Adaptivity of early stopping for PLS / CG[1]

Markus Reiß

Institut für Mathematik
Humboldt-Universität zu Berlin
www.mathematik.hu-berlin.de/∼mreiss

Luminy, 20 December 2017

[1] PLS: partial least squares / CG: conjugate gradients

Least squares

"White" Gaussian linear model:

$$\boxed{Y = X\beta + \varepsilon}$$

- $\beta \in \mathbb{R}^D$, $X \in \mathbb{R}^{n \times D}$, $\varepsilon \sim N(0, \sigma^2 E_n)$
- Least squares: $\hat{\beta}^{LS} = (X^\top X)^{-1} X^\top Y$
- $D \leqslant n$, $\lambda_1 \geqslant \cdots \geqslant \lambda_D > 0$ singular values of X (eigenvalues of $(X^\top X)^{1/2}$)
- Errors (conditional on X):

$$\boxed{\mathbb{E}[\|\hat{\beta}^{LS} - \beta\|^2] = \sigma^2 \textstyle\sum_{i=1}^{D} \lambda_i^{-2}} \quad \text{(strong norm error)}$$

$$\boxed{\mathbb{E}[\|X(\hat{\beta}^{LS} - \beta)\|^2] = \sigma^2 D} \quad \text{(prediction / weak norm error)}$$

Dimension reduction: PCR

Principal component regression: (Kendall 1957, Hotelling 1957)
$\hat{\beta}_m$: projection of $\hat{\beta}$ onto span$(v_1, \ldots, v_m)$
(v_i singular vectors of X)

$$\boxed{\mathbb{E}[\|\hat{\beta}_m - \beta\|^2] = \|\beta_m - \beta\|^2 + \sigma^2 \sum_{i=1}^{m} \lambda_i^{-2}} \quad \text{(strong error)}$$

$$\boxed{\mathbb{E}[\|X(\hat{\beta}_m - \beta)\|^2] = \|X(\beta_m - \beta)\|^2 + \sigma^2 m} \quad \text{(weak error)}$$

Critique:
The principal directions v_i of $X^\top X$ bear no information about their significance for predicting Y. (Jolliffe 1982)

Dimension reduction: PCR

Principal component regression: (Kendall 1957, Hotelling 1957)
$\hat{\beta}_m$: projection of $\hat{\beta}$ onto span$(v_1, \ldots, v_m)$
(v_i singular vectors of X)

$$\boxed{\mathbb{E}[\|\hat{\beta}_m - \beta\|^2] = \|\beta_m - \beta\|^2 + \sigma^2 \sum_{i=1}^m \lambda_i^{-2}} \quad \text{(strong error)}$$

$$\boxed{\mathbb{E}[\|X(\hat{\beta}_m - \beta)\|^2] = \|X(\beta_m - \beta)\|^2 + \sigma^2 m} \quad \text{(weak error)}$$

Critique:
The principal directions v_i of $X^\top X$ bear no information about their significance for predicting Y. (Jolliffe 1982)

Dimension reduction: PLS

Partial least squares: (Wold 1982, Rosipal, Krämer 2006)

$$\boxed{X^\top Y = (X^\top X)\beta + X^\top \varepsilon}$$

- $\hat{\beta}_0 := 0$
- $\hat{\beta}_1 := \alpha_1 X^\top Y$ such that $R_1^2 := \|Y - X\hat{\beta}_1\|^2 \to \min_{\alpha_1}!$
  Solution: $\alpha_1 = \|X^\top Y\|^2 / \|XX^\top Y\|^2$
  Structure: $X\hat{\beta}_1 = p_1(XX^\top)Y$, p_1 linear with Y-dependent coefficient
  Note: X orthogonal $\Rightarrow \hat{\beta}_1 = X^\top Y = \beta + X^\top \varepsilon = \hat{\beta}^{LS}$
- Update: $Y^{(-1)} := Y - X\hat{\beta}_1$
- $\hat{\beta}_2 := \hat{\beta}_1 + \alpha_2 X^\top Y^{(-1)}$ s.t. $R_2^2 := \|Y - X\hat{\beta}_2\|^2 \to \min_{\alpha_2}!$
  Structure: $X\hat{\beta}_2 = p_2(XX^\top)Y$, $p_2 \in \mathrm{Pol}_{2,0}$ quadratic polynomial with $p_2(0) = 0$ and random coefficients

  $\vdots$
- $\hat{\beta}_k := \hat{\beta}_{k-1} + \alpha_k X^\top Y^{(-(k-1))}$, $R_k^2 := \|Y - X\hat{\beta}_k\|^2 \to \min_{\alpha_k}!$
  Structure: $X\hat{\beta}_k = p_k(XX^\top)Y$, $p_k \in \mathrm{Pol}_{k,0}$

Dimension reduction: PLS

Partial least squares: (Wold 1982, Rosipal, Krämer 2006)

$$\boxed{X^\top Y = (X^\top X)\beta + X^\top \varepsilon}$$

- $\hat{\beta}_0 := 0$
- $\hat{\beta}_1 := \alpha_1 X^\top Y$ such that $R_1^2 := \|Y - X\hat{\beta}_1\|^2 \to \min_{\alpha_1}!$
  Solution: $\alpha_1 = \|X^\top Y\|^2 / \|XX^\top Y\|^2$
  Structure: $X\hat{\beta}_1 = p_1(XX^\top)Y$, p_1 linear with Y-dependent coefficient
  Note: X orthogonal $\Rightarrow \hat{\beta}_1 = X^\top Y = \beta + X^\top \varepsilon = \hat{\beta}^{LS}$
- Update: $Y^{(-1)} := Y - X\hat{\beta}_1$
- $\hat{\beta}_2 := \hat{\beta}_1 + \alpha_2 X^\top Y^{(-1)}$ s.t. $R_2^2 := \|Y - X\hat{\beta}_2\|^2 \to \min_{\alpha_2}!$
  Structure: $X\hat{\beta}_2 = p_2(XX^\top)Y$, $p_2 \in \mathrm{Pol}_{2,0}$ quadratic polynomial with $p_2(0) = 0$ and random coefficients

  $\vdots$
- $\hat{\beta}_k := \hat{\beta}_{k-1} + \alpha_k X^\top Y^{(-(k-1))}$, $R_k^2 := \|Y - X\hat{\beta}_k\|^2 \to \min_{\alpha_k}!$
  Structure: $X\hat{\beta}_k = p_k(XX^\top)Y$, $p_k \in \mathrm{Pol}_{k,0}$

Dimension reduction: PLS

Partial least squares: (Wold 1982, Rosipal, Krämer 2006)

$$\boxed{X^\top Y = (X^\top X)\beta + X^\top \varepsilon}$$

- $\hat{\beta}_0 := 0$
- $\hat{\beta}_1 := \alpha_1 X^\top Y$ such that $R_1^2 := \|Y - X\hat{\beta}_1\|^2 \to \min_{\alpha_1}$!
  Solution: $\alpha_1 = \|X^\top Y\|^2 / \|XX^\top Y\|^2$
  Structure: $X\hat{\beta}_1 = p_1(XX^\top)Y$, p_1 linear with Y-dependent coefficient
  Note: X orthogonal $\Rightarrow \hat{\beta}_1 = X^\top Y = \beta + X^\top \varepsilon = \hat{\beta}^{LS}$
- Update: $Y^{(-1)} := Y - X\hat{\beta}_1$
- $\hat{\beta}_2 := \hat{\beta}_1 + \alpha_2 X^\top Y^{(-1)}$ s.t. $R_2^2 := \|Y - X\hat{\beta}_2\|^2 \to \min_{\alpha_2}$!
  Structure: $X\hat{\beta}_2 = p_2(XX^\top)Y$, $p_2 \in \mathrm{Pol}_{2,0}$ quadratic polynomial with $p_2(0) = 0$ and random coefficients

  $\vdots$
- $\hat{\beta}_k := \hat{\beta}_{k-1} + \alpha_k X^\top Y^{(-(k-1))}$, $R_k^2 := \|Y - X\hat{\beta}_k\|^2 \to \min_{\alpha_k}$!
  Structure: $X\hat{\beta}_k = p_k(XX^\top)Y$, $p_k \in \mathrm{Pol}_{k,0}$

Dimension reduction: PLS

Partial least squares: (Wold 1982, Rosipal, Krämer 2006)

$$\boxed{X^\top Y = (X^\top X)\beta + X^\top \varepsilon}$$

- $\hat{\beta}_0 := 0$
- $\hat{\beta}_1 := \alpha_1 X^\top Y$ such that $R_1^2 := \|Y - X\hat{\beta}_1\|^2 \to \min_{\alpha_1}$!
  Solution: $\alpha_1 = \|X^\top Y\|^2 / \|XX^\top Y\|^2$
  Structure: $X\hat{\beta}_1 = p_1(XX^\top)Y$, p_1 linear with Y-dependent coefficient
  Note: X orthogonal $\Rightarrow \hat{\beta}_1 = X^\top Y = \beta + X^\top \varepsilon = \hat{\beta}^{LS}$
- Update: $Y^{(-1)} := Y - X\hat{\beta}_1$
- $\hat{\beta}_2 := \hat{\beta}_1 + \alpha_2 X^\top Y^{(-1)}$ s.t. $R_2^2 := \|Y - X\hat{\beta}_2\|^2 \to \min_{\alpha_2}$!
  Structure: $X\hat{\beta}_2 = p_2(XX^\top)Y$, $p_2 \in \mathrm{Pol}_{2,0}$ quadratic polynomial with $p_2(0) = 0$ and random coefficients
  $\vdots$
- $\hat{\beta}_k := \hat{\beta}_{k-1} + \alpha_k X^\top Y^{(-(k-1))}$, $R_k^2 := \|Y - X\hat{\beta}_k\|^2 \to \min_{\alpha_k}$!
  Structure: $X\hat{\beta}_k = p_k(XX^\top)Y$, $p_k \in \mathrm{Pol}_{k,0}$

Dimension reduction: PLS

Partial least squares: (Wold 1982, Rosipal, Krämer 2006)

$$\boxed{X^\top Y = (X^\top X)\beta + X^\top \varepsilon}$$

- $\hat\beta_0 := 0$
- $\hat\beta_1 := \alpha_1 X^\top Y$ such that $R_1^2 := \|Y - X\hat\beta_1\|^2 \to \min_{\alpha_1}$!
  Solution: $\alpha_1 = \|X^\top Y\|^2 / \|XX^\top Y\|^2$
  Structure: $X\hat\beta_1 = p_1(XX^\top)Y$, p_1 linear with Y-dependent coefficient
  Note: X orthogonal $\Rightarrow \hat\beta_1 = X^\top Y = \beta + X^\top\varepsilon = \hat\beta^{LS}$
- Update: $Y^{(-1)} := Y - X\hat\beta_1$
- $\hat\beta_2 := \hat\beta_1 + \alpha_2 X^\top Y^{(-1)}$ s.t. $R_2^2 := \|Y - X\hat\beta_2\|^2 \to \min_{\alpha_2}$!
  Structure: $X\hat\beta_2 = p_2(XX^\top)Y$, $p_2 \in \mathrm{Pol}_{2,0}$ quadratic polynomial with $p_2(0) = 0$ and random coefficients

  ⋮

- $\hat\beta_k := \hat\beta_{k-1} + \alpha_k X^\top Y^{(-(k-1))}$, $R_k^2 := \|Y - X\hat\beta_k\|^2 \to \min_{\alpha_k}$!
  Structure: $X\hat\beta_k = p_k(XX^\top)Y$, $p_k \in \mathrm{Pol}_{k,0}$

Dimension reduction: PLS

Partial least squares: (Wold 1982, Rosipal, Krämer 2006)

$$\boxed{X^\top Y = (X^\top X)\beta + X^\top \varepsilon}$$

- $\hat\beta_0 := 0$
- $\hat\beta_1 := \alpha_1 X^\top Y$ such that $R_1^2 := \|Y - X\hat\beta_1\|^2 \to \min_{\alpha_1}$!
  Solution: $\alpha_1 = \|X^\top Y\|^2 / \|XX^\top Y\|^2$
  Structure: $X\hat\beta_1 = p_1(XX^\top)Y$, p_1 linear with Y-dependent coefficient
  Note: X orthogonal $\Rightarrow \hat\beta_1 = X^\top Y = \beta + X^\top\varepsilon = \hat\beta^{LS}$
- Update: $Y^{(-1)} := Y - X\hat\beta_1$
- $\hat\beta_2 := \hat\beta_1 + \alpha_2 X^\top Y^{(-1)}$ s.t. $R_2^2 := \|Y - X\hat\beta_2\|^2 \to \min_{\alpha_2}$!
  Structure: $X\hat\beta_2 = p_2(XX^\top)Y$, $p_2 \in \mathrm{Pol}_{2,0}$ quadratic polynomial with $p_2(0) = 0$ and random coefficients

  $\vdots$
- $\hat\beta_k := \hat\beta_{k-1} + \alpha_k X^\top Y^{(-(k-1))}$, $R_k^2 := \|Y - X\hat\beta_k\|^2 \to \min_{\alpha_k}$!
  Structure: $X\hat\beta_k = p_k(XX^\top)Y$, $p_k \in \mathrm{Pol}_{k,0}$

Dimension reduction: PLS

Partial least squares: (Wold 1982, Rosipal, Krämer 2006)

$$\boxed{X^\top Y = (X^\top X)\beta + X^\top \varepsilon}$$

- $\hat{\beta}_0 := 0$
- $\hat{\beta}_1 := \alpha_1 X^\top Y$ such that $R_1^2 := \| Y - X\hat{\beta}_1 \|^2 \to \min_{\alpha_1}!$
  Solution: $\alpha_1 = \|X^\top Y\|^2 / \|XX^\top Y\|^2$
  Structure: $X\hat{\beta}_1 = p_1(XX^\top)Y$, p_1 linear with Y-dependent coefficient
  Note: X orthogonal $\Rightarrow \hat{\beta}_1 = X^\top Y = \beta + X^\top \varepsilon = \hat{\beta}^{LS}$
- Update: $Y^{(-1)} := Y - X\hat{\beta}_1$
- $\hat{\beta}_2 := \hat{\beta}_1 + \alpha_2 X^\top Y^{(-1)}$ s.t. $R_2^2 := \| Y - X\hat{\beta}_2 \|^2 \to \min_{\alpha_2}!$
  Structure: $X\hat{\beta}_2 = p_2(XX^\top)Y$, $p_2 \in \mathrm{Pol}_{2,0}$ quadratic polynomial with $p_2(0) = 0$ and random coefficients

  $\vdots$
- $\hat{\beta}_k := \hat{\beta}_{k-1} + \alpha_k X^\top Y^{(-(k-1))}$, $R_k^2 := \| Y - X\hat{\beta}_k \|^2 \to \min_{\alpha_k}!$
  Structure: $X\hat{\beta}_k = p_k(XX^\top)Y$, $p_k \in \mathrm{Pol}_{k,0}$

Partial least squares & conjugate gradients

Residual polynomials:

$r_k := 1 - p_k \in \mathrm{Pol}_{k,1}$ where $X\hat{\beta}_k = p_k(XX^\top)Y$. Then

$$\boxed{r_k = \mathrm{argmin}_{r \in \mathrm{Pol}_{k,1}} \|r(XX^\top)Y\|}$$

$\rightsquigarrow$ PLS is *conjugate gradient method* for solving $Y = X\hat{\beta}$.

- Nemirovski (1986)
- Hanke (1995)
- Phatak, de Hoog (2003)
- Blanchard, Mathé (2010)
- Blanchard, Krämer (2016)
- Singer, Krivobokova, de Groot, Munk (2016)

deterministic noise:

rate-optimal when stopped via discrepancy principle

statistical noise:

minimax bounds, not adaptive/sequential, **not easy**

Basic error analysis?

Weak error for PLS/CG

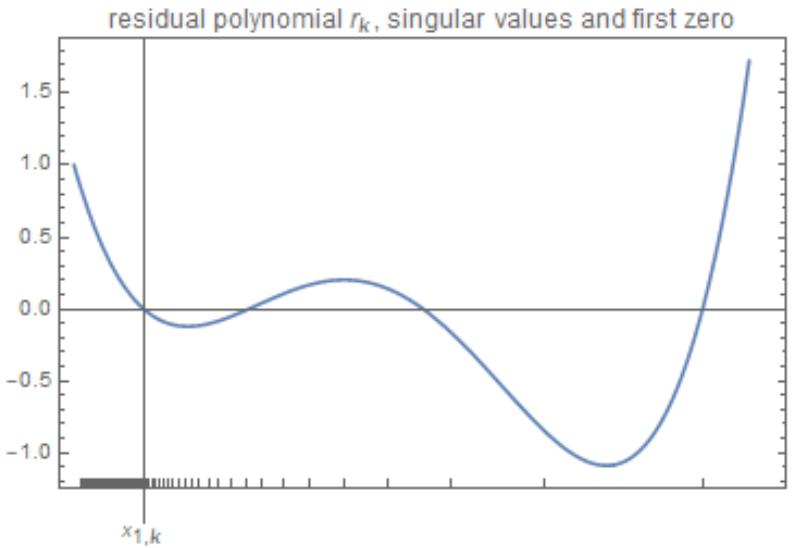
$$rY := r(XX^\top)Y: \quad X\hat{\beta}_k = (1 - r_k)Y, \quad r_k = \operatorname{argmin}_{r \in \mathrm{Pol}_{k,1}} \|rY\|$$

Crucial bound:

$$\boxed{R_k^2 := \|r_k Y\|^2 \leqslant \|r_{k,<}^{1/2} Y\|}$$

$$r_k(x) = \prod_{j=1}^{k}(1 - x/x_{j,k})$$

$$r_{k,<}(x) = r_k(x)\mathbf{1}(x < x_{1,k})$$



Weak norm error:

$$\begin{aligned}
&\|X(\hat{\beta}_k - \beta)\|^2 \\
&= \|(1 - r_k)Y - X\beta\|^2 \\
&= \|r_k Y\|^2 + \|\varepsilon\|^2 - 2\langle \varepsilon, r_k Y\rangle \\
&= \underbrace{\|r_{k,<}^{1/2}(X\beta)\|^2}_{\text{bias control}} + \underbrace{R_k^2 - \|r_{k,<}^{1/2} Y\|^2}_{\leqslant 0} + \underbrace{\|(1 - r_{k,<})^{1/2}\varepsilon\|^2}_{\text{stochastic error}} - \underbrace{2\langle \varepsilon, r_{k,>} Y\rangle}_{\text{cross term}}
\end{aligned}$$

Weak error for PLS/CG

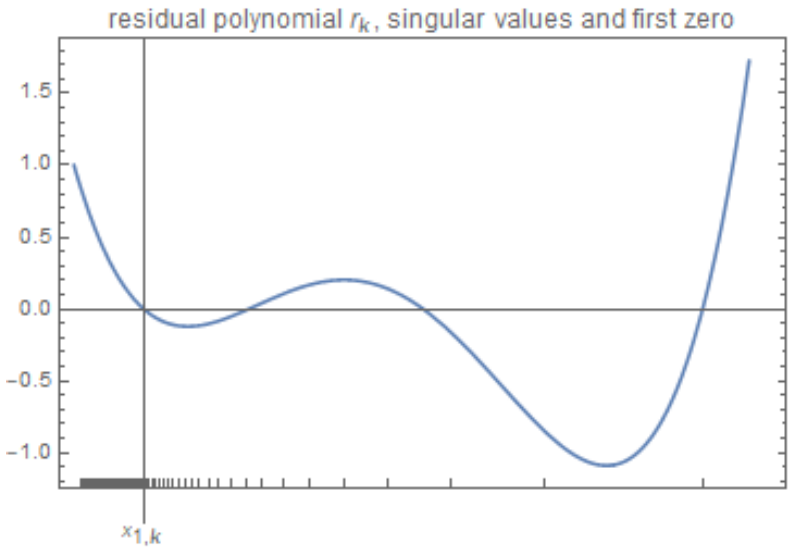
$$rY := r(XX^\top)Y: \quad X\hat{\beta}_k = (1 - r_k)Y, \quad r_k = \operatorname{argmin}_{r \in \mathrm{Pol}_{k,1}} \|rY\|$$

Crucial bound:

$$\boxed{R_k^2 := \|r_k Y\|^2 \leqslant \|r_{k,<}^{1/2} Y\|}$$

$$r_k(x) = \prod_{j=1}^{k}(1 - x/x_{j,k})$$

$$r_{k,<}(x) = r_k(x)\mathbf{1}(x < x_{1,k})$$



Weak norm error:

$$\begin{aligned}
&\|X(\hat{\beta}_k - \beta)\|^2 \\
&= \|(1 - r_k)Y - X\beta\|^2 \\
&= \|r_k Y\|^2 + \|\varepsilon\|^2 - 2\langle \varepsilon, r_k Y\rangle \\
&= \underbrace{\|r_{k,<}^{1/2}(X\beta)\|^2}_{\text{bias control}} + \underbrace{R_k^2 - \|r_{k,<}^{1/2} Y\|^2}_{\leqslant 0} + \underbrace{\|(1 - r_{k,<})^{1/2}\varepsilon\|^2}_{\text{stochastic error}} - \underbrace{2\langle \varepsilon, r_{k,>} Y\rangle}_{\text{cross term}}
\end{aligned}$$

Weak error decomposition

$$S_{k,\lambda} = \|(1 - r_{k,<})^{1/2}\varepsilon\|^2 \qquad \text{(stochastic error)}$$

$$B^2_{k,\lambda} = \|r^{1/2}_{k,<}(X\beta)\|^2 + R^2_k - \|r^{1/2}_{k,<}Y\|^2 \qquad \text{(bias-type error)}$$

Weak norm error:

$$\boxed{\|X(\hat{\beta}_k - \beta)\|^2 = B^2_{k,\lambda} + S_{k,\lambda} - 2\langle\varepsilon, r_{k,>}Y\rangle \leqslant 2(B^2_{k,\lambda} + S_{k,\lambda})}$$

Lemma. The stochastic error term $S_{k,\lambda}$ satisfies:

1. $S_{k,\lambda} = 0$ for $\varepsilon = 0$;
2. $S_{0,\lambda} = 0$ and $S_{D,\lambda} = \|\varepsilon\|^2$;
3. $k \mapsto S_{k,\lambda}$ is increasing.

Lemma. The bias-type error term $B^2_{k,\lambda}$ satisfies:

1. $B^2_{k,\lambda} \leqslant \|r^{1/2}_{k,<}(X\beta)\|^2$ and $B^2_{k,\lambda} \leqslant 0$ if $X\beta = 0$;
2. $B^2_{0,\lambda} = \|X\beta\|^2$ and $B^2_{D,\lambda} = 0$;
3. the upper bound $\|r^{1/2}_{k,<}(X\beta)\|^2$ is decreasing in k.

Weak error decomposition

$$S_{k,\lambda} = \|(1 - r_{k,<})^{1/2}\varepsilon\|^2 \qquad \text{(stochastic error)}$$

$$B^2_{k,\lambda} = \|r^{1/2}_{k,<}(X\beta)\|^2 + R^2_k - \|r^{1/2}_{k,<}Y\|^2 \qquad \text{(bias-type error)}$$

Weak norm error:

$$\boxed{\|X(\hat{\beta}_k - \beta)\|^2 = B^2_{k,\lambda} + S_{k,\lambda} - 2\langle\varepsilon, r_{k,>}Y\rangle \leqslant 2(B^2_{k,\lambda} + S_{k,\lambda})}$$

Lemma. The stochastic error term $S_{k,\lambda}$ satisfies:

1. $S_{k,\lambda} = 0$ for $\varepsilon = 0$;
2. $S_{0,\lambda} = 0$ and $S_{D,\lambda} = \|\varepsilon\|^2$;
3. $k \mapsto S_{k,\lambda}$ is increasing.

Lemma. The bias-type error term $B^2_{k,\lambda}$ satisfies:

1. $B^2_{k,\lambda} \leqslant \|r^{1/2}_{k,<}(X\beta)\|^2$ and $B^2_{k,\lambda} \leqslant 0$ if $X\beta = 0$;
2. $B^2_{0,\lambda} = \|X\beta\|^2$ and $B^2_{D,\lambda} = 0$;
3. the upper bound $\|r^{1/2}_{k,<}(X\beta)\|^2$ is decreasing in k.

Upper bounds

Interpolation: $t = k + \alpha$, $\alpha \in [0, 1)$: $\boxed{\hat{\beta}_t = (1-\alpha)\hat{\beta}_k + \alpha\hat{\beta}_{k+1}}$

Source condition: $\|(XX^\top)^{s/2}(X\beta)\| \leqslant R$

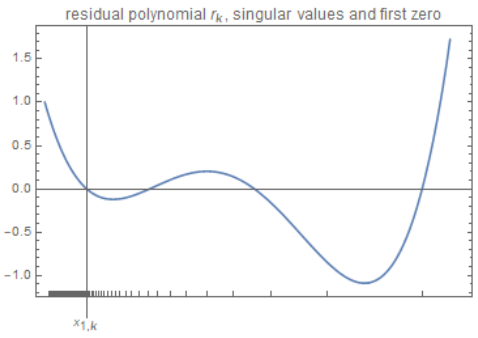
Main weak bounds:

$$B_{t,\lambda}^2 \leqslant \|r_{t,<}^{1/2}(X\beta)\|^2 \leqslant R^2 s^s |r_t'(0)|^{-s} \quad \text{(bias bound)}$$

$$S_{t,\lambda} \leqslant \|((|r_t'(0)|x) \wedge 1)^{1/2}\varepsilon\|^2 \quad \text{(stochastic error bound)}$$

Argument:
r_t on $[0, x_{1,t}]$ is convex and log-concave such that

$$(1 - |r_t'(0)|x)_+ \leqslant r_{t,<}(x) \leqslant \exp(-|r_t'(0)|x)$$



Upper bounds

Interpolation: $t = k + \alpha$, $\alpha \in [0, 1)$: $\boxed{\hat{\beta}_t = (1-\alpha)\hat{\beta}_k + \alpha\hat{\beta}_{k+1}}$

Source condition: $\|(XX^\top)^{s/2}(X\beta)\| \leqslant R$

Main weak bounds:

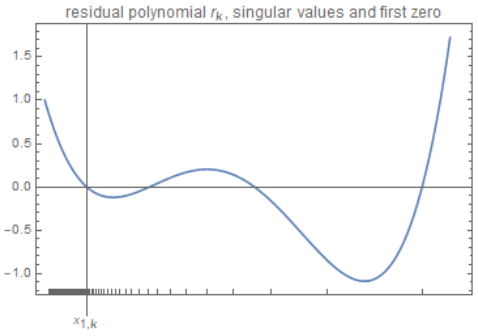
$$B_{t,\lambda}^2 \leqslant \|r_{t,<}^{1/2}(X\beta)\|^2 \leqslant R^2 s^s |r_t'(0)|^{-s} \quad \text{(bias bound)}$$

$$S_{t,\lambda} \leqslant \|((|r_t'(0)|x) \wedge 1)^{1/2}\varepsilon\|^2 \quad \text{(stochastic error bound)}$$

Argument:

r_t on $[0, x_{1,t}]$ is convex and log-concave such that

$$(1 - |r_t'(0)|x)_+ \leqslant r_{t,<}(x) \leqslant \exp(-|r_t'(0)|x)$$



Upper bounds

Interpolation: $t = k + \alpha$, $\alpha \in [0, 1)$: $\boxed{\hat{\beta}_t = (1-\alpha)\hat{\beta}_k + \alpha\hat{\beta}_{k+1}}$

Source condition: $\|(XX^\top)^{s/2}(X\beta)\| \leqslant R$

Main weak bounds:

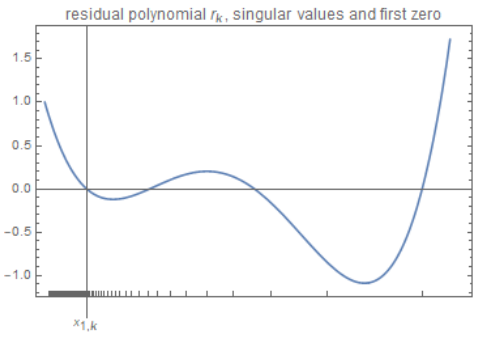
$$B_{t,\lambda}^2 \leqslant \|r_{t,<}^{1/2}(X\beta)\|^2 \leqslant R^2 s^s |r_t'(0)|^{-s} \quad \text{(bias bound)}$$

$$S_{t,\lambda} \leqslant \|((|r_t'(0)|x) \wedge 1)^{1/2}\varepsilon\|^2 \quad \text{(stochastic error bound)}$$

Argument:

r_t on $[0, x_{1,t}]$ is convex and log-concave such that

$$(1 - |r_t'(0)|x)_+ \leqslant r_{t,<}(x) \leqslant \exp(-|r_t'(0)|x)$$



Oracle and minimax upper bound

Weakly balanced oracle $\tau_{\mathfrak{w}}$:

$$\boxed{\tau_{\mathfrak{w}} = \inf\{t \geqslant 0 \,|\, B^2_{t,\lambda} \leqslant S_{t,\lambda}\}}$$

For X-singular values $\lambda_i \sim i^{-p}$, $p > 0$, use

$$\forall\, \rho \geqslant 0 : \ \mathbb{E}[\|((\rho x) \wedge 1)^{1/2}(XX^\top)\varepsilon\|^2] \leqslant C_p \sigma^2 \rho^{1/(2p)}$$

Theorem (Weak upper bound):
Under (s, R)-source condition and $\lambda_i \sim i^{-p}$ we have

$$\boxed{\mathbb{E}\left[\|X(\hat{\beta}_{\tau_{\mathfrak{w}}} - \beta)\|^2\right] \leqslant C_{p,s} R^{2/(2ps+1)} \sigma^{4s/(2s+1/p)}}$$

This rate is minimax optimal.
$p = 1/d$ gives standard Sobolev ellipsoids in d-dimensional domains.

Oracle and minimax upper bound

Weakly balanced oracle $\tau_{\mathfrak{w}}$:

$$\boxed{\tau_{\mathfrak{w}} = \inf\{t \geqslant 0 \mid B_{t,\lambda}^2 \leqslant S_{t,\lambda}\}}$$

For X-singular values $\lambda_i \sim i^{-p}$, $p > 0$, use

$$\forall\, \rho \geqslant 0 : \ \mathbb{E}[\|((\rho x) \wedge 1)^{1/2}(XX^\top)\varepsilon\|^2] \leqslant C_p \sigma^2 \rho^{1/(2p)}$$

Theorem (Weak upper bound):
Under (s, R)-source condition and $\lambda_i \sim i^{-p}$ we have

$$\boxed{\mathbb{E}\left[\|X(\hat{\beta}_{\tau_{\mathfrak{w}}} - \beta)\|^2\right] \leqslant C_{p,s} R^{2/(2ps+1)} \sigma^{4s/(2s+1/p)}}$$

This rate is minimax optimal.
$p = 1/d$ gives standard Sobolev ellipsoids in d-dimensional domains.

Adaptive choice of iteration number?

Early stopping via residuals

Residual norm / weak empirical risk / contrast:

$$R_t^2 := \|Y - X\hat{\beta}_t\|^2 = \|\varepsilon\|^2 + B_{t,\lambda}^2 - S_{t,\lambda} + 2\langle \varepsilon, r_{t,<}(X\beta)\rangle$$

Empirical risk minimisation: open; heavy for D large

Weakly balanced oracle $\tau_{\mathfrak{w}}$:

$$\tau_{\mathfrak{w}} = \inf\{t \geqslant 0 \mid B_{t,\lambda}^2 \leqslant S_{t,\lambda}\} = \inf\{t \geqslant 0 \mid R_t^2 \leqslant \|\varepsilon\|^2 + 2\langle \varepsilon, r_{t,<}X\beta\rangle\}$$

Early stopping iteration τ: $\varepsilon \sim N(0, \sigma^2 E_D) \Rightarrow \mathbb{E}[\|\varepsilon\|^2] = D\sigma^2$

$$\tau = \inf\{t \geqslant 0 \mid R_t^2 \leqslant D\sigma^2\}$$

Early stopping via residuals

Residual norm / weak empirical risk / contrast:

$$\boxed{R_t^2 := \|Y - X\hat{\beta}_t\|^2 = \|\varepsilon\|^2 + B_{t,\lambda}^2 - S_{t,\lambda} + 2\langle \varepsilon, r_{t,<}(X\beta)\rangle}$$

Empirical risk minimisation: open; heavy for D large

Weakly balanced oracle $\tau_{\mathfrak{w}}$:

$$\boxed{\tau_{\mathfrak{w}} = \inf\{t \geqslant 0 \mid B_{t,\lambda}^2 \leqslant S_{t,\lambda}\} = \inf\{t \geqslant 0 \mid R_t^2 \leqslant \|\varepsilon\|^2 + 2\langle \varepsilon, r_{t,<}X\beta\rangle\}}$$

Early stopping iteration τ: $\varepsilon \sim N(0, \sigma^2 E_D) \Rightarrow \mathbb{E}[\|\varepsilon\|^2] = D\sigma^2$

$$\boxed{\tau = \inf\{t \geqslant 0 \mid R_t^2 \leqslant D\sigma^2\}}$$

Early stopping via residuals

Residual norm / weak empirical risk / contrast:

$$R_t^2 := \|Y - X\hat{\beta}_t\|^2 = \|\varepsilon\|^2 + B_{t,\lambda}^2 - S_{t,\lambda} + 2\langle \varepsilon, r_{t,<}(X\beta)\rangle$$

Empirical risk minimisation: open; heavy for D large

Weakly balanced oracle $\tau_{\mathfrak{w}}$:

$$\tau_{\mathfrak{w}} = \inf\{t \geqslant 0 \,|\, B_{t,\lambda}^2 \leqslant S_{t,\lambda}\} = \inf\{t \geqslant 0 \,|\, R_t^2 \leqslant \|\varepsilon\|^2 + 2\langle \varepsilon, r_{t,<}X\beta\rangle\}$$

Early stopping iteration τ: $\varepsilon \sim N(0, \sigma^2 E_D) \Rightarrow \mathbb{E}[\|\varepsilon\|^2] = D\sigma^2$

$$\tau = \inf\{t \geqslant 0 \,|\, R_t^2 \leqslant D\sigma^2\}$$

Weak norm oracle inequality

Lemma. $\|X(\hat\beta_t - \hat\beta_s)\|^2 \leqslant |R_t^2 - R_s^2|$ with equality for integer t, s.

Proposition (Distance to balanced oracle estimator).

$$\mathbb{E}[\|X(\hat\beta_\tau - \hat\beta_{\tau_{\mathrm{w}}})\|^2] \leqslant \sigma^2\sqrt{2D} + 2\,\mathbb{E}[|\langle \varepsilon, r_{\tau_w,<}(X\beta)\rangle|].$$

For $\lambda_i \sim i^{-p}$ we have

$$\mathbb{E}[\|X(\hat\beta_\tau - \hat\beta_{\tau_{\mathrm{w}}})\|^2] \leqslant \sigma^2\sqrt{2D} + C_{p,\lambda}\left(\mathbb{E}[S_{\tau_{\mathrm{w}},\lambda}]^{\frac{4p+1}{4p+2}}\sigma^{\frac{2}{4p+2}} + \mathbb{E}[S_{\tau_{\mathrm{w}},\lambda}]^{1/2}\sigma\sqrt{\log D}\right).$$

Theorem (weak norm oracle-type inequality).
For any (λ_i) we have with a numerical constant $C > 0$

$$\boxed{\mathbb{E}\left[\|X(\hat\beta_\tau - \hat\beta_{\tau_{\mathrm{w}}})\|^2\right] \leqslant C\left(\mathbb{E}\left[\inf_{t\geqslant 0}(A_{t,\lambda} + S_{t,\lambda})\right] + \sigma^2\sqrt{D}\right)}$$

Remark. $\hat\beta_\tau$ is weak minimax adaptive for smoothness $s \in [0, \bar s]$.
Remainder $\sigma^2\sqrt{D}$ unavoidable for PCR.(Blanchard, Hoffmann, MR 2016)

Weak norm oracle inequality

Lemma. $\|X(\hat\beta_t - \hat\beta_s)\|^2 \leqslant |R_t^2 - R_s^2|$ with equality for integer t, s.

Proposition (Distance to balanced oracle estimator).

$$\mathbb{E}[\|X(\hat\beta_\tau - \hat\beta_{\tau_{\mathfrak{w}}})\|^2] \leqslant \sigma^2\sqrt{2D} + 2\,\mathbb{E}[|\langle \varepsilon, r_{\tau_w,<}(X\beta)\rangle|].$$

For $\lambda_i \sim i^{-p}$ we have

$$\mathbb{E}[\|X(\hat\beta_\tau - \hat\beta_{\tau_{\mathfrak{w}}})\|^2] \leqslant \sigma^2\sqrt{2D} + C_{p,\lambda}\left(\mathbb{E}[S_{\tau_{\mathfrak{w}},\lambda}]^{\frac{4p+1}{4p+2}}\sigma^{\frac{2}{4p+2}} + \mathbb{E}[S_{\tau_{\mathfrak{w}},\lambda}]^{1/2}\sigma\sqrt{\log D}\right).$$

Theorem (weak norm oracle-type inequality).
For any (λ_i) we have with a numerical constant $C > 0$

$$\mathbb{E}\left[\|X(\hat\beta_\tau - \hat\beta_{\tau_{\mathfrak{w}}})\|^2\right] \leqslant C\left(\mathbb{E}\left[\inf_{t\geqslant 0}(A_{t,\lambda} + S_{t,\lambda})\right] + \sigma^2\sqrt{D}\right)$$

Remark. $\hat\beta_\tau$ is weak minimax adaptive for smoothness $s \in [0, \bar s]$.
Remainder $\sigma^2\sqrt{D}$ unavoidable for PCR.(Blanchard, Hoffmann, MR 2016)

Weak norm oracle inequality

Lemma. $\|X(\hat\beta_t - \hat\beta_s)\|^2 \leqslant |R_t^2 - R_s^2|$ with equality for integer t, s.

Proposition (Distance to balanced oracle estimator).

$$\mathbb{E}[\|X(\hat\beta_\tau - \hat\beta_{\tau_{\mathfrak{w}}})\|^2] \leqslant \sigma^2\sqrt{2D} + 2\,\mathbb{E}[|\langle \varepsilon, r_{\tau_w,<}(X\beta)\rangle|].$$

For $\lambda_i \sim i^{-p}$ we have

$$\mathbb{E}[\|X(\hat\beta_\tau - \hat\beta_{\tau_{\mathfrak{w}}})\|^2] \leqslant \sigma^2\sqrt{2D} + C_{p,\lambda}\left(\mathbb{E}[S_{\tau_{\mathfrak{w}},\lambda}]^{\frac{4p+1}{4p+2}}\sigma^{\frac{2}{4p+2}} + \mathbb{E}[S_{\tau_{\mathfrak{w}},\lambda}]^{1/2}\sigma\sqrt{\log D}\right).$$

Theorem (weak norm oracle-type inequality).
For any (λ_i) we have with a numerical constant $C > 0$

$$\boxed{\mathbb{E}\left[\|X(\hat\beta_\tau - \hat\beta_{\tau_{\mathfrak{w}}})\|^2\right] \leqslant C\left(\mathbb{E}\left[\inf_{t\geqslant 0}(A_{t,\lambda} + S_{t,\lambda})\right] + \sigma^2\sqrt{D}\right)}$$

Remark. $\hat\beta_\tau$ is weak minimax adaptive for smoothness $s \in [0, \bar s]$.
Remainder $\sigma^2\sqrt{D}$ unavoidable for PCR.(Blanchard, Hoffmann, MR 2016)

Strong norm oracle inequality

Lemma. $\|\hat{\beta}_{k+1} - \hat{\beta}_k\|^2 = (|r'_{k+1}(0)| - |r'_k(0)|)\|X(\hat{\beta}_{k+1} - \hat{\beta}_k)\|^2$

Theorem (minimax adaptive).
$\hat{\beta}_\tau$ is rate-optimal in strong norm over the same source conditions $s \in [0, \bar{s}]$. (to be polished)

Discussion:

- Only one-sided oracle inequality (counterexample!).
- Two-stage procedure:
  1. Compute iterates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_T$ with $T \geqslant \tau$, $S_{T,\lambda} \geqslant \sigma^2\sqrt{D}$.
  2. "Standard" model selection on $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_T$.

  Theory: strongly balanced oracle $\tau_{\mathfrak{s}}$ in $[0, T]$ w.h.p., *but* nonlinear model selection???

Strong norm oracle inequality

Lemma. $\|\hat{\beta}_{k+1} - \hat{\beta}_k\|^2 = (|r'_{k+1}(0)| - |r'_k(0)|)\|X(\hat{\beta}_{k+1} - \hat{\beta}_k)\|^2$

Theorem (minimax adaptive).
$\hat{\beta}_\tau$ is rate-optimal in strong norm over the same source conditions $s \in [0, \bar{s}]$. (to be polished)

Discussion:

- Only one-sided oracle inequality (counterexample!).
- Two-stage procedure:
  1. Compute iterates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_T$ with $T \geqslant \tau$, $S_{T,\lambda} \geqslant \sigma^2\sqrt{D}$.
  2. "Standard" model selection on $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_T$.

  Theory: strongly balanced oracle τ_s in $[0, T]$ w.h.p.,
  but nonlinear model selection???

Strong norm oracle inequality

Lemma. $\|\hat{\beta}_{k+1} - \hat{\beta}_k\|^2 = (|r'_{k+1}(0)| - |r'_k(0)|)\|X(\hat{\beta}_{k+1} - \hat{\beta}_k)\|^2$

Theorem (minimax adaptive).
$\hat{\beta}_\tau$ is rate-optimal in strong norm over the same source conditions $s \in [0, \bar{s}]$. (to be polished)

Discussion:

- Only one-sided oracle inequality (counterexample!).
- Two-stage procedure:
  1. Compute iterates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_T$ with $T \geqslant \tau$, $S_{T,\lambda} \geqslant \sigma^2\sqrt{D}$.
  2. "Standard" model selection on $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_T$.

  Theory: strongly balanced oracle $\tau_{\mathfrak{s}}$ in $[0, T]$ w.h.p.,
  but nonlinear model selection???

Summary

- PLS/CG (often $=$) popular nonlinear methods.
- Early stopping is regularisation.
- Nonlinear error decomposition, optimal rates.
- Stopping rule τ mimics weakly balanced oracle $\tau_{\mathfrak{w}}$.
- Weak norm oracle-type inequality, remainder $\sigma^2\sqrt{D}$.
- Adaptive minimax over range of smoothness classes.
- Challenges (contributions welcome!):
  - "Standard" model selection
  - Analysis of two-stage procedure
  - Oracle inequality for true risk, not for $2(B_{t,\lambda}^2 + S_{t,\lambda})$
  - Number of iterations (on average)
  - Deterministic vs. statistical noise

Summary

- PLS/CG (often $=$) popular nonlinear methods.
- Early stopping is regularisation.
- Nonlinear error decomposition, optimal rates.
- Stopping rule τ mimics weakly balanced oracle $\tau_{\mathfrak{w}}$.
- Weak norm oracle-type inequality, remainder $\sigma^2\sqrt{D}$.
- Adaptive minimax over range of smoothness classes.
- Challenges (contributions welcome!):
  - "Standard" model selection
  - Analysis of two-stage procedure
  - Oracle inequality for true risk, not for $2(B_{t,\lambda}^2 + S_{t,\lambda})$
  - Number of iterations (on average)
  - Deterministic vs. statistical noise

Happy birthday and a long life, Oleg & Sasha!

Summary

- PLS/CG (often $=$) popular nonlinear methods.
- Early stopping is regularisation.
- Nonlinear error decomposition, optimal rates.
- Stopping rule τ mimics weakly balanced oracle $\tau_{\mathfrak{w}}$.
- Weak norm oracle-type inequality, remainder $\sigma^2\sqrt{D}$.
- Adaptive minimax over range of smoothness classes.
- Yet another challenge

Thanks to all for listening.