# Computation time/accuracy trade-off and linear regression

Christophe BIERNACKI & Maxime BRUNIN & Alain CELISSE

Laboratoire Paul Painlevé, Université de Lille, Science et Technologie
INRIA Lille-Nord Europe, MODAL team

9 december 2016

# Outline

# Linear regression

Linear regression model

$$Y = X\theta^* + \epsilon,$$

with $X \in \mathcal{M}_{n,d}(\mathbb{R})$ with $\text{rg}(X) = d$ ; $\theta^* \in \mathbb{R}^d$ is unknown ; $\epsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right)$.

We are in the case $n > d$. We usually use the Ordinary Least Squares (OLS) $\hat{\theta}$ to estimate $\theta^*$.

**Goal**

Find an estimator that performs better in terms of MSE or/and in terms of computation time than $\hat{\theta}$.

# Gradient descent algorithm

Using the approach (Raskutti, Wainwright, and Yu 2014), we use a gradient descent algorithm with fixed step $\alpha$ to minimize the convex and differentiable function $g(\theta) = \frac{1}{2n} \|Y - X\theta\|_2^2$ for $\theta \in \mathbb{R}^d$. We get closed formulas of $\hat{\theta}^{(k)}$, the estimator of $\theta^*$ at the iteration $k$, and $\hat{Y}^{(k)} = X\hat{\theta}^{(k)}$.

$\forall k \geq 0$,

$$\begin{aligned}\hat{\theta}^{(k)} &= \sum_{i=0}^{k-1} \left(I_d - \frac{\alpha}{n} X^T X\right)^i \left(\hat{\theta}^{(1)} - \theta_0\right) + \theta_0 \\ &= \left(I_d - \left(I_d - \frac{\alpha}{n} X^T X\right)^k\right) \hat{\theta} + \left(I_d - \frac{\alpha}{n} X^T X\right)^k \theta_0.\end{aligned}$$

$\forall k \geq 0$,

$$\hat{Y}^{(k)} = \left(I_n - \left(I_n - \frac{\alpha}{n} X X^T\right)^k\right) \hat{Y} + \left(I_n - \frac{\alpha}{n} X X^T\right)^k Y^{(0)}.$$

# Accuracy of our estimator

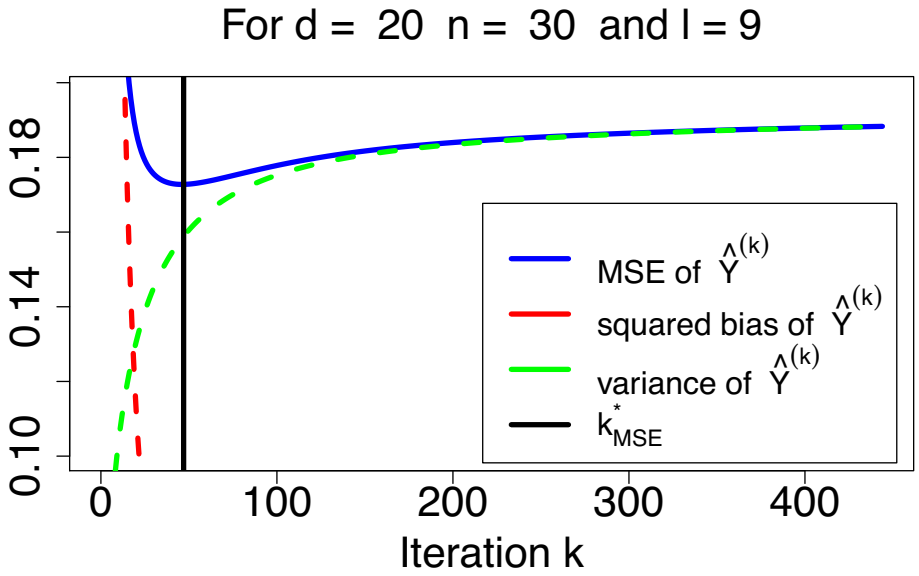
We assess the accuracy of $\hat{\theta}^{(k)}$ by $\begin{cases} \Delta\left(\hat{Y}^{(k)}\right) = \frac{1}{n}\left\|\hat{Y}^{(k)} - Y^*\right\|_2^2. \\ \text{or} \\ \text{MSE}\left(\hat{Y}^{(k)}\right) = E\left[\Delta\left(\hat{Y}^{(k)}\right)\right]. \end{cases}$

with $Y^* = X\theta^*$.

## Property

$\text{MSE}\left(\hat{Y}^{(k^*_{\text{MSE}})}\right) < \text{MSE}\left(\hat{Y}\right)$ *with* $k^*_{\text{MSE}} = \underset{k\in\mathbb{N}}{\text{argmin}}\left\{\text{MSE}\left(\hat{Y}^{(k)}\right)\right\}$ *and* $\hat{Y} = X\hat{\theta}$.

# Trade-off bias variance (1/2)

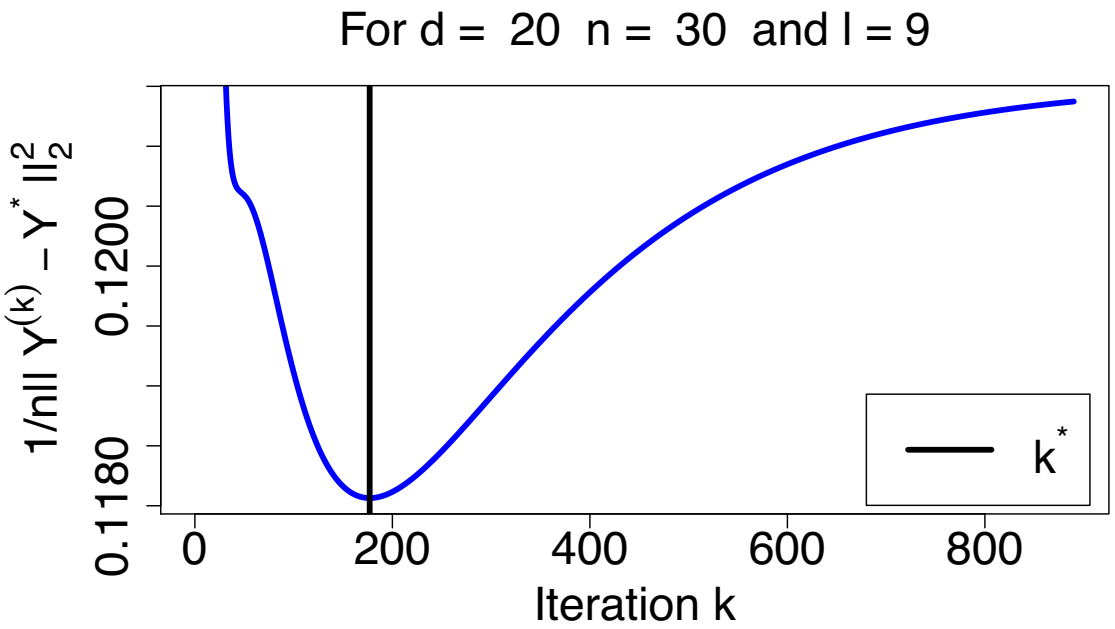



$k^*_{\mathrm{MSE}}$ occurs when squared bias of $\hat{Y}^{(k)} \neq$ variance of $\hat{Y}^{(k)} \Rightarrow$ we have to control the ratio between the variance of $\hat{Y}^{(k)}$ and the squared bias of $\hat{Y}^{(k)}$ to estimate $k^*_{\mathrm{MSE}}$.

$$\mathrm{MSE}\left(\hat{Y}^{(k)}\right) = \frac{1}{n}\left\|S^k P^T\left(Y^{(0)} - Y^*\right)\right\|_2^2 + \frac{\sigma^2}{n}\mathrm{Tr}\left(\left(I_n - S^k\right)^2\right),$$

with $K = \frac{1}{n}XX^T = P\Lambda P^T$ ; $S = I_n - \alpha\Lambda$ ; $0 < \alpha < \min\left\{1, \frac{1}{\hat{\lambda}_1}\right\}$ ; $\hat{\lambda}_1 = \|K\|_2$.

# Trade-off bias variance (2/2)




with $\alpha = l/10 \times \min\left\{1, \frac{1}{\hat{\lambda}_1}\right\}$ ; $k^* = \underset{k \in \mathbb{N}}{\operatorname{argmin}} \left\{ \frac{1}{n} \left\| \hat{Y}^{(k)} - Y^* \right\|_2^2 \right\}$.

# Stopping time to estimate $k^* = \underset{k \in \mathbb{N}}{\operatorname{argmin}} \left\{ \frac{1}{n} \left\| \hat{Y}^{(k)} - Y^* \right\|_2^2 \right\}$

We study the random variable $\frac{1}{n} \left\| \hat{Y}^{(k)} - Y^* \right\|_2^2$ $(Y^* = X\theta^*)$.

$\forall k \geq 0$,

$$\frac{1}{n} \left\| \hat{Y}^{(k)} - Y^* \right\|_2^2 \leq \underbrace{\frac{2}{n} \left\| E\left[ \hat{Y}^{(k)} \right] - Y^* \right\|_2^2}_{B_k^2} + \underbrace{\frac{2}{n} \left\| \hat{Y}^{(k)} - E\left[ \hat{Y}^{(k)} \right] \right\|_2^2}_{V_k}.$$

**Lemma (Raskutti, Wainwright, and Yu)**

*If* $\|\theta^*\|_2 \leq 1$ *and* $\theta_0 = 0$, $\forall k \geq 1$,

$$B_k^2 \leq \frac{1}{\mathrm{e} k \alpha} = B_k^{2,sup}.$$

**Lemma**

*On an event* $\mathcal{A}_q$ *with high probability,* $\forall k \in [\![1, k_{\hat{\sigma}}]\!]$,

$$V_k \leq 5\sigma^2 k\alpha \left[ R_K\left( \frac{1}{\sqrt{k\alpha}} \right) \right]^2 = V_k^{sup}.$$

# Stopping time to estimate $k^* = \underset{k\in\mathbb{N}}{\operatorname{argmin}} \left\{ \frac{1}{n} \left\| \hat{Y}^{(k)} - Y^* \right\|_2^2 \right\}$

To control the ratio between $V_k^{\sup}$ and $B_k^{2,\sup}$, we define the stopping time $k_{\hat{\sigma}}$

$$k_{\hat{\sigma}} = \min \left\{ k \in \mathbb{N} : R_K\left(\frac{1}{\sqrt{k\alpha}}\right) > \sqrt{\frac{c}{5e}} \frac{1}{\hat{\sigma} k \alpha} \right\} - 1$$
$$= \min \left\{ k \in \mathbb{N} : V_k^{\sup} > c \left(\frac{\sigma}{\hat{\sigma}}\right)^2 B_k^{2,\sup} \right\} - 1.$$

We define $\varepsilon_\sigma$

$$\varepsilon_\sigma = \inf \left\{ \varepsilon > 0 : R_K(\varepsilon) \leq \sqrt{\frac{c}{5e}} \frac{\varepsilon^2}{\sigma} \right\}.$$

with:

- $\hat{\sigma}^2 = \frac{1}{n-d} \left\| Y - X\hat{\theta} \right\|_2^2$.
- $R_K(\varepsilon) = \sqrt{\frac{1}{n} \sum_{i=1}^n \min\left\{ \hat{\lambda}_i, \varepsilon^2 \right\}}$ ; $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \hat{\lambda}_n \geq 0$ are the eigenvalues of $\frac{1}{n} X X^T$.
- $\Omega_q = \{ | \hat{\sigma} - \sigma | \leq q\sigma \} = \{ \sigma_{\text{inf}} \leq \hat{\sigma} \leq \sigma_{\text{sup}} \}$ ; $\sigma_{\text{inf}} = (1-q)\sigma$ and $\sigma_{\text{sup}} = (1+q)\sigma$.

# Theorem

## Theorem (transposition of theorem 1 (Raskutti, Wainwright, and Yu 2014))

*Given the stopping time* $k_{\hat{\sigma}}$, $\exists c_1, c_2 \in \mathbb{R}_+^*$ *such on an event* $\mathcal{A}_q$ :
$P(\mathcal{A}_q) \geq P(\Omega_q) - c_1 \exp\left(-c_2 n \varepsilon_{\sigma_{\textbf{inf}}}^4\right)$ ( $\Omega_q = \{|\hat{\sigma} - \sigma| \leq q\sigma\}$).

(a) $\forall k \in [\![1, k_{\hat{\sigma}}]\!]$,

$$\frac{1}{n}\left\|\hat{Y}^{(k)} - Y^*\right\|_2^2 \leq B_k^{2,\textbf{sup}} + V_k^{\textbf{sup}} \leq \frac{1}{\mathrm{e}k\alpha}\left(1 + c\left(\frac{\sigma}{\hat{\sigma}}\right)^2\right).$$

(b) $\forall k \geq 0$ *(interesting when* $k > k_{\hat{\sigma}}$*)*,

$$E\left[\frac{1}{n}\left\|\hat{Y} - Y^*\right\|_2^2\right] \geq E\left[\frac{1}{n}\left\|\hat{Y}^{(k)} - Y^*\right\|_2^2\right] \geq \frac{\sigma^2}{4}\underbrace{(k\alpha)^2\left(R_K\left(\frac{1}{\sqrt{k\alpha}}\right)\right)^4}_{f(k)},$$

*with* $f$ *is a non-decreasing function and* $f(k) \underset{k\to+\infty}{\longrightarrow} 1$.

# Comparison of the complexity in time of $\hat{\theta}^{(k_{\hat{\sigma}})}$ and $\hat{\theta}$

The complexity in time of $\hat{\theta}^{(k_{\hat{\sigma}})}$ is

$$\underbrace{nd^2 + 2nd + O\left(d^3\right) + 3d^2 + (k+1)d}_{\hat{\theta}^{(k)}} + \underbrace{nd + 3d^2 + n + d}_{\hat{\sigma}^2} + \underbrace{\left[1 + \frac{\log(\overline{k})}{\log(2)}\right] d}_{k_{\hat{\sigma}}}$$

$$= nd^2 + 3nd + O\left(d^3\right) + 6d^2 + n + \left[k + \frac{\log(\overline{k})}{\log(2)} + 3\right] d,$$

with $k_{\hat{\sigma}} \leq \overline{k}$ and $\hat{\sigma}^2 = \frac{1}{n-d}\left\|Y - X\hat{\theta}\right\|_2^2$.

The complexity in time of $\left(\hat{\theta}, \hat{\sigma}^2\right)$ is $\underbrace{nd^2 + 2nd + O\left(d^3\right) + d^2}_{\hat{\theta}} + \underbrace{nd + n}_{\hat{\sigma}^2}$.

The complexity in time of $\left(\hat{\theta}^{(k_{\hat{\sigma}})}, \hat{\sigma}^2\right)$ is bigger than the complexity in time of $\left(\hat{\theta}, \hat{\sigma}^2\right)$.

# Comparison of the complexity in space of $\hat{\theta}^{(k_{\hat{\sigma}})}$ and $\hat{\theta}$

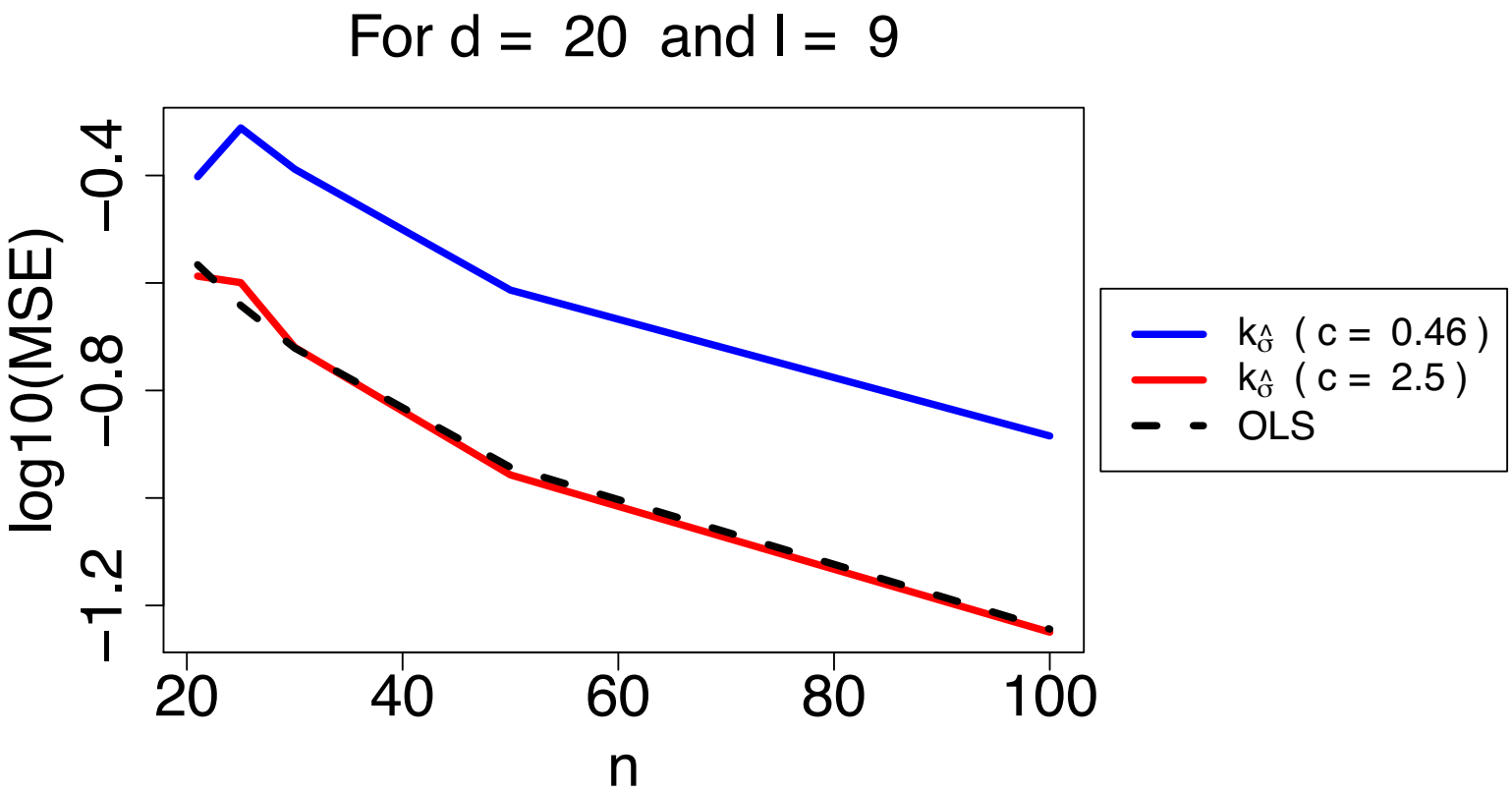
The complexity in space of $\hat{\theta}^{(k_{\hat{\sigma}})}$ is

$$\underbrace{nd + 3d^2 + 6d}_{\hat{\theta}^{(k)}} + \underbrace{d^2 + 3d + n + 1}_{\hat{\sigma}^2} + \underbrace{2d}_{k_{\hat{\sigma}}}$$
$$= nd + 4d^2 + n + 11d + 1.$$

The complexity in space of $\left(\hat{\theta}, \hat{\sigma}^2\right)$ is $\underbrace{nd + 2d^2 + 2d}_{\hat{\theta}} + \underbrace{n + 1}_{\hat{\sigma}^2}$.

The complexity in space of $\left(\hat{\theta}^{(k_{\hat{\sigma}})}, \hat{\sigma}^2\right)$ is bigger than the complexity in space of $\left(\hat{\theta}, \hat{\sigma}^2\right)$.

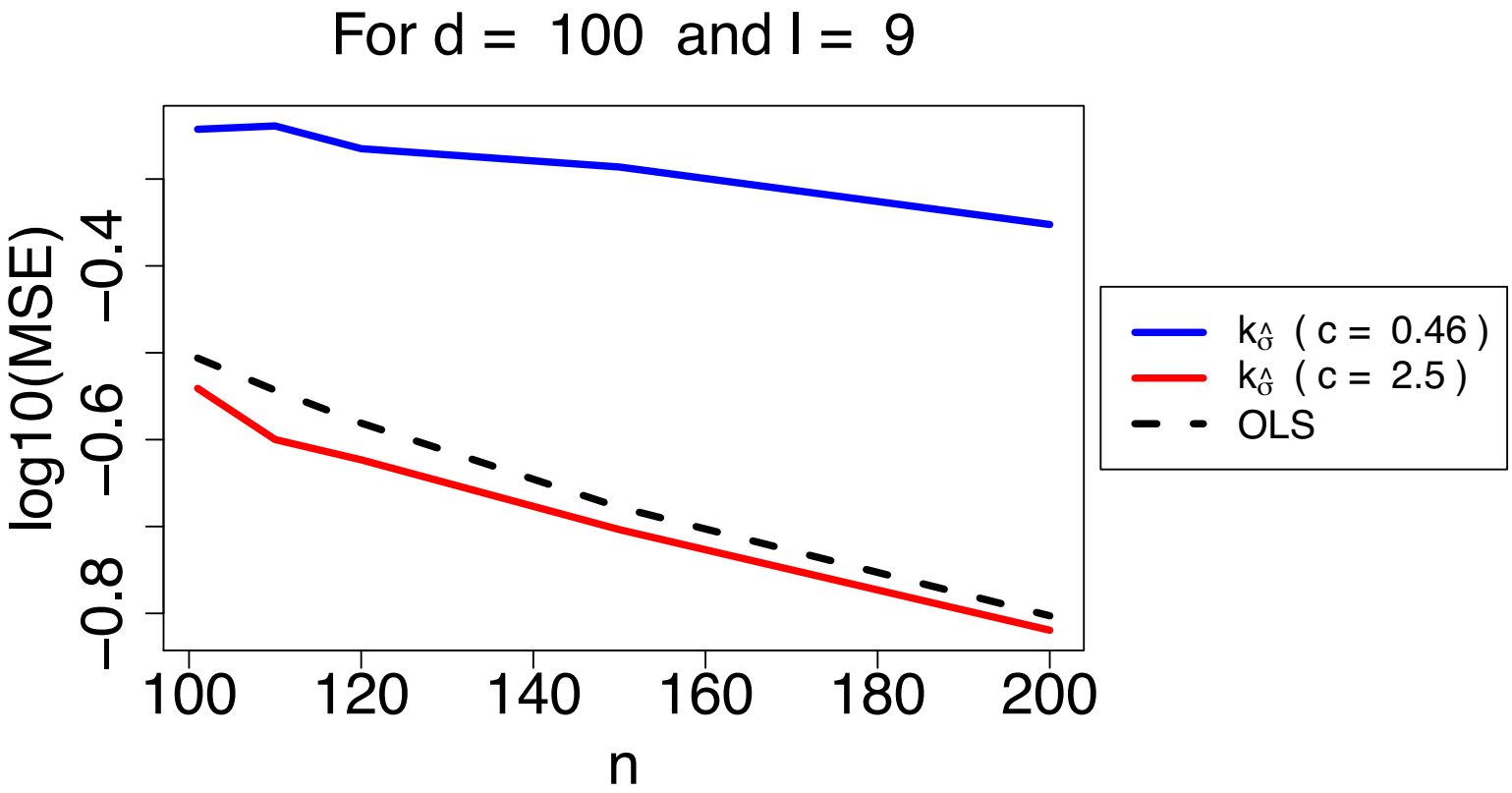
# MSE as a function of $n$ for $d = 20$



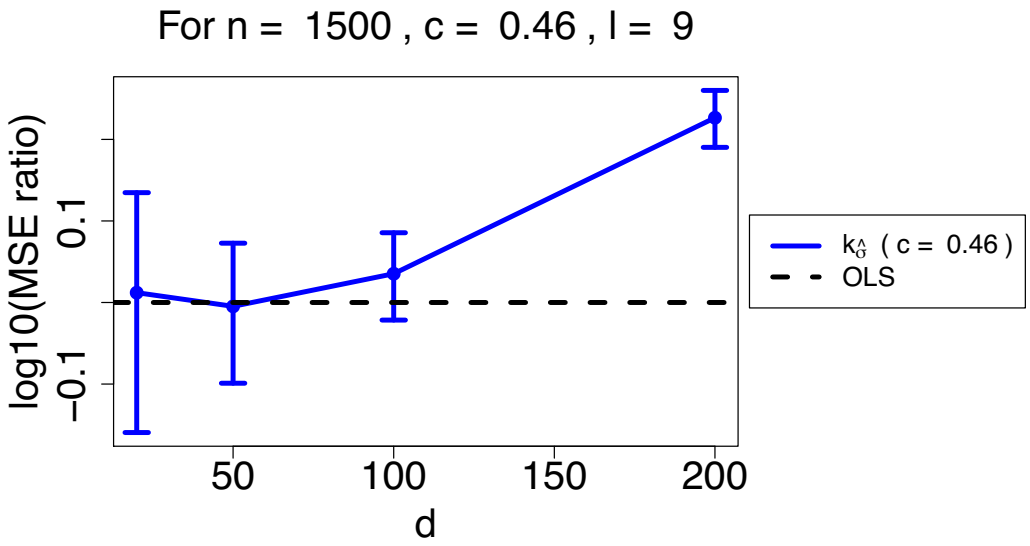

with $\alpha = l/10 \times \min\left\{1, \frac{1}{\hat{\lambda}_1}\right\}$.

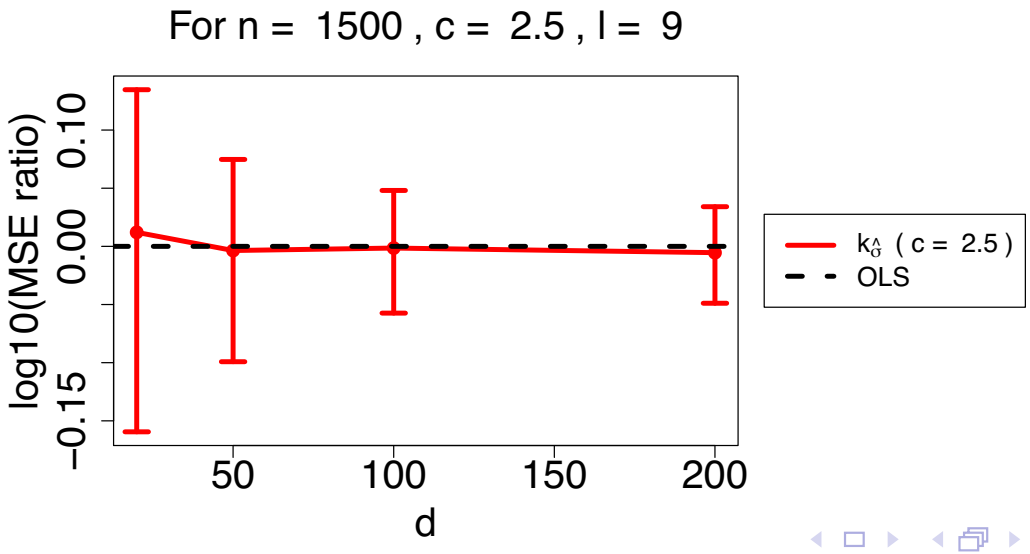
# MSE as a function of $n$ for $d = 100$

Raskutti, Wainwright, and Yu chooses $c = 0.46$ but it is not optimal.

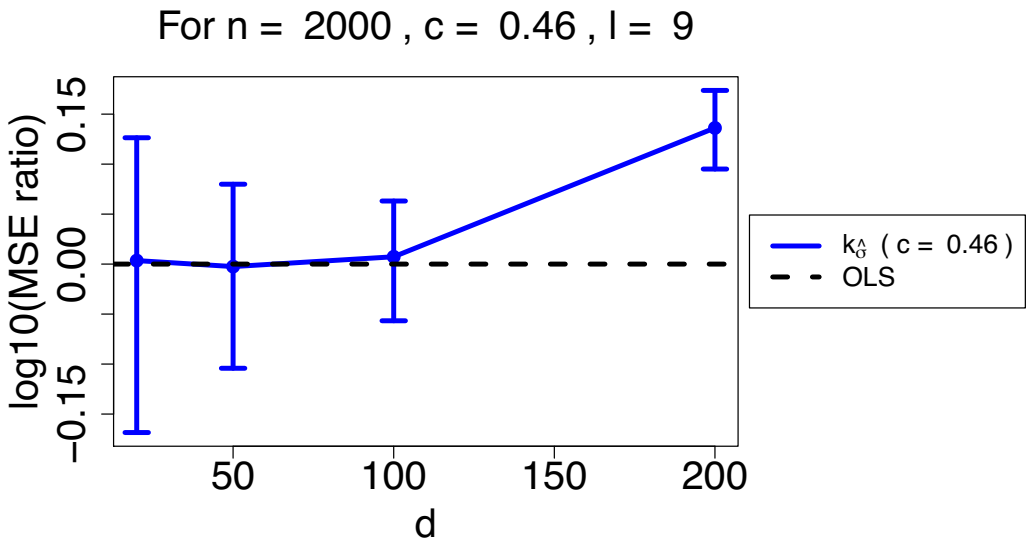


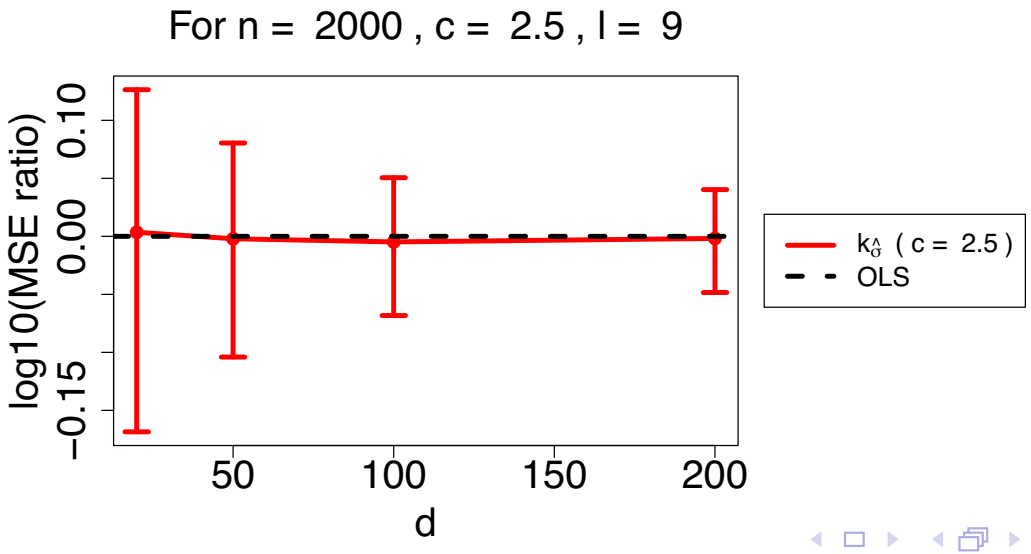
# MSE ratio as a function of $d$ for $n = 1500$



# MSE ratio as a function of $d$ for $n = 2000$

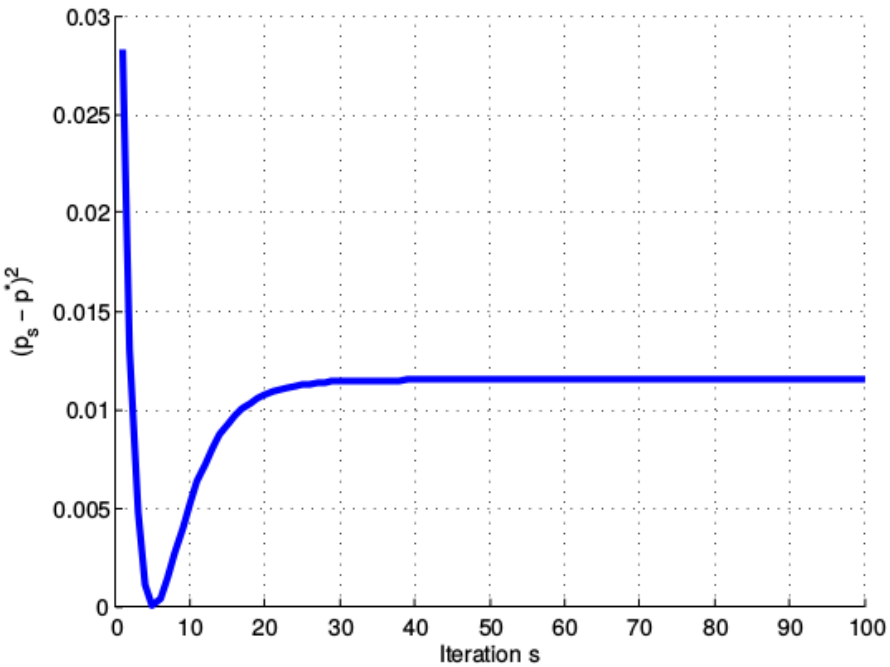


# General use of stopping time

Stopping time enables to reduce computation time (without loosing in accuracy) in problems when $\hat{\theta}$ has no closed formula and needs a lot of iterations to be computed.

For instance, in the problem of two Gaussian univariate mixture where only the proportion $p^*$ is unknown, we use a EM whose estimate at the $s^{th}$ iteration is $\hat{p}^{(s)}$ and $\hat{p}^{(s)} \underset{s \to +\infty}{\longrightarrow} \hat{p}$.



# Conclusion

**Results**:

- We prove that $\frac{1}{n}\left\|\hat{Y}^{(k)} - Y^*\right\|_2^2$ decreases at least until $k_{\hat{\sigma}}$.
- We should choose $c > 1$ to ensure $\mathrm{MSE}\left(\hat{Y}^{(k_{\hat{\sigma}})}\right) < \mathrm{MSE}\left(\hat{Y}\right)$.

**Drawback**:

The complexity in time and in space of $\hat{\theta}^{(k_{\hat{\sigma}})}$ is bigger than the complexity in time and in space of $\hat{\theta}$ (specific to linear regression).

**Perspectives**:

- Find how to choose parameter $c$ to ensure $\mathrm{MSE}\left(\hat{Y}^{(k_{\hat{\sigma}})}\right) < \mathrm{MSE}\left(\hat{Y}\right)$.
- Prove the theorem in other problems (e.g. logistic regression).

# Complexity of $\hat{\theta}^{(k_{\hat{\sigma}})}$

We compute $\hat{\theta}^{(k)}$ using the formula

$$\begin{aligned}\hat{\theta}^{(k)} &= \sum_{i=0}^{k-1} \left(I_d - \frac{\alpha}{n} X^T X\right)^i \left(\hat{\theta}^{(1)} - \theta_0\right) + \theta_0 \\ &= UQU^T \left(-\frac{\alpha}{n} X^T X \theta_0 + \frac{\alpha}{n} X^T Y\right) + \theta_0,\end{aligned}$$

with $X^T X = U D_3 U^T$ ; $Q = \mathrm{Diag}\left(\frac{1-\mu_d^k}{1-\mu_d}, \ldots, \frac{1-\mu_1^k}{1-\mu_1}\right)$ ;
$1 > \mu_1 \geq \mu_2 \geq \cdots \geq \mu_d > 0$ are the eigenvalues of $V = I_d - \frac{\alpha}{n} X^T X$.

# Complexity of $\hat{\theta}^{(k_{\hat{\sigma}})}$

The complexity of $\hat{\theta}^{(k)}$ is

| Terms/Complexity | Time | Space |
|---|---|---|
| $X^T$ | $nd$ | $nd$ |
| $X^T X$ | $nd^2$ | $d^2$ |
| SVD of $X^T X$ | $O(d^3)$ | $d^2 + d$ |
| $X^T Y$ | $nd$ | $d$ |
| $-\frac{\alpha}{n} X^T X \theta_0 + \frac{\alpha}{n} X^T Y$ | $d$ | $d$ |
| $Q \times U^T$ | $d^2$ | $d^2$ |
| $QU^T \times \left(-\frac{\alpha}{n} X^T X \theta_0 + \frac{\alpha}{n} X^T Y\right)$ | $d^2$ | $d$ |
| $Q$ | $kd$ | $d$ |
| $U \times QU^T \left(-\frac{\alpha}{n} X^T X \theta_0 + \frac{\alpha}{n} X^T Y\right)$ | $d^2$ | $d$ |

The complexity of $\hat{\theta}^{(k)}$ is

- in time: $nd^2 + 2nd + O(d^3) + 3d^2 + (k+1)d$.
- in space: $nd + 3d^2 + 6d$.

# Complexity of $\hat{\theta}^{(k_{\hat{\sigma}})}$

The complexity of $k_\sigma$ is $\left[1 + \frac{\log(\bar{k})}{\log(2)}\right] d$ in time and $2d$ in space.

After the computation of $\hat{\theta}^{(k)}$, the complexity of $\hat{\theta}$ is

| Terms/Complexity | Time | Space |
|---|---|---|
| $D_3^{-1}$ | $d$ | $d$ |
| $D_3^{-1} \times U^T$ | $d^2$ | $d^2$ |
| $D_3^{-1} U^T \times X^T Y$ | $d^2$ | $d$ |
| $U \times D_3^{-1} U^T X^T Y$ | $d^2$ | $d$ |

with $X^T X = U D_3 U^T$; $k_\sigma \leq \bar{k}$.

After the computation of $\hat{\theta}^{(k)}$, the complexity of $\hat{\theta}$ is

- in time: $3d^2 + d$.
- in space: $d^2 + 3d$.

# Complexity of $\hat{\theta}^{(k_{\hat{\sigma}})}$

The complexity of $\hat{\sigma}^2$ $\left(\hat{\sigma}^2 = \frac{1}{n-d}\left\|Y - X\hat{\theta}\right\|_2^2\right)$ is

| Terms/Complexity | Time | Space |
|---|---|---|
| $X\hat{\theta}$ | $nd$ | $n$ |
| $\left\|Y - X\hat{\theta}\right\|_2^2$ | $n$ | 1 |

The complexity of $\hat{\sigma}^2$ is $nd + n$ in time and $n + 1$ in space.

# Complexity of $\hat{\theta}$

The complexity of $\hat{\theta}$ is

| Terms/Complexity | Time | Space |
|---|---|---|
| $X^T$ | $nd$ | $nd$ |
| $X^T X$ | $nd^2$ | $d^2$ |
| $(X^T X)^{-1}$ | $O(d^3)$ | $d^2$ |
| $X^T Y$ | $nd$ | $d$ |
| $(X^T X)^{-1} \times X^T Y$ | $d^2$ | $d$ |

The complexity of $\hat{\theta}$ is

- in time: $nd^2 + 2nd + O(d^3) + d^2$.
- in space: $nd + 2d^2 + 2d$.

# Bibliography

Raskutti, Garvesh, Martin J. Wainwright, and Bin Yu (2014). "Early Stopping and Non-Parametric Regression: An Optimal Data-Dependent Stopping Rule". In: *J. Mach. Learn. Res.* 15.1, pp. 335–366. ISSN: 1532-4435.