p}$$

- Then, with probability $\geq 1 - \frac{2}{p^{K-1}}$

(i) $$\| \hat{\beta}^{\hat{\ell}} - \beta^* \|^2 \leq 2K \left(\frac{6a}{a-1} \right)^2 \sigma^2 |\beta^*|_0 \log(p)$$

(ii) $$\| X\hat{\beta}^{\hat{\ell}} - X\beta^* \|^2 \leq 2(1+\delta) K \left(\frac{6a}{a-1} \right)^2 \sigma^2 |\beta^*|_0 \log(p)$$

Proof of Corollary 6.3:

- All we need is to prove that (6.11) holds. Indeed, then (i) holds and $|\hat{\beta}^{\hat{\ell}}_{\bar{m}^*}|_0 \leq |\beta^*|_0$ so the restricted isometry property (6.10) ensures (ii). (from (i))
- In the analysis of Lasso estimator, we have already seen that $\mathbb{P}[B > \frac{a}{a-1} |X^T \varepsilon|_\infty] \geq 1 - \frac{1}{p^{K-1}}$
- So, it remains to prove that

$$\mathbb{P}\left[A \geq \frac{\|\beta^*\|}{3\sqrt{|\beta^*|_0}} \right] \geq 1 - \frac{1}{p^{K-1}} \qquad (*)$$

- Gaussian concentration: $\varepsilon \to \|X^T X\beta^* + X^T\varepsilon\|$ is $|X|_{op}$-Lipschitz so there exist $\zeta, \zeta' \sim \text{Exp}(1)$ such that

$$\mathbb{E}[\|X^T Y\|] - \sigma|X|_{op}\sqrt{2\zeta} \leq \|X^T Y\| \leq \mathbb{E}[\|X^T Y\|] + \sigma|X|_{op}\sqrt{2\zeta'}$$

As in Lecture 1:

$$\mathbb{E}[\|X^T Y\|^2] \leq \mathbb{E}\Big[\big(\mathbb{E}[\|X^T Y\|] + \sigma|X|_{op}\sqrt{2\zeta'}\big)^2\Big]$$
$$\overset{\text{Jensen}}{\leq} \big(\mathbb{E}[\|X^T Y\|] + \sigma|X|_{op}\sqrt{2}\big)^2$$

So

$$\|X^T Y\| + \sigma|X|_{op}(\sqrt{2} + \sqrt{2\zeta}) \geqslant \mathbb{E}[\|X^T Y\|] + \sigma|X|_{op}\sqrt{2}$$
$$\geqslant \mathbb{E}[\|X^T Y\|^2]^{1/2}$$
$$= \sqrt{\|X^T X\beta^*\|^2 + \underbrace{\mathbb{E}[\|X^T\varepsilon\|^2]}_{\geq 0}} \geqslant \|(X^T X\beta^*)_{m^*}\|$$
$$\geqslant \|\beta^*_{m^*}\| - \|(\Lambda\beta^*)_{m^*}\| = \|\beta^*_{m^*}\| - \|\Lambda_{m^*m^*}\beta^*_{m^*}\|$$
$$(6.10) \geqslant (1-\delta)\|\beta^*_{m^*}\| = (1-\delta)\|\beta^*\|$$

Hence $(*)$ holds.

□

Take home message:

- We have obtained some results similar as for the Lasso
- The number $\hat{t}$ of iterations is controlled.

(3) Group sparsity: Iterative Group Thresholding

→ Section 6.2 of lecture notes.