Early stopping for conjugate gradients in statistical inverse problems

Joint work in progress with Markus Reiß

Laura Hucker

StatMathAppli 2023 in Fréjus // September 21, 2023

Institut für Mathematik, Humboldt-Universität zu Berlin

Supported by
DFG Research Unit 5381 – Mathematical Statistics in the Information Age
Project 2: Optimal actions and stopping in sequential learning

Setting

Statistical inverse problem (see e.g. Cavalier 2011)

$$Y = g + \xi = Af + \delta \dot{W}$$

with $A \in \mathcal{L}(H_1, H_2)$, noise level $\delta > 0$ and Gaussian white noise $\dot{W}$

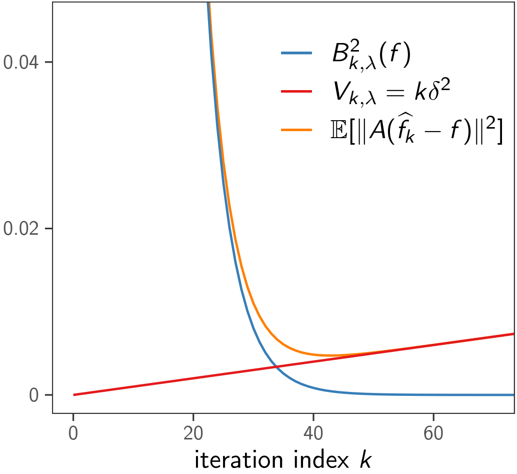
Discretised model

$$Y = g + \xi = Af + \delta Z$$

with

- $A \in \mathbb{R}^{D \times P}$ with rank $D \leqslant P$ and singular values $\lambda_1 > \cdots > \lambda_D > 0$,
- $f \in \mathbb{R}^P$ signal of interest,
- $Z \sim \mathcal{N}(0, \mathrm{I}_D)$, $\delta > 0$

Regularisation of iterative methods by early stopping

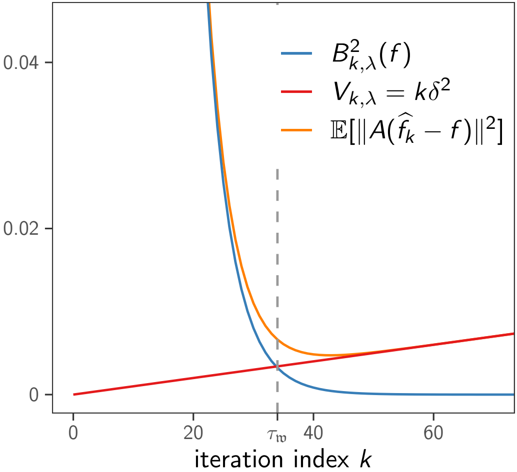


Bias-variance decomposition of the prediction error for $\widehat{f}_k = \sum_{i=1}^{k} \lambda_i^{-1} \langle Y, u_i \rangle v_i$

$B^2_{k,\lambda}(f) = \sum_{i=k+1}^{D} \lambda_i^2 \langle f, v_i \rangle^2$,
$V_{k,\lambda} = k\delta^2$

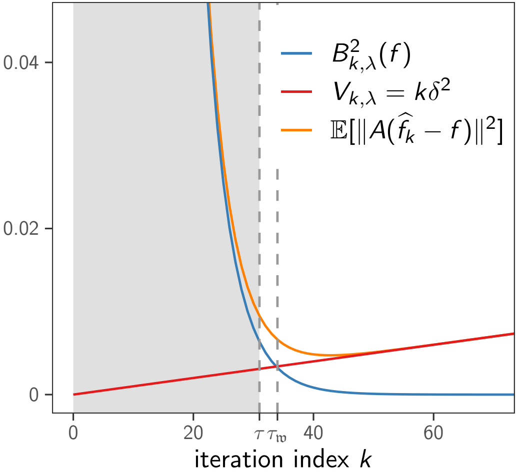
(truncated SVD, see Blanchard et al. 2018a)

Regularisation of iterative methods by early stopping



Bias-variance decomposition
of the prediction error
with **weakly balanced oracle**
$\tau_{\mathfrak{w}} := \inf\{k \mid B^2_{k,\lambda}(f) \leqslant V_{k,\lambda}\}$

Regularisation of iterative methods by early stopping



Bias-variance decomposition of the prediction error with **weakly balanced oracle**

$\tau_{\mathfrak{w}} := \inf\{k \mid B^2_{k,\lambda}(f) \leqslant V_{k,\lambda}\}$

Goal: computational and statistical efficiency by choosing data-driven τ depending on previous iterates only

Conjugate gradients for the normal equation (CGNE)

$A^\top A f = A^\top (Y - \xi)$ (normal equation)

$(1/2)\langle Af, Af\rangle - \langle Af, Y\rangle \to \min_f!$ (minimisation problem)

Algorithm (Hestenes and Stiefel 1952)

$\widehat{f}_0 \leftarrow 0,\ Y^{(-0)} \leftarrow Y,\ p_1 \leftarrow A^\top Y,\ k = 1$

while $A^\top Y^{(-(k-1))} \neq 0$ **do**

  $\widehat{f}_k \leftarrow \widehat{f}_{k-1} + \alpha_k p_k$ s.t. $\|Y - A\widehat{f}_k\|^2 \to \min_{\alpha_k}!$

  $Y^{(-k)} \leftarrow Y - A\widehat{f}_k$

  $\gamma_k \leftarrow \|A^\top Y^{(-k)}\|^2 / \|A^\top Y^{(-(k-1))}\|^2$

  $p_{k+1} \leftarrow A^\top Y^{(-k)} + \gamma_k p_k$

  $k \leftarrow k + 1$

end while

Note: $\widehat{f}_k$ depends **nonlinearly** on Y.

An alternative definition of CGNE

CGNE as Krylov subspace iteration method:

$$\|Y - \widehat{g}_k\| = \|Y - A\widehat{f}_k\| = \min_{\widehat{f} \in \mathcal{K}_k(A^\top Y, A^\top A)} \|Y - A\widehat{f}\| \quad \text{with}$$

$$\mathcal{K}_k(A^\top Y, A^\top A) = \mathrm{span}\{A^\top Y, (A^\top A)A^\top Y, \ldots, (A^\top A)^{k-1}A^\top Y\}$$

An alternative definition of CGNE

CGNE as polynomial based iterative method:

$$\|Y - \widehat{g}_k\| = \|Y - A\widehat{f}_k\| = \min_{\widehat{f} \in \mathcal{K}_k(A^\top Y, A^\top A)} \|Y - A\widehat{f}\| \quad \text{with}$$

$$\mathcal{K}_k(A^\top Y, A^\top A) = \operatorname{span}\{A^\top Y, (A^\top A)A^\top Y, \ldots, (A^\top A)^{k-1}A^\top Y\}$$

Definition (Conjugate gradient iterate at iteration k)

$$\widehat{g}_k := (1 - r_k(AA^\top))Y, \quad \text{where} \quad r_k := \arg\min_{p_k \in \mathrm{Pol}_{k,1}} \|p_k(AA^\top)Y\|^2$$

Key property

$(r_k)_{k \geqslant 0}$ *is orthogonal w.r.t.* $d\widehat{\mu}(\lambda) = \sum_{i=1}^{D} \lambda_i^2 \langle Y, u_i \rangle^2 \delta_{\lambda_i^2}$.

Notation: $hv = h(AA^\top)v$ for functions h and $v \in \mathbb{R}^D$

Decomposition of the prediction error of CGNE

Proposition

We have

$$\|\widehat{g}_t - g\|^2 = A_{t,\lambda} + S_{t,\lambda} - 2\langle \xi, r_{t,>} Y\rangle \leqslant 2(A_{t,\lambda} + S_{t,\lambda})$$

with

$$S_{t,\lambda} := \|(1 - r_{t,<})^{1/2}\xi\|^2, \qquad \text{(weak stochastic error)}$$

$$A_{t,\lambda} := \|r_{t,<}^{1/2} g\|^2 + R_t^2 - \|r_{t,<}^{1/2} Y\|^2, \qquad \text{(weak approximation error)}$$

$$r_t := (1-\alpha) r_k + \alpha r_{k+1}, \qquad \text{(interpol. residual polynomial)}$$

$$r_{t,<}(x) = r_t(x)1(x < x_{1,t}), \quad r_{t,>}(x) = r_t(x)1(x > x_{1,t}),$$

where $t = k + \alpha$, $k = 0, \ldots, D-1$, $\alpha \in (0,1]$ *and* $x_{1,t}$ *is the smallest zero of* r_t.

Properties of the weak stochastic and approximation error

The **weak stochastic error** $S_{t,\lambda}$ satisfies

- $S_{0,\lambda} = 0$, $S_{D,\lambda} = \|\xi\|^2$,
- $t \mapsto S_{t,\lambda}$ is nondecreasing.

The **weak approximation error** $A_{t,\lambda}$ satisfies

- $A_{0,\lambda} = \|g\|^2$, $A_{D,\lambda} = 0$,
- $A_{t,\lambda} \leqslant \|r_{t,<}^{1/2} g\|^2$, where $t \mapsto \|r_{t,<}^{1/2} g\|^2$ is nonincreasing.

Properties of the weak stochastic and approximation error

The **weak stochastic error** $S_{t,\lambda}$ satisfies

- $S_{0,\lambda} = 0$, $S_{D,\lambda} = \|\xi\|^2$,
- $t \mapsto S_{t,\lambda}$ is nondecreasing.

The **weak approximation error** $A_{t,\lambda}$ satisfies

- $A_{0,\lambda} = \|g\|^2$, $A_{D,\lambda} = 0$,
- $A_{t,\lambda} \leqslant \|r_{t,<}^{1/2} g\|^2$, where $t \mapsto \|r_{t,<}^{1/2} g\|^2$ is nonincreasing.

- r_t is nonnegative, decreasing, convex and log-concave on $[0, x_{1,t}]$.
- $R_t^2 := \|r_t Y\|^2 \leqslant \|r_{t,<}^{1/2} Y\|^2$ (cf. Nemirovskii 1986)
- $t \mapsto R_t^2$ is monotonically decreasing.

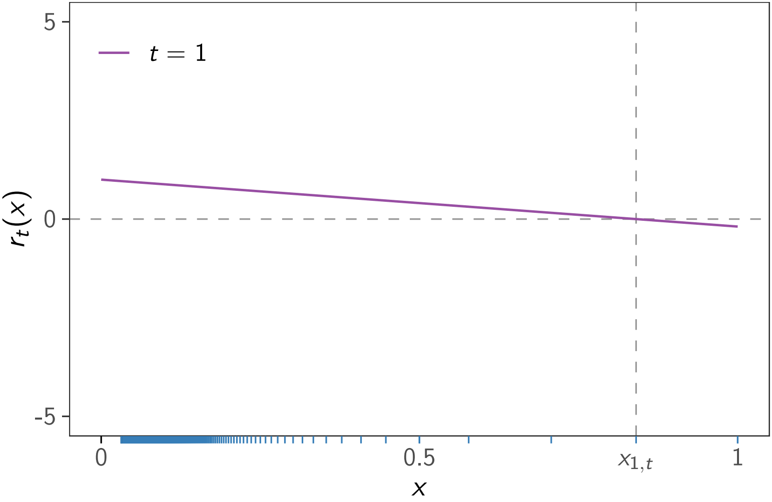


Figure: Residual polynomial with smallest zero $x_{1,t}$ and singular values of A

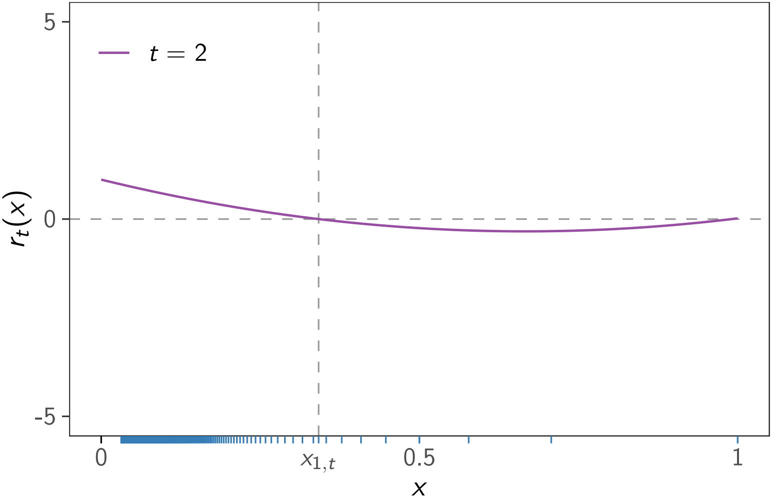


Figure: Residual polynomial with smallest zero $x_{1,t}$ and singular values of A

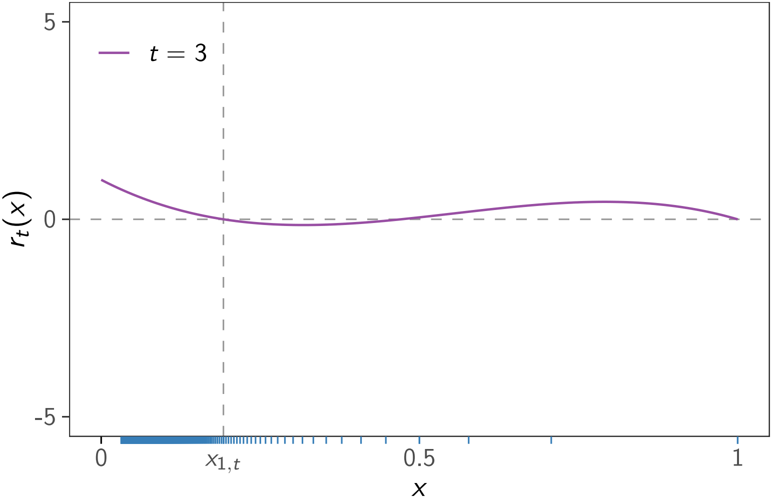


Figure: Residual polynomial with smallest zero $x_{1,t}$ and singular values of A

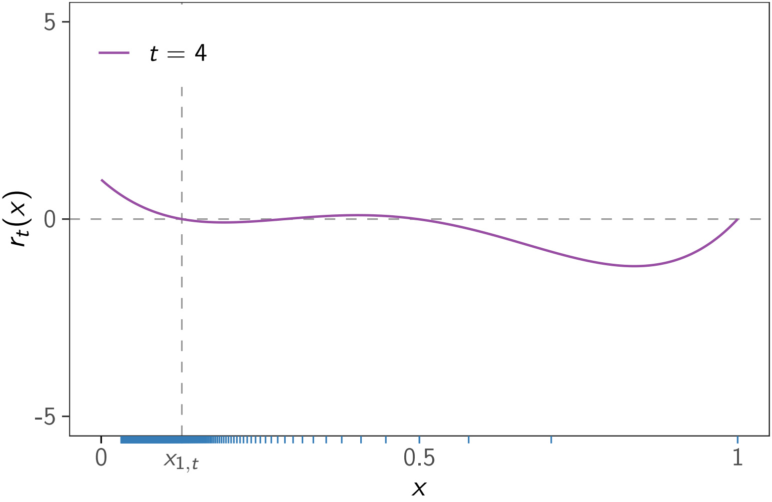


Figure: Residual polynomial with smallest zero $x_{1,t}$ and singular values of A

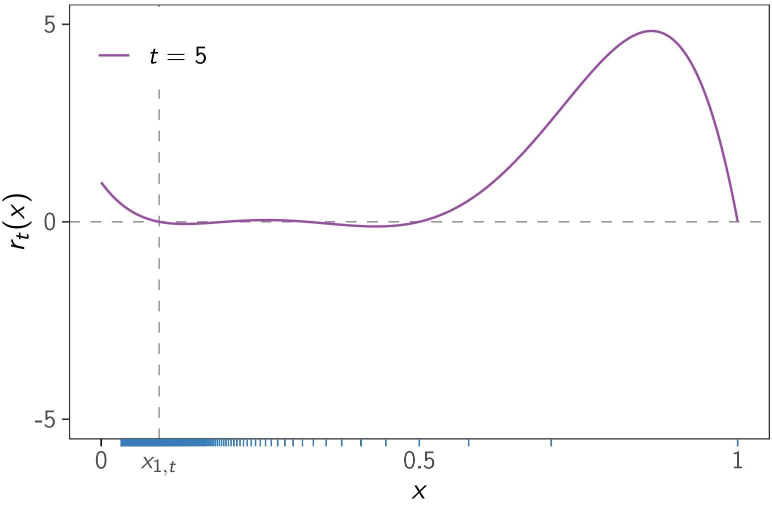


Figure: Residual polynomial with smallest zero $x_{1,t}$ and singular values of A

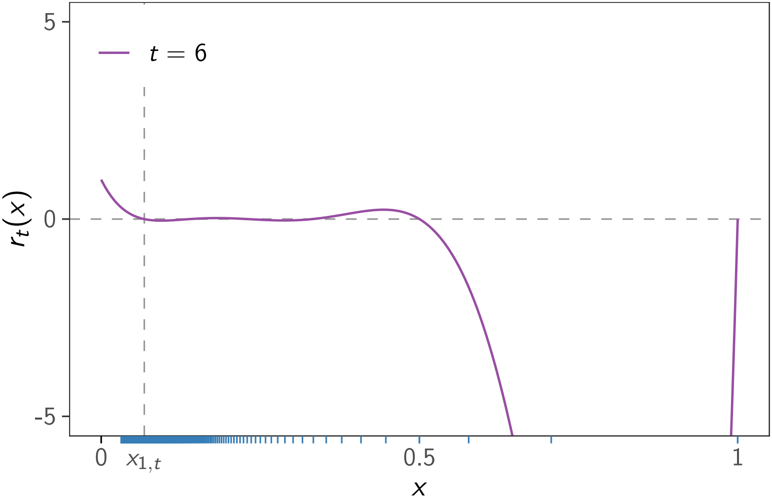


Figure: Residual polynomial with smallest zero $x_{1,t}$ and singular values of A

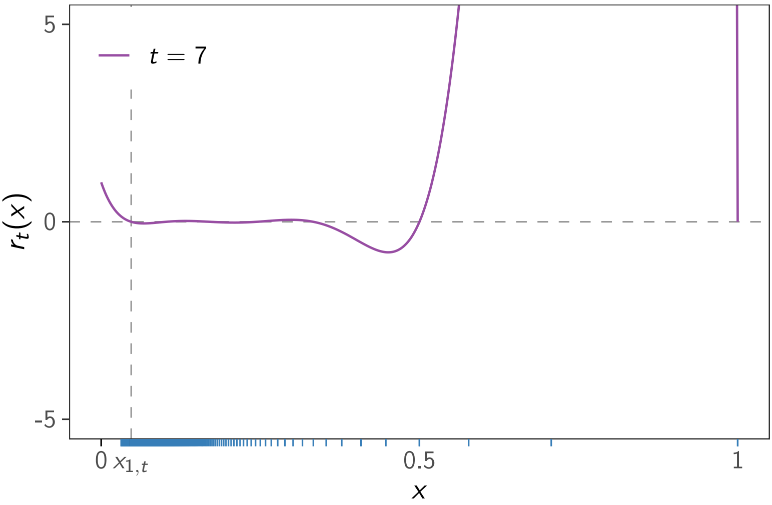


Figure: Residual polynomial with smallest zero $x_{1,t}$ and singular values of A

A weak oracle minimax bound

Weakly balanced oracle

$$\tau_{\mathfrak{w}} := \inf \{ t \in [0, D] \mid A_{t,\lambda} \leqslant S_{t,\lambda} \}$$

A weak oracle minimax bound

Weakly balanced oracle

$$\tau_{\mathfrak{w}} := \inf \{ t \in [0, D] \mid A_{t,\lambda} \leqslant S_{t,\lambda} \}$$

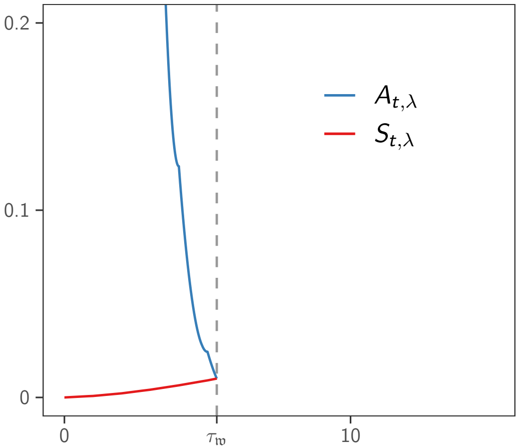
Proposition

Suppose g satisfies $\|g\|_{\mu+1/2}^2 := \sum_{i=1}^{D} \lambda_i^{-4(\mu+1/2)} \langle g, u_i \rangle^2 \leqslant R^2$, where $\mu, R > 0$, and the singular values are $\lambda_i = i^{-p}$, $i = 1, \dots, D$, $p > 1/2$. Then

$$\mathbb{E}\left[\|\widehat{g}_{\tau_{\mathfrak{w}}} - g\|^2 \right] \leqslant C_{p,\mu} R^{2/(4\mu p + 2p + 1)} \delta^{(8\mu p + 4p)/(4\mu p + 2p + 1)}.$$

This rate is **minimax optimal** (cf. Johnstone 2017).
In particular, CGNE is as good as truncated SVD.

Mimic the weakly balanced oracle by early stopping



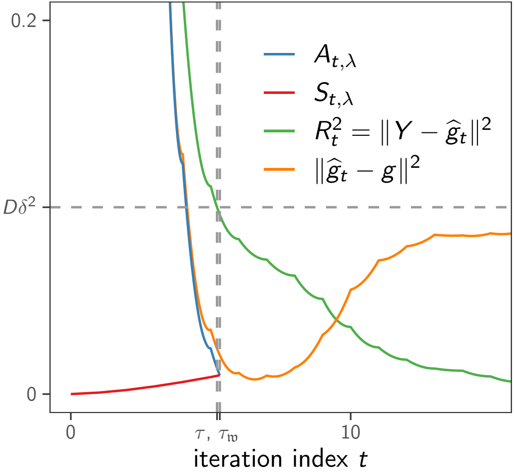
Weakly balanced oracle

$$\tau_{\mathfrak{w}} = \inf \{ t \in [0, D] \mid A_{t,\lambda} \leqslant S_{t,\lambda} \}$$

Equivalent formulation

$$\tau_{\mathfrak{w}} = \inf \{ t \mid R_t^2 \leqslant \|\xi\|^2 + 2\langle \xi, r_{t,<g} \rangle \}$$

Mimic the weakly balanced oracle by early stopping



Weakly balanced oracle

$\tau_{\mathfrak{w}} = \inf \{t \in [0, D] \mid A_{t,\lambda} \leqslant S_{t,\lambda}\}$

Equivalent formulation

$\tau_{\mathfrak{w}}$
$= \inf \{t \mid R_t^2 \leqslant \|\xi\|^2 + 2\langle \xi, r_{t,<} g\rangle\}$

Early stopping rule

$\tau := \inf \{t \in [0, D] \mid R_t^2 \leqslant D\delta^2\}$

Balanced oracle inequality for the prediction error

Lemma

We have

$$\mathbb{E}\left[\|\widehat{g}_{\tau} - \widehat{g}_{\tau_{\mathfrak{w}}}\|^2\right] \leqslant 2\mathbb{E}\left[|\langle \xi, r_{\tau_{\mathfrak{w}},<} g\rangle|\right] + \delta^2\sqrt{2D}.$$

Balanced oracle inequality for the prediction error

Lemma

We have

$$\mathbb{E}\left[\|\widehat{g}_\tau - \widehat{g}_{\tau_{\mathfrak{w}}}\|^2\right] \leqslant 2\mathbb{E}\left[|\langle \xi, r_{\tau_{\mathfrak{w}},<} g\rangle|\right] + \delta^2\sqrt{2D}.$$

Theorem

We have

$$\mathbb{E}\left[|\langle \xi, r_{\tau_{\mathfrak{w}},<} g\rangle|\right] \leqslant C\delta^2\left(\mathbb{E}\left[\delta^{-2} S_{\tau_{\mathfrak{w}},\lambda}\right] + \mathbb{E}\left[\delta^{-2} S_{\tau_{\mathfrak{w}},\lambda}\right]^{1/2}\sqrt{\log D}\right).$$

This implies

$$\mathbb{E}\left[\|\widehat{g}_\tau - g\|^2\right] \leqslant C\left(\mathbb{E}\left[S_{\tau_{\mathfrak{w}},\lambda}\right] + \delta^2\sqrt{D}\right).$$

Minimax optimality of early stopping for the prediction error

Corollary

Suppose g satisfies $\|g\|^2_{\mu+1/2} \leqslant R^2$ and the singular values are $\lambda_i = i^{-p}$, $p > 1/2$. Then

$$\mathbb{E}\left[\|\widehat{g}_\tau - g\|^2\right] \leqslant C_{p,\mu}\left(R^{2/(4\mu p+2p+1)}\delta^{(8\mu p+4p)/(4\mu p+2p+1)} + \delta^2\sqrt{D}\right).$$

This gives the minimax rate for all regularities $\mu > 0$ with $D^{\mu p+p/2+1/4} \lesssim \delta^{-1}$.

Transfer to the reconstruction error

Definition

$$\widehat{f}_t := A^\top (AA^\top)^{-1} \widehat{g}_t = A^\top (AA^\top)^{-1} (1 - r_t(AA^\top)) Y$$

Transfer to the reconstruction error

Definition

$$\widehat{f}_t := A^\top (AA^\top)^{-1} \widehat{g}_t = A^\top (AA^\top)^{-1} (1 - r_t(AA^\top)) Y$$

Lemma (Bound on the reconstruction error)

We have *(cf. Engl et al. 2000, Lemma 7.11)*

$$\|\widehat{f}_t - f\|^2 \leqslant 2\|r_{t,<}(A^\top A) f\|^2 + 4|r_t'(0)| S_{t,\lambda} + 2|r_t'(0)| A_{t,\lambda}.$$

Minimax optimality for the reconstruction error

Early stopping rule

$$\tau = \inf\left\{t \in [0, D] \;\middle|\; R_t^2 \leqslant D\delta^2\right\}$$

Theorem

Suppose f *satisfies* $\|f\|_\mu^2 := \sum_{i=1}^D \lambda_i^{-4\mu} \langle f, u_i\rangle^2 \leqslant R^2$*, where* $\mu, R > 0$*, and the singular values are* $\lambda_i = i^{-p}$*,* $i = 1, \ldots, D$*,* $p > 1/2$*. Then*

$$\mathbb{E}\left[\|\widehat{f}_\tau - f\|^2\right] \leqslant C_{p,\mu}\left(R^{(4p+2)/(4\mu p+2p+1)}\delta^{8\mu p/(4\mu p+2p+1)} + \delta^2 D^{p+1/2}\right).$$

Minimax optimality for the reconstruction error

Early stopping rule

$$\tau = \inf \left\{ t \in [0, D] \;\middle|\; R_t^2 \leqslant D\delta^2 \right\}$$

Theorem

Suppose f satisfies $\|f\|_\mu^2 := \sum_{i=1}^D \lambda_i^{-4\mu} \langle f, u_i \rangle^2 \leqslant R^2$, where $\mu, R > 0$, and the singular values are $\lambda_i = i^{-p}$, $i = 1, \dots, D$, $p > 1/2$. Then

$$\mathbb{E}\left[\|\widehat{f}_\tau - f\|^2 \right] \leqslant C_{p,\mu} \left(R^{(4p+2)/(4\mu p+2p+1)} \delta^{8\mu p/(4\mu p+2p+1)} + \delta^2 D^{p+1/2} \right).$$

✓ Early stopping at the data-driven τ achieves **computational and statistical efficiency**.

A numerical illustration for the prediction error

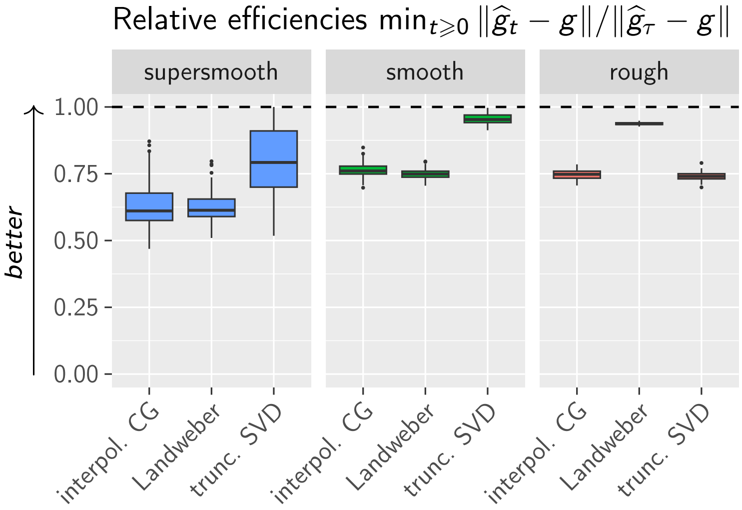


Figure: 100 Monte-Carlo runs, $\lambda_i = i^{-1/2}$, $\delta = 0.01$, $D = P = 1000$, for the signals of interest see Stankewitz (2020)

A numerical illustration for the prediction error

	interpol. CG	Landweber	trunc. SVD
supersmooth	5.32	30.70	39.76
smooth	12.64	270.09	309.81
rough	17.27	905.43	908.49

Table: Means of the early stopping rules

A numerical illustration for the prediction error

	interpol. CG	Landweber	trunc. SVD
supersmooth	5.32	30.70	39.76
smooth	12.64	270.09	309.81
rough	17.27	905.43	908.49

Table: Means of the early stopping rules

Thanks a lot for your attention!

Blanchard, G. and Mathé, P. (2012). “Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration”. In: *Inverse Problems* 28.11, p. 115011.

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

Blanchard, G., Hoffmann, M., and Reiß, M. (2018b). “Optimal Adaptation for Early Stopping in Statistical Inverse Problems”. In: *SIAM/ASA Journal on Uncertainty Quantification* 6.3, pp. 1043–1075.

Blanchard, G. and Krämer, N. (2010). “Optimal Learning Rates for Kernel Conjugate Gradient Regression”. In: *Proceedings of the 23rd International Conference on Neural Information Processing Systems*. Vol. 1. Red Hook, NY, USA: Curran Associates Inc., pp. 226–234.

Cavalier, L. (2011). "Inverse Problems in Statistics". In: *Inverse Problems and High-Dimensional Estimation*. Ed. by P. Alquier, E. Gautier, and G. Stoltz. Vol. 203. Lecture Notes in Statistics. Berlin, Heidelberg: Springer, pp. 3–96.

Celisse, A. and Wahl, M. (2021). "Analyzing the discrepancy principle for kernelized spectral filter learning algorithms". In: *Journal of Machine Learning Research* 22.76, pp. 1–59.

Engl, H. W., Hanke, M., and Neubauer, A. (2000). *Regularization of Inverse Problems*. Vol. 375. Mathematics and Its Applications. Dordrecht: Kluwer Academic Publishers.

Hestenes, M. R. and Stiefel, E. (1952). "Methods of Conjugate Gradients for Solving Linear Systems". In: *Journal of Research of the National Bureau of Standards* 49.6, pp. 409–436.

- Johnstone, I. M. (2017). *Gaussian estimation: Sequence and wavelet models. Draft version.* URL: `https://imjohnstone.su.domains//GE_08_09_17.pdf` (visited on 01/16/2023).
- Nemirovskii, A. S. (1986). "The regularizing properties of the adjoint gradient method in ill-posed problems". In: *USSR Computational Mathematics and Mathematical Physics* 26.2, pp. 7–16.
- Stankewitz, B. (2020). "Smoothed residual stopping for statistical inverse problems via truncated SVD estimation". In: *Electronic Journal of Statistics* 14.2, pp. 3396–3428.