Iterative Algorithms

Lecture 4

Reminder

- Model: $y_i = \langle \beta^*, x_i \rangle + \varepsilon_i$, $i = 1, \dots, n$

with $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

- Notation:

$Y := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$; $f^* := \begin{bmatrix} f^*(x_1) \\ \vdots \\ f^*(x_n) \end{bmatrix}$; $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$

and $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}$

$\Rightarrow \quad Y = X\beta^* + \varepsilon = f^* + \varepsilon.$

- Hidden structure: we assume that

$\underbrace{|\beta^*|_0 := \text{card}\{j : \beta_j^* \neq 0\} \text{ is small}}_{\text{Coordinate sparse assumption.}}$

- For $S \subset \{1, \dots, p\}$, we set

$\bar{S} = \{X\beta : \text{supp}(\beta) \subset S\}$

and $\hat{f}^{(S)} := X \hat{\beta}^{(S)}$ with

$$\hat{\beta}^{(S)} \in \underset{\beta \in \bar{S}}{\text{argmin}} \|Y - X\beta\|^2$$

- Structure learning:

$$\hat{S} \in \underset{S \subset \{1, \dots, p\}}{\text{argmin}} \text{crit}(S) \quad (\Pi S)$$

where $\text{crit}(S) = \|Y - \hat{f}^{(S)}\|^2 + \text{pen}(S)\sigma^2$

with $\text{pen}(S) = \underset{\substack{\uparrow \\ \text{constant}}}{K} |S| \log \frac{e p}{|S|}$

fulfills

$$R(\hat{f}^{(\hat{S})}) \lesssim \underbrace{\frac{\sigma^2}{n} |\beta^*|_0 \log \frac{e p}{|\beta^*|_0}}_{\text{minimax optimal}}$$

- Problem

solving (ΠS) is computationally prohibitive.

Trick 1 : replace (NS) by a surrogate convex optimisation problem. ⇝ Lasso estimator.

issue : shrinkage bias

Trick 2 : iterative algorithms

→ computationally more efficient

→ no shrinkage.

today :

1) good old forward-backward algorithm
2) iterative hard-thresholding

① Forward-backward

Recipe : try to minimise (NS) by adding / removing one variable at a time.

Solving (NS) is NP-hard in general.

So, in general, there is <u>no</u> hope to exactly solve (NS) with such a heuristic.

Forward-Backward

- Init: $\hat{S} = \emptyset$
- Iterate: until convergence
  - Forward step:
    - $j_+ \in \underset{j \notin \hat{S}}{\operatorname{argmin}} \operatorname{crit}(\hat{S} \cup \{j\})$
    - if $\operatorname{crit}(\hat{S} \cup \{j_+\}) < \operatorname{crit}(\hat{S})$ then
      $\hat{S} \leftarrow \hat{S} \cup \{j_+\}$
  - Backward step:
    - $j_- \in \underset{j \in \hat{S}}{\operatorname{argmin}} \operatorname{crit}(\hat{S} \setminus \{j\})$
    - if $\operatorname{crit}(\hat{S} \setminus \{j_-\}) \leq \operatorname{crit}(\hat{S})$ then
      $\hat{S} \leftarrow \hat{S} \setminus \{j_-\}$
- Output: $\hat{\beta}^{(\hat{S})}$

→ computationally very efficient

→ theoretical garanties quite similar to those for the Lasso estimator.

→ in practice, it seems to be a bit greedy compared to the Lasso optimisation.

2 Iterative hard-thresholding

Inspired by proximal optimisation

a) Reminder on proximal optimisation

- Assume that you want to minimise

$$\hat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} F(\beta)$$

with F convex and smooth.

No close-form solution ?

Taylor

$$F(\beta) = F(\beta') + \langle \nabla F(\beta'), \beta - \beta' \rangle + O(\|\beta - \beta'\|^2)$$

$\Downarrow$ iterate

$$\beta^{t+1} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \Big\{ \underbrace{F(\beta^t)}_{\text{no } \beta \text{ here :)}} + \langle \nabla F(\beta^t), \beta - \beta^t \rangle + \frac{1}{2\eta} \|\beta - \beta^t\|^2 \Big\}$$

$$= \operatorname*{argmin}_{\beta} \frac{1}{2\eta} \|\beta - (\beta^t - \eta \nabla F(\beta^t))\|^2$$

solution: $\beta^{t+1} = \beta^t - \eta \nabla F(\beta^t)$

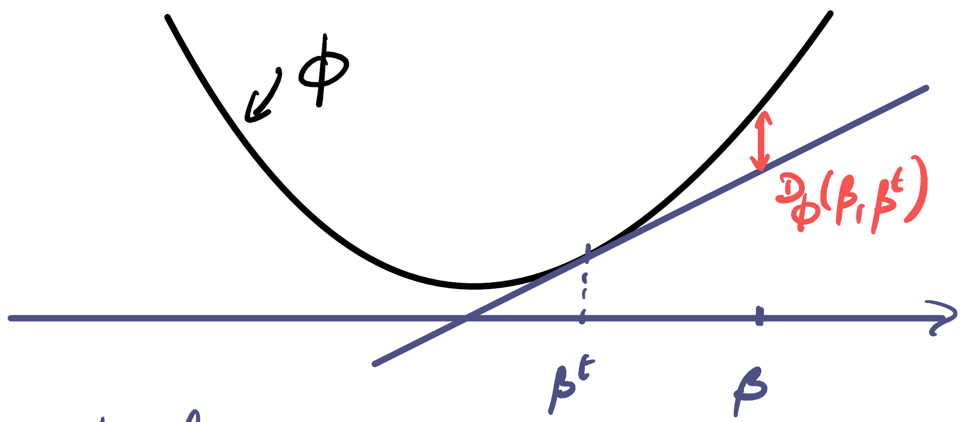
$\rightsquigarrow$ good old gradient descent.

Digression: we can replace $O(\|\beta - \beta^t\|^2)$ by something different.

For example, we can choose

$$D_\Phi(\beta, \beta^t) = \Phi(\beta) - \Phi(\beta^t) - \langle \nabla \Phi(\beta^t), \beta - \beta^t \rangle$$

with Φ convex



and solve

$$\beta^{t+1} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \Big\{ F(\beta^t) + \langle \nabla F(\beta^t), \beta - \beta^t \rangle + \frac{1}{\eta} D_\Phi(\beta, \beta^t) \Big\}$$

solution:

$$\nabla \Phi(\beta_{t+1}) = \nabla \Phi(\beta_t) - \eta \nabla F(\beta_t)$$

(Mirror Descent)

. For solving non-smooth optimisation problems like

$$\hat{\beta} \in \underset{\beta}{\operatorname{argmin}} \{ F(\beta) + \lambda |\beta|_1 \} \quad (*) \quad ?$$

⇝ apply the Taylor approximation to F:

$$\beta^{t+1} \in \underset{\beta}{\operatorname{argmin}} \{ F(\beta^t) + \langle \nabla F(\beta^t), \beta - \beta^t \rangle + \frac{1}{2\eta} \|\beta - \beta^t\|^2 + \lambda |\beta|_1 \}$$

$$= \underset{\beta}{\operatorname{argmin}} \{ \frac{1}{2} \|\beta - (\beta^t - \eta \nabla F(\beta^t))\|^2 + \lambda \eta |\beta|_1 \}$$

$$= S_{\lambda \eta}(\beta^t - \eta \nabla F(\beta^t))$$

⇝ for F convex it solves $(*)$ for η small enough

(can be used to solve the Lasso problem)

where

$$S_\mu(\underbrace{\alpha}_{\in \mathbb{R}^p}) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \{ \frac{1}{2} \|\beta - \alpha\|^2 + \mu |\beta|_1 \}$$

$$= \begin{bmatrix} \alpha_1 \left(1 - \frac{\mu}{|\alpha_1|}\right)_+ \\ \vdots \\ \alpha_p \left(1 - \frac{\mu}{|\alpha_p|}\right)_+ \end{bmatrix} \in \mathbb{R}^p$$

(soft-thresholding operator)

b) Iterative Hard-Thresholding:

The Lasso estimator $\hat{\beta}^{Lasso} \in \arg\min_{\beta} \{ \|Y - X\beta\|^2 + \lambda |\beta|_1 \}$ has been derived as a convex surrogate of $\hat{\beta}^{ns} \in \arg\min_{\beta} \{ \underbrace{\|Y - X\beta\|^2}_{F(\beta)} + \lambda^2 |\beta|_0 \}$ (ns)

Apply proximal optimisation to the (ns) problem:

$$
\begin{aligned}
\beta^{t+1} &\in \arg\min_{\beta} \{ F(\beta^t) + \langle \nabla F(\beta^t), \beta - \beta^t \rangle + \tfrac{1}{2\eta} \|\beta - \beta^t\|^2 + \lambda^2 |\beta|_0 \} \\
&= \arg\min_{\beta} \{ \|\beta - (\beta^t - \eta \nabla F(\beta^t))\|^2 + 2\eta \lambda^2 |\beta|_0 \} \\
&= H_{\lambda\sqrt{2\eta}} (\beta^t - \eta \nabla F(\beta^t))
\end{aligned}
$$

where $H_{\mu}(\underset{\in \mathbb{R}^p}{\alpha}) := \arg\min_{\beta} \{ \|\beta - \alpha\|^2 + \mu^2 |\beta|_0 \}$

$$
= \begin{bmatrix} \alpha_1 \mathbb{1}_{|\alpha_1| > \mu} \\ \vdots \\ \alpha_p \mathbb{1}_{|\alpha_p| > \mu} \end{bmatrix}
$$

} Hard-Thresholding operator.

:) · computationally efficient
· no shrinkage

:(· Since (ns) is non-convex, there is no convergence garantee.

Theory?

for · $\hat{\beta}^0 \equiv 0$
· $\eta = 1/2$ (for simple formulaes)
· a regularisation parameter λ_t decreasing from a high-value to the optimal level
$\lambda_\infty = \underbrace{K}_{\text{constant}} \sigma \sqrt{\log p}$

we have some garantees similar to those of Lasso estimator after $O(\log n)$ steps.

Sketch of analysis: for $\eta = 1/2$

$$\beta^{t+1} = H_{\lambda_{t+1}}\left((I - X^TX)\beta^t + X^TY\right)$$

$$= H_{\lambda_{t+1}}\left(\underbrace{\beta^*}_{\text{target}} + \underbrace{(I - X^TX)(\beta^t - \beta^*)}_{\text{contraction?}} + \underbrace{X^T\varepsilon}_{Z}\right)$$

$Y = X\beta^* + \varepsilon$

Contraction? If

· $\max \|(I - X^TX)u\| \leq \underbrace{(1-\delta)}_{<1} \|u\|$
u sparse

· β^t remains sparse

Then $\|(I - X^TX)\underbrace{(\beta^t - \beta^*)}_{\text{sparse}}\| \leq (1-\delta)\|\beta^t - \beta^*\|$.

So for t large:

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

:)

Benefit of successive hard-thresholding:

- IHT alternates gradient steps and hard-thresholding:

$$\hat{\beta}^{t+1} \leftarrow H_{\lambda_t}(\hat{\beta}^t - \eta \nabla F(\hat{\beta}^t))$$

- Why not simply doing a full gradient descent and then apply hard-thresholding just at the end?

$$\tilde{\beta} \leftarrow H_{\lambda_\infty}(\hat{\beta}^{GD})$$ where $\hat{\beta}^{GD}$ = solution to gradient descent started from 0.

For $F(\beta) = \|Y - X\beta\|^2$: $\hat{\beta}^{GD} = \underbrace{X^+}_{\text{Moore-Penrose pseudo inverse}} Y = \beta^* + X^+ \varepsilon$

We have $\tilde{\beta} = H_{\lambda_\infty}(\beta^* + \underbrace{X^+\varepsilon}_{\mathcal{N}(0,\, X^+(X^+)^T)})$ to be compared to $\hat{\beta}^t \underset{t \text{ large}}{\approx} H_{\lambda_\infty}(\beta^* + \underbrace{X^T\varepsilon}_{\mathcal{N}(0,\, X^TX)})$

$X^+(X^+)^T$ can have a huge operator norm in high-dimension

$|X^TX|_{op} = |X|_{op}^2$

- Successive hard-thresholding has a regularisation effect.

Convex criterion or iterative algorithm?

Let us consider the more complex setting $\underbrace{y^{(i)}}_{\in \mathbb{R}^T} = \underbrace{(A^*)^T}_{\in \mathbb{R}^{p \times T}} \underbrace{x^{(i)}}_{\in \mathbb{R}^p} + \underbrace{\varepsilon^{(i)}}_{\mathcal{N}(0, \sigma^2 I_T)}$, $i = 1, \dots, n$

and assume that

i) **row-sparsity**: only a few row of A^* are non-zero

(since $A^T x = \sum_{j=1}^{p} (A_{j:})^T x_j$, it means that only a few coordinate of x are active)

ii) **Low-rank**: $\text{rank}(A^*)$ is small.

$\rightsquigarrow$ mixture of coordinate-wise and spectral structures

Can we take benefit of Low-rank and row-sparse simultaneously?

- **With model selection**: yes, but prohibitive computational cost.

Benchmark: if $r^* = \text{rank}(A^*)$ and $k^* = \text{card}\{j : A^*_{j:} \neq 0\}$, then

$$\mathbb{E}\left[\|X\hat{A}^{(ns)} - XA^*\|_F^2\right] \leq c\left(\underbrace{r^*(T+k^*)}_{\substack{\text{low rank} \\ \text{with } k^* \text{ rows}}} + \underbrace{k^* \log \frac{ep}{k^*}}_{\substack{\text{complexity of rows} \\ \text{identification}}}\right)\sigma^2 \qquad \text{(Theorem 8.7)}$$

where

$$\hat{A}^{(ns)} \simeq \underset{A \in \mathbb{R}^{p \times T}}{\text{argmin}} \left\{ \|Y - XA\|_F^2 + \lambda \,\text{card}\{j : A_{j:} \neq 0\} + \mu \,\text{rank}(A) \right\}$$

with λ, μ well-chosen

convex-relaxation ?

- $\text{card}\{j : A_{j:} \neq 0\} = \sum_{j=1}^{p} \mathbb{1}_{\|A_{j:}\| \neq 0} \rightsquigarrow \sum_{j=1}^{p} \|A_{j:}\|$ (group lasso)

- $\text{rank}(A) = \sum_k \mathbb{1}_{\sigma_k(A) \neq 0} \rightsquigarrow |A|_* = \sum_k \sigma_k(A)$ (nuclear norm)

$$\hat{A}^{cvx} \in \underset{A \in \mathbb{R}^{p \times T}}{\text{argmin}} \left\{ \underbrace{\|Y - XA\|_F^2 + \lambda \sum_{j=1}^{p} \|A_{j:}\| + \mu |A|_*}_{\text{convex :)}} \right\}$$

$$\hat{A}^{cvx} \in \underset{A \in \mathbb{R}^{p \times T}}{\arg\min} \left\{ \|Y - XA\|_F^2 + \lambda \sum_{j=1}^{p} \|A_{j:}\| + \mu \|A\|_* \right\}$$

it can be computed (ADMM)

no improvement compared to low rank or row-sparse alone
(proved for a similar problem)

why?

⇝ bias cumulate...

Iterative algorithm ?

- **Idea 1**: decompose $A = UV$ with $U \in \mathbb{R}^{p \times r}$ and $V \in \mathbb{R}^{r \times T}$. ($\rightsquigarrow$ rank $\leq r$)

Problem: $UV = (\alpha U)(\frac{1}{\alpha} V) \rightsquigarrow$ size of U and V must be stabilized.

- **Idea 2**: consider

$$F(U,V) = \underbrace{\| Y - X \overset{A}{\overbrace{UV}} \|_F^2}_{\text{data fit}} + \underbrace{\frac{1}{2} \| U^T U - V V^T \|_F^2}_{\text{scale stabilization}}$$

- **Idea 3**: proximal iterations related to

$$\min_{U,V} F(U,V) + \lambda \sum_j \mathbb{1}_{\|U_{j:}\| \neq 0} :$$

$$\begin{bmatrix} U^{t+1} \\ V^{t+1} \end{bmatrix} \leftarrow \begin{bmatrix} H_\lambda^G (U^t - 2 \nabla_U F(U^t, V^t)) \\ V^t - 2 \nabla_V F(U^t, V^t) \end{bmatrix}$$

with H_λ^G = group thresholding operator

(set to 0 rows with $\|U_{j:}\|^2 > \lambda$)

Theorem (informal)

Under some assumptions (unavoidable), for t large enough ($\simeq c \log n$), we have with high probability

$$\| X U^t V^t - X A^* \|_F^2 \lesssim (r^*(T + k^*) + k^* \log p) \sigma^2$$

benchmark for $\hat{A}^{(ns)}$

Take Home Messages:

→ iterative algorithms have received a renewed interest in recent years as

- they are computationally efficient
- they do not suffer from shrinkage bias
- they can solve some problems where convex penalisation fails

→ Unlike the classical analysis, the optimisation analysis and the statistical analysis cannot be treated apart:

Classical analysis: For $\hat{\beta}^t \to \hat{\beta}^\infty$ (for example $\hat{\beta}^\infty = \hat{\beta}^{lasso}$)

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

analysed apart

Iterative algorithms:

$\hat{\beta}^t \not\to \dots$: no convergence known, so we cannot split apart the analysis. We must prove a "contraction" at each step.