Early stopping for L^2-boosting in high-dimensional linear models

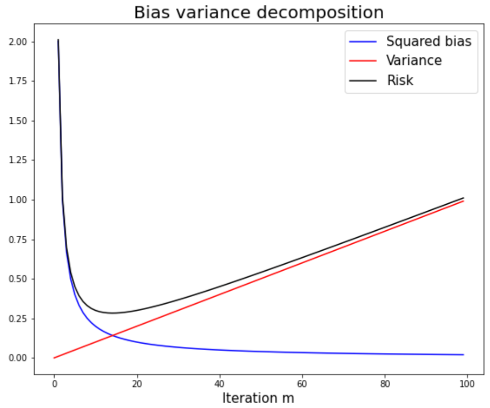
Bernhard Stankewitz

AIP 2023, Göttingen

Department of Decision Sciences
Bocconi University

Early stopping vs. model selection

For a given iterative estimation procedure $(\widehat{F}^{(m)})_{m\geq 0}$, choose a data driven iteration $\widehat{m}$ that neither over- nor underfits the data.



Model selection

for all $(m) \leq m_{\max}$ **do**

  compute $\widehat{F}^{(m)}$ and criterion(m)

end for

$\widehat{m} \leftarrow \text{argmin}_{m \leq m_{\max}}$ criterion(m)

Early stopping

while condition(m) is false **do**

  compute $\widehat{F}^{(m)}$ and condition(m)

  $m \leftarrow m + 1$

end while

$\widehat{m} \leftarrow m$

Can computational and statistical complexity be treated at the same time?

- Used everywhere in machine learning. Limited theoretical understanding.
- Positive results as in Blanchard and Mathé [BM12], Blanchard, Hoffmann, and Reiß [BHR18a; BHR18b], Celisse and Wahl [CW21] yield substantial computational gains.
- Negative results lead to important questions about statistical optimality under information/computational constraints, see Blanchard, Hoffmann, and Reiß [BHR18a].[1]
- Many possible applications and open questions.

[1] G. Blanchard, M. Hoffmann, and M. Reiß. "Early stopping for statistical inverse problems via truncated SVD estimation". In: *Electronic Journal of Statistics* 12.2 (2018), pp. 3204-3231.

L^2-boosting

Consider i.i.d. observations from a high dimensional linear model

$$Y_i = f^*(X_i) + \varepsilon_i = \sum_{j=1}^{p} \beta_j^* X_i^{(j)} + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1}$$

where $p \gg n$ with $\log(p)/n \to 0$ for $n \to \infty$ and we assume:

(A1) **(SubGE)**: Conditional on the design, the noise terms are centered subgaussians with a joint parameter $\overline{\sigma}^2 > 0$.

Examples

(a) **(Gaussian Regression):** For $\varepsilon_1, \dots \varepsilon_n \sim N(0, \sigma^2)$ i.i.d., we have $\overline{\sigma}^2 = \sigma^2$.

(b) **(Classification):** For classification, we consider i.i.d. observations

$$Y_i \sim \mathrm{Ber}(f^*(X_i)), \qquad i = 1, \dots, n. \tag{2}$$

Then, the noise terms are given by $\varepsilon_i = Y_i - f^*(X_i)$.

L^2-boosting

Consider i.i.d. observations from a high dimensional linear model

$$Y_i = f^*(X_i) + \varepsilon_i = \sum_{j=1}^{p} \beta_j^* X_i^{(j)} + \varepsilon_i, \qquad i = 1, \dots, n. \tag{3}$$

Algorithm 1.1 (Orthogonal matching pursuit (OMP))

1: $\widehat{F}^{(0)} \leftarrow 0, \widehat{J}_0 \leftarrow \emptyset$

2: **for** $m = 0, 1, 2, \dots$ **do**

3: $\quad \widehat{j}_{m+1} \leftarrow \operatorname{argmax}_{j \leq p} \left| \left\langle Y - \widehat{F}^{(m)}, \frac{X^{(j)}}{\|X^{(j)}\|_n} \right\rangle_n \right|$

4: $\quad \widehat{J}_{m+1} \leftarrow \widehat{J}_m \cup \{\widehat{j}_{m+1}\}$

5: $\quad \widehat{F}^{(m+1)} \leftarrow \widehat{\Pi}_{\widehat{J}_{m+1}} Y$

6: **end for**

- Inner product $\langle a, b \rangle_n := n^{-1} \sum_{i=1}^n a_i b_i$ with norm $\|a\|_n := \langle a, a \rangle_n^{1/2}$ for $a, b \in \mathbb{R}^n$.
- $\widehat{\Pi}_J : \mathbb{R}^n \to \mathbb{R}^n$ orthogonal projection onto $\operatorname{span}(X^{(j)}, j \in J)$.
- Analysis of greedy algorithms Temlyakov [Tem00]. In a statistical setting Bühlmann [Bü06].

Early stopping

Early stopping according to the discrepancy principle

$$\tau := \inf\{m \geq 0 : \|Y - \widehat{F}^{(m)}\|_n^2 \leq \kappa\} \quad \text{for some critical value} \quad \kappa \approx \|\varepsilon\|_n^2. \tag{4}$$

The empirical risk has the decomposition

$$\|\widehat{F}^{(m)} - f^*\|_n^2 = \|(I - \widehat{\Pi}_m)f^*\|_n^2 + \|\widehat{\Pi}_m \varepsilon\|_n^2 =: b_m^2 + s_m. \tag{5}$$

Early stopping

Early stopping according to the discrepancy principle

$$\tau := \inf\{m \geq 0 : \|Y - \widehat{F}^{(m)}\|_n^2 \leq \kappa\} \quad \text{for some critical value} \quad \kappa \approx \|\varepsilon\|_n^2. \tag{4}$$

The empirical risk has the decomposition

$$\|\widehat{F}^{(m)} - f^*\|_n^2 = \|(I - \widehat{\Pi}_m)f^*\|_n^2 + \|\widehat{\Pi}_m\varepsilon\|_n^2 =: b_m^2 + s_m. \tag{5}$$

The residuals can be written as

$$\begin{aligned} \|Y - \widehat{F}^{(m)}\|_n^2 &= \|(I - \widehat{\Pi}_m)f^*\|_n^2 + 2\langle (I - \widehat{\Pi}_m)f, \varepsilon\rangle_n + \|\varepsilon\|_n^2 - \|\widehat{\Pi}_m\varepsilon\|_n^2 \\ &=: b_m^2 + 2c_m + \|\varepsilon\|_n^2 - s_m. \end{aligned} \tag{6}$$

Early stopping

Early stopping according to the discrepancy principle

$$\tau := \inf\{m \geq 0 : \|Y - \widehat{F}^{(m)}\|_n^2 \leq \kappa\} \quad \text{for some critical value} \quad \kappa \approx \|\varepsilon\|_n^2. \tag{4}$$

The empirical risk has the decomposition

$$\|\widehat{F}^{(m)} - f^*\|_n^2 = \|(I - \widehat{\Pi}_m)f^*\|_n^2 + \|\widehat{\Pi}_m \varepsilon\|_n^2 =: b_m^2 + s_m. \tag{5}$$

The residuals can be written as

$$\begin{aligned} \|Y - \widehat{F}^{(m)}\|_n^2 &= \|(I - \widehat{\Pi}_m)f^*\|_n^2 + 2\langle (I - \widehat{\Pi}_m)f, \varepsilon\rangle_n + \|\varepsilon\|_n^2 - \|\widehat{\Pi}_m \varepsilon\|_n^2 \\ &=: b_m^2 + 2c_m + \|\varepsilon\|_n^2 - s_m. \end{aligned} \tag{6}$$

Intuition for early stopping

Therefore, the stopping condition $\|Y - \widehat{F}^{(m)}\|_n^2 \leq \kappa$ is equivalent to

$$b_m^2 + 2c_m \leq s_m + \kappa - \|\varepsilon\|_n^2. \tag{7}$$

For $\kappa \approx \|\varepsilon\|_n^2$, τ mimics the *balanced oracle* $m^{\mathfrak{b}} := \inf\{m \geq 0 : b_m^2 \leq s_m\}$.

A general oracle inequality for the empirical risk

Discrepancy principle with noise estimation

$$\tau := \inf\{m \geq 0 : \|Y - \widehat{F}^{(m)}\|_n^2 \leq \kappa_m\} \quad \text{with} \quad \kappa_m := \widehat{\sigma}^2 + \frac{C_\tau m \log p}{n}, \qquad m \geq 0. \tag{8}$$

A general oracle inequality for the empirical risk

Discrepancy principle with noise estimation

$$\tau := \inf\{m \geq 0 : \|Y - \widehat{F}^{(m)}\|_n^2 \leq \kappa_m\} \quad \text{with} \quad \kappa_m := \widehat{\sigma}^2 + \frac{C_\tau m \log p}{n}, \qquad m \geq 0. \tag{8}$$

Theorem (Oracle inequality for the empirical risk)

Under Assumption **(SubGE)**, *the empirical risk at the stopping time τ in Equation (8) with $C_\tau \geq 8\overline{\sigma}^2$ satisfies*

$$\begin{aligned}
\|\widehat{F}^{(\tau)} - f^*\|_n^2 &\leq \min_{m\geq 0}\left(7\|\widehat{F}^{(m)} - f^*\|_n^2 + \frac{(8\overline{\sigma}^2 + C_\tau)m \log p}{n}\right) + |\widehat{\sigma}^2 - \|\varepsilon\|_n^2| \\
&\leq 7\|\widehat{F}^{(m^\mathfrak{b})} - f^*\|_n^2 + \frac{(8\overline{\sigma}^2 + C_\tau)m^\mathfrak{b} \log p}{n} + |\widehat{\sigma}^2 - \|\varepsilon\|_n^2|
\end{aligned}$$

with probability converging to one.

Adaption for the population risk

Analogous to the empirical quantities:

- $\langle f, g\rangle_{L^2} := \mathbb{E}(f(X_1)g(X_1))$ with norm $\|f\|_{L^2} := \langle f, f\rangle_{L^2}^{1/2}$ for functions $f, g \in L^2(\mathbb{P}^{X_1})$, where $\mathbb{P}^{X_1}$ denotes the distribution of one observation of the covariates.
- $\Pi_J : L^2(\mathbb{P}^{X_1}) \to L^2(\mathbb{P}^{X_1})$ denote the orthogonal projection with respect to $\langle \cdot, \cdot\rangle_{L^2}$ onto the span of the covariates $\{X_1^{(j)} : j \in J\}$.

Setting $\Pi_m := \Pi_{\widehat{J}_m}$, the population risk decomposes into

$$\|\widehat{F}^{(m)} - f^*\|_{L^2}^2 = \|(I - \Pi_m)f^*\|_{L^2}^2 + \|\widehat{F}^{(m)} - \Pi_m f^*\|_{L^2}^2 = B_m^2 + S_m, \tag{9}$$

with $B_m^2 := \|(I - \Pi_m)f^*\|_{L^2}^2$ and $S_m := \|\widehat{F}^{(m)} - \Pi_m f^*\|_{L^2}^2$.

(A2) **(Sparse):** We assume one of the two following assumptions holds:

(i) β^* is s-sparse for some $s \in \mathbb{N}_0$, i.e. $|\{j \le p : |\beta_j^*| \neq 0\}| \le s$. Additionally,

$$s\|\beta^*\|_1^2 = s\Big(\sum_{j=1}^{p} |\beta_j^*|\Big)^2 = o\Big(\frac{n}{\log p}\Big), \quad \|f^*\|_{L^2}^2 \le C_{f^*} \quad \text{and} \quad \min_{j \in S} |\beta_j^*| \ge \underline{\beta}.$$

(ii) β^* is γ-sparse for some $\gamma \in [1, \infty)$, i.e., $\|\beta^*\|_2 \le C_{\ell 2}$ and

$$\sum_{j \in J} |\beta_j^*| \le C_\gamma \Big(\sum_{j \in J} |\beta_j^*|^2\Big)^{\frac{\gamma-1}{2\gamma-1}} \qquad \text{for all } J \subset \{1, \dots, p\},$$

where $C_{\ell 2}, C_\gamma > 0$ are numerical constants.

(A3) **(SubGD):** The design variables are centered subgaussians in $\mathbb{R}^p$ with unit variance, i.e., there exists some $\rho > 0$ such that for all $x \in \mathbb{R}^p$ with $\|x\| = 1$,

$$\mathbb{E}e^{u\langle x, X_1\rangle} \le e^{\frac{u^2\rho^2}{2}}, \quad u \in \mathbb{R} \qquad \text{and} \qquad \mathrm{Var}(X_1^{(j)}) = 1 \quad \text{for all } j \le p.$$

(A4) **(CovB):** The covariance matrix $\Gamma := \mathrm{Cov}(X_1)$ of one design observation satisfies

$$\lambda_{\min}(\Gamma) \ge c_\lambda > 0 \tag{10}$$

and the sum of partial covariance terms are *sufficiently bounded*.

Under Assumptions **(SubGE)**, **(Sparse)**, **(SubGD)** and **(CovB)**,

$$B_m^2 \lesssim \begin{cases} \exp\left(\frac{-c_{\text{Bias}} m}{s}\right) & \beta^* \ s\text{-sparse}, \\ m^{1-2\gamma} & \beta^* \ \gamma\text{-sparse} \end{cases} \quad \text{and} \quad S_m \lesssim \frac{(\overline{\sigma}^2 + \rho^4) m \log p}{n} \tag{11}$$

using theory developed in Ing [Ing20].[2]

The quantities balance at

$$m^*_{s,\gamma} := \begin{cases} C_{\text{supp}} s, & \beta^* \ s\text{-sparse}, \\ \left(\frac{n}{(\overline{\sigma}^2 + \rho^4) \log p}\right)^{\frac{1}{2\gamma}}, & \beta^* \ \gamma\text{-sparse} \end{cases} \tag{12}$$

with

$$\|\widehat{F}^{(m^*_{s,\gamma})} - f^*\|_{L^2}^2 \lesssim \begin{cases} \frac{\overline{\sigma}^2 s \log p}{n}, & \beta^* \ s\text{-sparse}, \\ \left(\frac{(\overline{\sigma}^2 + \rho^4) \log p}{n}\right)^{1 - \frac{1}{2\gamma}}, & \beta^* \ \gamma\text{-sparse} \end{cases} \tag{13}$$
$$=: \mathcal{R}(s, \gamma).$$

[2] C. Ing. "Model selection for high-dimensional linear regression with dependent observations". In: *The Annals of Statistics* 48 (2020), pp.1959-1980.

Theorem (Optimal adaptation for the population risk)

Under Assumptions **(SubGE)**, **(Sparse)**, **(SubGD)** *and* **(CovB)**, *choose* $\widehat{\sigma}^2$ *in Equation* (8) *such that there is a constant* $C_{Noise} > 0$ *for which*

$$|\widehat{\sigma}^2 - \|\varepsilon\|_n^2| \leq C_{Noise}\mathcal{R}(s, \gamma)$$

with probability converging to one. Then, the population risk at the stopping time τ *with* $C_\tau = c(\overline{\sigma}^2 + \rho^4)$ *for any* $c > 0$ *satisfies*

$$\|\widehat{F}^{(\tau)} - f^*\|_{L^2}^2 \leq C_{PopRisk}\mathcal{R}(s, \gamma)$$

with probability converging to one for a constant $C_{PopRisk} > 0$.

Theorem (Optimal adaptation for the population risk)

Under Assumptions **(SubGE)**, **(Sparse)**, **(SubGD)** *and* **(CovB)**, *choose* $\widehat{\sigma}^2$ *in Equation* (8) *such that there is a constant* $C_{Noise} > 0$ *for which*

$$|\widehat{\sigma}^2 - \|\varepsilon\|_n^2| \leq C_{Noise}\mathcal{R}(s, \gamma)$$

with probability converging to one. Then, the population risk at the stopping time τ *with* $C_\tau = c(\overline{\sigma}^2 + \rho^4)$ *for any* $c > 0$ *satisfies*

$$\|\widehat{F}^{(\tau)} - f^*\|_{L^2}^2 \leq C_{PopRisk}\mathcal{R}(s, \gamma)$$

with probability converging to one for a constant $C_{PopRisk} > 0$.

Preliminary result

Sequential adaptation works when $\|\varepsilon\|_n^2$ can be estimated well.

Noise estimation

Proposition (Fast noise estimation)

Under Assumptions **(SubGE)**, **(Sparse)** *and* **(CovB)** *with Gaussian design* $(X_i)_{i \leq n} \sim N(0, \Gamma)$ *i.i.d., set* $\xi > 1$ *and* $\lambda_0 = C_{\lambda_0}(\xi + 1)/(\xi - 1)\sqrt{\log(p)/n}$ *with* $C_{\lambda_0} \geq 2C_\varepsilon \overline{\sigma}/\underline{\sigma}$. *Then, the Scaled Lasso noise estimator* $\widehat{\sigma}^2$ *from Sun and Zhang [SZ12][3] satisfies*

$$|\widehat{\sigma}^2 - \|\varepsilon\|_n^2| \leq C \begin{cases} \dfrac{\overline{\sigma}^2 s \log p}{n}, & \beta^* \text{ s-sparse}, \\ \left(\dfrac{\overline{\sigma}^2 \log p}{n}\right)^{1-1/(2\gamma)}, & \beta^* \ \gamma\text{-sparse} \end{cases}$$

with probability converging to one.

- $\|\varepsilon\|_n^2$ is easier to estimate than $\mathrm{Var}(\varepsilon_1)$.
- Only need to solve one convex optimization problem.
- Together with the preliminary results, we obtain full sequential adaptation.

[3] T. Sun and C. H. Zhang "Scaled sparse linear regression". In: *Biometrika* 99.4 (2012), pp. 879-898.

An improved two-step procedure

Perform second step based on a high-dimensional Akaike-information criterion

$$\tau_{\text{two-step}} := \operatorname*{argmin}_{m \leq \tau} \mathrm{AIC}(m) \quad \text{with} \quad \mathrm{AIC}(m) := \|Y - \widehat{F}^{(m)}\|_n^2 + \frac{C_{\mathrm{AIC}}\, m \log p}{n}, \quad m \geq 0. \tag{14}$$

An improved two-step procedure

Perform second step based on a high-dimensional Akaike-information criterion

$$\tau_{\text{two-step}} := \operatorname*{argmin}_{m \le \tau} \mathrm{AIC}(m) \quad \text{with} \quad \mathrm{AIC}(m) := \|Y - \widehat{F}^{(m)}\|_n^2 + \frac{C_{\mathrm{AIC}}\, m \log p}{n}, \quad m \ge 0. \tag{14}$$

Theorem (Two-step procedure)

Under Assumptions **(SubGE)**, **(Sparse)**, **(SubGD)** *and* **(CovB)**, *choose* $\widehat{\sigma}^2$ *such that*

$$\widehat{\sigma}^2 \le \|\varepsilon\|_n^2 + C\mathcal{R}(s, \gamma)$$

with probability converging to one. Then, for any choice $C_\tau \ge 0$ *in* (8) *with* $c \ge 0$ *and* $C_{AIC} = C(\overline{\sigma}^2 + \rho^4)$ *with* $C > 0$ *large enough, the two-step procedure satisfies that with probability converging to one,* $\tau_{\text{two-step}} \ge \tilde{m}_{s,\gamma,G}$ *from Equation* (**??**) *for some* $G > 0$*. On the corresponding event,*

$$\|\widehat{F}^{(\tau_{\text{two-step}})} - f^*\|_{L^2}^2 \le C_{Risk}\mathcal{R}(s, \gamma)$$

for some constant $C_{Risk} > 0$.

A small simulation example

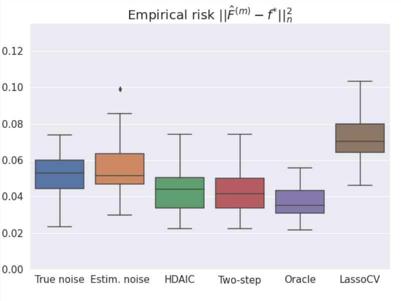


Figure 1: Empirical risk for different methods.

True noise	19.8 sec
Estimated noise	32.0 sec
Two-step	49.6 sec
HDAIC	411.6 sec
Lasso CV	164.3 sec

Table 1: Computation times for different methods.

[Sta22] Early stopping for L^2-boosting in high dimensional linear models. 2022. https://arxiv.org/abs/2210.07850

Thank you!

Early stopping for L^2-boosting in high-dimensional linear models

Bernhard Stankewitz[1],

[1] *Department of Mathematics, Humboldt-University of Berlin, e-mail:* stankebe@math.hu-berlin.de

Abstract: Increasingly high-dimensional data sets require that estimation methods do not only satisfy statistical guarantees but also remain computationally feasible. In this context, we consider L^2-boosting via orthogonal matching pursuit in a high-dimensional linear model and analyze a data-driven early stopping time τ of the algorithm, which is sequential in the sense that its computation is based on the first τ iterations only. This approach is much less costly than established model selection criteria, that require the computation of the full boosting path. We prove that sequential early stopping preserves statistical optimality in this setting in terms of a fully general oracle inequality for the empirical risk and recently established optimal convergence rates for the population risk. Finally, an extensive simulation study shows that at an immensely reduced computational cost, the performance of these type of methods is on par with other state of the art algorithms such as the cross-validated Lasso or model selection via a high dimensional Akaike criterion based on the full boosting path.

MSC2020 subject classifications: Primary 62G05, 62J07; secondary 62F35.
Keywords and phrases: Early stopping, Discrepancy principle, Adaptive estimation, Oracle inequalities, L2-Boosting, Orthogonal matching pursuit.

1. Introduction

Iterative estimation procedures typically have to be combined with a data-driven choice $\widehat{m}$ of the effectively selected iteration in order to avoid under- as well as over-fitting. In the context of increasingly high-dimensional data sets, which require that estimation methods do not only provide statistical guarantees but also ensure computational feasibility, established model selection criteria for $\widehat{m}$ such as *cross-validation*, *unbiased risk estimation*, *Akaike's information criterion* or *Lepski's balancing principle* suffer from a disadvantage: They involve computing the full iteration path up to some large $m_{\max}$, which is computationally costly, even if the final choice $\widehat{m}$ is much smaller than $m_{\max}$. In comparison, *sequential early stopping*, i.e., halting the procedure at an iteration $\widehat{m}$ depending only on the iterates $m \le \widehat{m}$, can substantially reduce computational complexity while maintaining guarantees in terms of adaptivity. For inverse problems, results were established in Blanchard and Mathé [5], Blanchard et al. [3, 4], Stankewitz [20] and Jahn [14]. A Poisson inverse problem was treated in Mika and Szkutnik [16] and general kernel learning in Celisse and Wahl [8].

In this work, we analyze sequential early stopping for an iterative boosting algorithm applied to data $Y = (Y_i)_{i \le n}$ from a high-dimensional linear model

$$Y_i = f^*(X_i) + \varepsilon_i = \sum_{j=1}^{p} \beta_j^* X_i^{(j)} + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1.1}$$

References

[BHR18a] G. Blanchard, M. Hoffmann, and M. Reiß. **"Early stopping for statistical inverse problems via truncated SVD estimation".** In: *Electronic Journal of Statistics* 12.2 (2018), pp. 3204–3231.

[BHR18b] G. Blanchard, M. Hoffmann, and M. Reiß. **"Optimal adaptation for early stopping in statistical inverse problems".** In: *SIAM/ASA Journal of Uncertainty Quantification* 6.3 (2018), pp. 1043–1075.

[BM12] G. Blanchard and P. Mathé. **"Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration".** In: *Inverse Problems* 28.11 (2012), pp. 115011/1–115011/23.

[Bü06] P. Bühlmann. **"Boosting for high-dimensional linear models".** In: *The annals of statistics* 34.2 (2006), pp. 559–583.

[CW21] A. Celisse and M. Wahl. **"Analyzing the Discrepancy Principle for Kernelized Spectral Filter Learning Algorithms".** In: *Journal of Machine Learning Research* 22.76 (2021), pp. 1–59.

References ii

[Ing20] C. Ing. **"Model selection for high-dimensional linear regression with dependent observations".** In: *The Annals of Statistics* 48 (2020), pp. 1959–1980.

[SZ12] T. Sun and C.-H. Zhang. **"Scaled sparse linear regression".** In: *Biometrika* 99.4 (2012), pp. 879–898.

[Sta22] B. Stankewitz. **Early stopping for L2-boosting in high-dimensional linear models.** 2022. URL: https://arxiv.org/abs/2210.07850.

[Tem00] V. N. Temlyakov. **"Weak greedy algorithms".** In: *Advances in Computational Mathematics* 12 (2000), pp. 213–227.