# Simultaneous adaptation for several criteria using an extended Lepskii principle

G. Blanchard

Université Paris-Saclay

Workshop “Statistique et Informatique pour la science des données”, IHES
27/01/2020

Based on joint work with: P. Mathé (Weierstrass Institute, Berlin), N. Mücke (U. Stuttgart, TU Berlin)



# Setting: linear regression in Hilbert space

We consider the observation model

$$Y_i = \langle f_\circ, X_i \rangle + \xi_i,$$

where

- $X_i$ takes its values in a Hilbert space $\mathcal{H}$, with $\|X_i\| \leq 1$ a.s.;
- $\xi_i$ is a random variable with $\mathbb{E}[\xi_i | X_i] = 0$, $\mathbb{E}\left[\xi^2 | X_i\right] \leq \sigma^2$, $|\xi| \leq M$ a.s.;
- $(X_i, \xi_i)_{1 \leq i \leq n}$ are i.i.d.

The goal is to estimate $f_\circ$ (in a sense to be specified) from the data.
Note that if $\dim(\mathcal{H}) = \infty$, this is essentially a non-parametric model.

# Why this model?

- Hilbert-space valued variables appear in standard models of **Functional Data Analysis**, where the observed data are modeled (idealized) as function-valued.
- Such models also appear in **reproducing kernel Hilbert space (RKHS) methods** in machine learning:
  - assume observations $X_i$ take valued in some space $\mathcal{X}$
  - let $\Phi : \mathcal{X} \to \mathcal{H}$ be a "feature mapping" in a Hilbert space $\mathcal{H}$, and $\widetilde{X} = \Phi(X)$, then one considers the model

    $$Y_i = \langle f_\circ, \widetilde{X}_i \rangle + \xi_i = \widetilde{f}_\circ(X_i) + \xi_i,$$

    where $\widetilde{f} \in \widetilde{H} := \{x \mapsto \langle f, \Phi(x) \rangle; f \in \mathcal{H}\}$ is a nonparametric model of functions (nonlinear in $x$!).
  - Usually all computations don't require explicit knowledge of $\Phi$ but only access to the **kernel** $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$.



# Why this model (II) - inverse learning

Of interest is also the **inverse learning** problem:

- $X_i$ takes value in $\mathcal{X}$;
- if $A$ is a linear operator from a Hilbert space $\mathcal{H}$ to a real function space on $\mathcal{X}$;
- inverse regression learning model:

  $$Y_i = (Af_\circ)(X_i) + \xi_i.$$

- If $A$ is a Carleman operator (i.e. evaluation functionals $f \mapsto (Af)(x)$ are continuous for all $x$), then this can be isometrically reduced to a reproducing kernel learning setting (De Vito, Rosasco, Caponnetto 2006; Blanchard and Mücke, 2017).

# Two notions of risk

We will consider two notions of error (risk) for a candidate estimate $\widehat{f}$ of $f_\circ$:

- Squared prediction error:

$$\mathcal{E}(\widehat{f}) := \mathbb{E}\left[\left(\langle \widehat{f}, X\rangle - Y\right)^2\right].$$

- The associated (excess error) risk is

$$\mathcal{E}(\widehat{f}) - \mathcal{E}(f_\circ) = \mathbb{E}\left[\left(\left\langle \widehat{f} - f_\circ, X\right\rangle\right)^2\right] = \left\|\widehat{f}^* - f_\circ^*\right\|_{2,X}^2,$$

- Reconstruction error risk:

$$\left\|\widehat{f} - f_\circ\right\|_{\mathcal{H}}^2.$$

The goal is to find a suitable estimator $\widehat{f}$ of $f_\circ$ from the data having "optimal" convergence properties with respect to these two risks.



# Finite-dimensional case

- The final dimensional case: $\mathcal{X} = \mathbb{R}^p$, $f_\circ$ now denoted $\beta_\circ$
- In usual matrix form:

$$\boldsymbol{Y} = \boldsymbol{X}\beta_\circ + \boldsymbol{\xi}.$$

  - $X_i^T$ form the lines of the $(n, p)$ design matrix $\boldsymbol{X}$
  - $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$
  - $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^T$
- "Reconstruction" risk corresponds to $\left\|\beta_\circ - \widehat{\beta}\right\|^2$.
- Prediction risk corresponds to

$$\mathbb{E}\left[\langle \beta_\circ - \widehat{\beta}, X\rangle^2\right] = \left\|\Sigma^{1/2}(\beta_\circ - \widehat{\beta})\right\|^2,$$

  where $\Sigma := \mathbb{E}\left[XX^T\right]$.
- In Hilbert space, same relation with $\Sigma := \mathbb{E}[X \otimes X^*]$.

# The founding fathers of machine learning

A.M. Legendre

C.F. Gauß

The "ordinary" least squares (OLS) solution:

$$\widehat{\beta}_{OLS} = (X^T X)^{-1} X^T Y.$$



# Convergence of OLS in finite dimension

- The "ordinary" least squares (OLS) solution:
$$\widehat{\beta}_{OLS} = (X^T X)^{-1} X^T Y.$$
- We want to understand the behavior of $\widehat{\beta}_{OLS}$, when the data size $n$ grows large. Will we be close to the truth $\beta_\circ$?
- Recall
$$\widehat{\beta}_{OLS} = \left(X^T X\right)^{-1} X^T Y = \Big(\underbrace{\frac{1}{n} X^T X}_{:=\widehat{\Sigma}}\Big)^{-1} \Big(\underbrace{\frac{1}{n} X^T Y}_{:=\widehat{\gamma}}\Big) = \widehat{\Sigma}^{-1} \widehat{\gamma},$$
- Observe by a vectorial LLN, as $n \to \infty$:
$$\widehat{\Sigma} := \frac{1}{n} X^T X = \frac{1}{n} \sum_{i=1}^{n} \underbrace{X_i X_i^T}_{=:Z_i'} \longrightarrow \mathbb{E}\left[X_1 X_1^T\right] =: \Sigma;$$
$$\widehat{\gamma} := \frac{1}{n} X^T Y = \frac{1}{n} \sum_{i=1}^{n} \underbrace{X_i Y_i}_{=:Z_i} \longrightarrow \mathbb{E}[X_1 Y_1] = \Sigma \beta_\circ =: \gamma;$$
- Hence $\widehat{\beta} = \widehat{\Sigma}^{-1} \widehat{\gamma} \to \Sigma^{-1} \gamma = \beta_\circ$. (Assuming $\Sigma$ invertible.)

# From OLS to Hilbert-space regression

- For ordinary linear regression with $\mathcal{X} = \mathbb{R}^p$ (fixed $p$, $n \to \infty$):
  - LLN implies $\widehat{\beta}_{OLS} (= \widehat{\Sigma}^{-1}\widehat{\gamma}) \to \beta_\circ (= \Sigma^{-1}\gamma)$;
  - CLT+Delta Method imply asymptotic normality and convergence in $\mathcal{O}(n^{-\frac{1}{2}})$.
- **How to generalize to $\mathcal{X} = \mathcal{H}$?**
- **Main issue:** $\Sigma = \mathbb{E}[X \otimes X^*]$ does not have a continuous inverse. ($\to$ ill-posed problem)
- Need to consider a suitable approximation $\zeta(\widehat{\Sigma})$ of $\Sigma^{-1}$ **(regularization)**, where

$$\widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^{m} X_i \otimes X_i^*$$

is the empirical second moment operator.



# Regularization methods

- Main idea: replace $\widehat{\Sigma}^{-1}$ by an approximate inverse, such as
- **Ridge regression/Tikhonov**:

$$\widehat{f}_{Ridge(\lambda)} = (\widehat{\Sigma} + \lambda I_p)^{-1}\widehat{\gamma}$$

- **PCA projection/spectral cut-off**: restrict $\widehat{\Sigma}$ on its $k$ first eigenvectors

$$\widehat{f}_{PCA(k)} = (\widehat{\Sigma})^{-1}_{|k}\widehat{\gamma}$$

- **Gradient descent/Landweber Iteration/$L^2$ boosting**:

$$\widehat{f}_{LW(k)} = \widehat{f}_{LW(k-1)} + (\widehat{\gamma} - \widehat{\Sigma}\widehat{f}_{LW(k-1)})$$
$$= \sum_{i=0}^{k}(I - \widehat{\Sigma})^k\widehat{\gamma},$$

(assuming $\|\widehat{\Sigma}\|_{op} \le 1$).

# General form spectral linearization

Bauer, Rosasco, Pereverzev 2007

- **General form** regularization method:

$$\widehat{f}_\lambda = \zeta_\lambda(\widehat{\Sigma})\widehat{\gamma}$$

  for some well-chosen function $\zeta_\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ acting on the spectrum and "approximating" the function $x \mapsto x^{-1}$.

- $\lambda > 0$: regularization parameter; $\lambda \to 0 \Leftrightarrow$ less regularization
- Notation of (autoadjoint) functional calculus, i.e.

$$\widehat{\Sigma} = Q^T \mathrm{diag}(\mu_1, \mu_2, \ldots) Q \Rightarrow \zeta(\widehat{\Sigma}) := Q^T \mathrm{diag}(\zeta(\mu_1), \zeta(\mu_2), \ldots) Q$$

- Examples (revisited):
  - **Tikhonov**: $\zeta_\lambda(t) = (t+\lambda)^{-1}$
  - **Spectral cut-off**: $\zeta_\lambda(t) = t^{-1}\mathbf{1}\{t \geq \lambda\}$
  - **Landweber iteration**: $\zeta_k(t) = \sum_{i=0}^{k}(1-t)^i$.



# Assumptions on regularization function

Standard assumptions on the regularization family $\zeta_\lambda : [0,1] \to \mathbb{R}$ are:

(i) There exists a constant $D < \infty$ such that

$$\sup_{0<\lambda\leq 1} \sup_{0<t\leq 1} |t\zeta_\lambda(t)| \leq D\,,$$

(ii) There exists a constant $E < \infty$ such that

$$\sup_{0<\lambda\leq 1} \sup_{0<t\leq 1} \lambda|\zeta_\lambda(t)| \leq E\,,$$

(iii) **Qualification:** for **residual** $r_\lambda(t) := 1 - t\zeta_\lambda(t)$,

$$\forall \lambda \leq 1: \qquad \sup_{0<t\leq 1} |r_\lambda(t)| t^\nu \leq \gamma_\nu \lambda^\nu,$$

holds for $\nu = 0$ and $\nu = q > 0$.

# Structural Assumptions (I)

- Denote $(\mu_i)_{i\geq 1}$ the sequence of positive eigenvalues of $\Sigma$ in nonincreasing order.
- **Assumptions on spectrum decay:** for $s \in (0,1); \alpha > 0$:

$$\mathbf{IP}^{<}(s,\alpha): \quad \mu_i \leq \alpha i^{-\frac{1}{s}}$$

- This implies quantitative estimates of the "effective dimension"

$$\mathcal{N}(\lambda) := \mathrm{Tr}(\ (\Sigma+\lambda)^{-1}\Sigma\ ) \lesssim \lambda^{-s}.$$



# Structural Assumptions (II)

- Denote $(\mu_i)_{i\geq 1}$ the sequence of positive eigenvalues of $\Sigma$ in nonincreasing order.
- **Source condition** for the signal: for $r > 0$, define

$$\mathbf{SC}(r,R): \quad f_\circ = \Sigma^r h_\circ \text{ for some } h_\circ \text{ with } \|h_\circ\| \leq R,$$

or equivalently, as a **Sobolev-type regularity**

$$\mathbf{SC}(r,R): \quad f_\circ \in \left\{ f \in \mathcal{H} : \sum_{i\geq 1} \mu_i^{-2r} f_i^2 \leq R^2 \right\},$$

where $f_i$ are the coefficients of $h$ in the eigenbasis of $\Sigma$.

- Under $(\mathbf{SC})(r,R)$ it is assumed that the **qualification** $q$ of the regularization method satisfies $q \geq r + \frac{1}{2}$.

# A general upper bound risk estimate

**Theorem**

*Assume the source condition* (**SC**)$(r, R)$ *holds.*
*If* $\lambda$ *is such that* $\lambda \gtrsim (\mathcal{N}(\lambda) \vee \log(\eta)^2)/n$, *then with probability at least* $1-\eta$, *it holds:*

$$\left\|(\Sigma+\lambda)^{1/2}\left(f_\circ - \widehat{f}_\lambda\right)\right\|_{\mathcal{H}} \lesssim \log(\eta)^2\left(R\lambda^{r+\frac{1}{2}} + \sigma\sqrt{\frac{\mathcal{N}(\lambda)}{n}} + \frac{1}{n\sqrt{\lambda}} + \mathcal{O}(n^{-\frac{1}{2}})\right).$$

This gives rise to estimates in both norms of interest since

$$\left\|f_\circ - \widehat{f}_\lambda\right\|_{\mathcal{H}} \leq \lambda^{-\frac{1}{2}}\left\|(\Sigma+\lambda)^{1/2}\left(f_\circ - \widehat{f}_\lambda\right)\right\|_{\mathcal{H}},$$

and

$$\left\|f_\circ^* - \widehat{f}_\lambda^*\right\|_{L^2(P_X)} = \left\|\Sigma^{\frac{1}{2}}(f_\circ - \widehat{f}_\lambda)\right\|_{\mathcal{H}} \leq \left\|(\Sigma+\lambda)^{1/2}\left(f_\circ - \widehat{f}_\lambda\right)\right\|_{\mathcal{H}}.$$



# Upper bound on rates

Optimizing the obtained bound over $\lambda$ (i.e. balancing the main terms) one obtains

**Theorem**

*Assume* $r, R, s, \alpha$ *are fixed positive constants and assume* $\mathbb{P}_{XY}$ *satisfies* **(IP$^<$)**$(s,\alpha)$, **(SC)**$(r, R)$ *and* $\|X\| \leq 1$, $\|Y\| \leq M$, $\mathrm{Var}[Y|X]_\infty \leq \sigma^2$ *a.s. Define*

$$\widehat{\beta}_n = \zeta_{\lambda_n}(\widehat{\Sigma})\widehat{\gamma},$$

*using a regularization family* $(\zeta_\lambda)$ *satisfying the standard assumptions with qualification* $q \geq r + \frac{1}{2}$, *and the parameter choice rule*

$$\lambda_n = (R^2\sigma^2/n)^{-\frac{1}{2r+1+s}}.$$

*Then it holds for any* $p \geq 1$:

$$\limsup_{n\to\infty} \mathbb{E}^{\otimes n}\left(\left\|f_\circ - \widehat{f}_{\lambda_n}\right\|^p\right)^{1/p} \Big/ R\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{r}{2r+1+s}} \leq C_\blacktriangle;$$

$$\limsup_{n\to\infty} \mathbb{E}^{\otimes n}\left(\left\|f_\circ^* - \widehat{f}_{\lambda_n}\right\|_{2,X}^p\right)^{1/p} \Big/ R\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{r+1/2}{2r+1+s}} \leq C_\blacktriangle.$$

# Towards adaptivity: existing approaches

- Cross-validation (or hold-out) will yield a tuning of the parameter which is **adaptive in the prediction risk**, it is based on a unbiased estimate of the risk **(URE)** principle.
- Standard Lepski's principle parameter selection can be applied for any fixed norm (provided a good estimate of the "variance" term $\sigma\sqrt{\mathcal{N}(\lambda)/n}$ is available)
- Despite the **existence** of a regularization parameter $\lambda$ being optimal for both norms, there is no guarantee that **any** (close to) optimal parameter for prediction risk (eg. selected by cross-validation) will be close to optimal in reconstruction risk, or vice-versa.
- We want to construct a **simultaneously (for both norms) adaptive** data-driven parameter selection.



# Generalized Lepskii's principle

We consider the following "deterministic" assumption to highlight the construction.

**Assumption**

Let $\Lambda \subset \mathbb{R}_+$ be a finite set of candidate regularization parameters,

$$\Lambda := \{\lambda_j, \quad \lambda_0 > \lambda_1 > \ldots > \lambda_m = \lambda_{\min} > 0\},$$

The (known) family of elements of $\mathcal{H}$, $(f_\lambda)_{\lambda\in\Lambda}$, satisfies for any $\lambda \in \Lambda$:

$$\left\| (\Sigma + \lambda)^{1/2}(f_\circ - f_\lambda) \right\|_{\mathcal{H}} \leq C\sqrt{\lambda}(\mathcal{A}(\lambda) + \mathcal{S}(\lambda)),$$

where

- the function $\lambda \in \Lambda \mapsto \mathcal{A}(\lambda) \in \mathbb{R}_+$ is **non-decreasing** with $\mathcal{A}(0) = 0$ and possibly **unknown**;
- the function $\lambda \in \Lambda \mapsto \sqrt{\lambda}\mathcal{S}(\lambda) \in \mathbb{R}_+$ is **non-increasing** and **known**.

# Generalized Lepskii's principle (II)

- Set

$$\mathcal{M}(\Lambda) := \left\{ \lambda \in \Lambda \ : \ \left\| (\Sigma + \lambda')^{1/2} (f_\lambda - f_{\lambda'}) \right\|_{\mathcal{H}} \le 4C\sqrt{\lambda'}\mathcal{S}(\lambda'), \quad \forall \lambda' \in \Lambda, \text{ s.t. } \lambda' \le \lambda \right\}.$$

- The balancing parameter is given as

$$\hat{\lambda} := \max \ \mathcal{M}(\Lambda) \ ;$$

(this quantity is always well-defined since $\lambda_{\min} \in \mathcal{M}(\Lambda)$.)



# Generalized Lepskii's principle: bound

**Theorem**

*Under the assumptions made previously, if*

$$\lambda_* := \max\{\lambda \in \Lambda : \mathcal{A}(\lambda) \le \mathcal{S}(\lambda)\},$$

*and* $\widehat{\lambda}$ *is the parameter choice defined previously, then:*

- *It holds*

$$\left\| (\Sigma + \lambda_*)^{\frac{1}{2}} (f_\circ - f_{\widehat{\lambda}}) \right\|_{\mathcal{H}} \lesssim \sqrt{\lambda_*}\mathcal{S}(\lambda_*);$$

- *Assuming it holds* $\mathcal{S}(\lambda_k) \le C_{\mathcal{S}}\mathcal{S}(\lambda_{k-1})$ *for* $k = 1, \dots, m$*, then:*

$$\left\| f_\circ - f_{\widehat{\lambda}} \right\|_{\mathcal{H}} \lesssim \min_{\lambda \in \Lambda} (\mathcal{A}(\lambda) + \mathcal{S}(\lambda));$$

$$\left\| \Sigma^{\frac{1}{2}} (f_\circ - f_{\widehat{\lambda}}) \right\|_{\mathcal{H}} \lesssim \min_{\lambda \in \Lambda} \sqrt{\lambda}(\mathcal{A}(\lambda) + \mathcal{S}(\lambda)).$$

# Applying Lepski's principle

Looking at the main error bound obtained earlier, with high probability the assumption

$$\left\| (\Sigma + \lambda)^{1/2}(f_\circ - f_\lambda) \right\|_{\mathcal{H}} \leq C\sqrt{\lambda}(\mathcal{A}(\lambda) + \mathcal{S}(\lambda))$$

is satisfied with

$$\mathcal{A}(\lambda) := \left( R\lambda^{r+\frac{1}{2}} + \mathcal{O}(n^{-\frac{1}{2}}) \right),$$

$$\mathcal{S}(\lambda) := \sigma\sqrt{\frac{\mathcal{N}(\lambda)}{n}} + \frac{1}{n\sqrt{\lambda}}.$$

**Remaining issues:**

- $\Sigma$ is not known;
- $\mathcal{N}(\lambda) = \mathrm{Tr}(\ (\Sigma + \lambda)^{-1}\Sigma)$ is not known;
- the noise variance $\sigma^2$ might not be known (issue ignored for now).



# Replacing $\Sigma, \mathcal{N}(\lambda)$ by empirical quantities

**Proposition**

*If $\lambda$ is such that $\lambda \gtrsim (\mathcal{N}(\lambda) \vee \log(\eta)^2)\,/\,n$, then with probability at least $1 - \eta$, it holds:*

$$\left\| (\Sigma + \lambda)^{\frac{1}{2}}(\widehat{\Sigma} + \lambda)^{-\frac{1}{2}} \right\| \lesssim 1 + \log(\eta^{-1}).$$

**Proposition**

*If $\lambda \gtrsim n^{-1}$, it holds with probability at least $1 - \eta$, for $\widehat{\mathcal{N}}(\lambda) := \mathrm{Tr}(\widehat{\Sigma}(\widehat{\Sigma} + \lambda)^{-1})$:*

$$\max\left( \frac{\mathcal{N}(\lambda) \vee 1}{\widehat{\mathcal{N}}(\lambda) \vee 1}, \frac{\widehat{\mathcal{N}}(\lambda) \vee 1}{\mathcal{N}(\lambda) \vee 1} \right) \lesssim (1 + \log\eta^{-1})^2.$$

# Fully empirical procedure ($\sigma$, $M$ known)

- Put $L := 2\log(8\log n/(\eta\log q))$ and let

$$\widehat{\Lambda} := \left\{\lambda_i = q^{-i}, i \in \mathbb{N}, \text{ s.t. } \lambda_i \geq 100(\widehat{\mathcal{N}}(\lambda) \vee L^2/n)\right\}.$$

- Define the parameter choice

$$\widehat{\lambda} = \max\Big\{\lambda \in \widehat{\Lambda} : \forall \lambda' \in \widehat{\Lambda}, \text{ s.t. } \lambda' \leq \lambda :$$
$$\left\|(\widehat{\Sigma} + \lambda')^{\frac{1}{2}}(\widehat{f}_\lambda - \widehat{f}_{\lambda'})\right\| \leq cL\sqrt{\lambda'}\widehat{\mathcal{S}}(\lambda')\Big\},$$

where

$$\widehat{\mathcal{S}}(\lambda) := \frac{\sigma\sqrt{2(\widehat{\mathcal{N}}(\lambda) \vee 1) + M/5}}{\sqrt{\lambda n}}.$$



# Result for the empirical selection procedure

**Theorem**

*Assume the source condition* $(\mathbf{SC})(r, R)$ *holds.*
*Then for the generalized-Lepski parameter choice* $\widehat{\lambda}$*, with probability at least* $1 - \eta$*:*

$$\left\|(\Sigma + \lambda)^{\frac{1}{2}}(\widehat{f}_{\widehat{\lambda}} - f_\circ)\right\| \lesssim L^3 \min_{\lambda \in [\lambda_{\min}, 1]} \left(R\lambda^{r+\frac{1}{2}} + \sigma\sqrt{\frac{\mathcal{N}(\lambda)}{n}} + \frac{1}{n\sqrt{\lambda}} + \mathcal{O}(n^{-\frac{1}{2}})\right).$$

*where*

$$\lambda_{\min} = \min\left\{\lambda \in [0, 1] : \lambda \gtrsim (\mathcal{N}(\lambda) \vee L^2/n)\right\}.$$

**Conclusion**: as a direct byproduct we get the same rates (up to $\log\log n$ factor) as the optimal choice of $\lambda$ in the original bound, for **both norms of interest**.

# Can we estimate the noise variance $\sigma^2$?

- Observe that in general, there is no identifiability in the model

$$y_i = f(x_i) + \sigma\xi_i,$$

if the function $f$ can be "arbitrary".

- There is a hope when we assumed that $f$ has some regularity (here: linearity)

- **Idea:**
  - Take $\lambda$ small so that the "bias" $\mathcal{A}(\lambda)$ is expected to be much lower than the "variance" $\mathcal{S}(\lambda)$ (e.g., close to $\widehat{\lambda}_{\mathrm{min}}$.
  - Split the sample into two subsamples giving rise to $\widehat{f}_\lambda^{(1)}, \widehat{f}_\lambda^{(2)}$.
  - The hope is that by considering $\left\| \widehat{f}_\lambda^{(1)} - \widehat{f}_\lambda^{(2)} \right\|^2$ in a suitable norm, we cancel the bias and observe twice the "variance".

- Need somewhat precise concentration (upper and lower) for this quantity.



# Estimation of the variance $\sigma^2$

- Assume we have two independent sample of the same size $n$, giving rise to estimators $\widehat{f}_\lambda^{(1)}, \widehat{f}_\lambda^{(2)}$ (using the same regularization parameter $\lambda > 0$).

- Consider the statistic

$$\begin{aligned}\Delta^2 &:= \left\| \frac{1}{2}(\widehat{\Sigma}^{(1)} + \widehat{\Sigma}^{(2)} + \lambda)^{\frac{1}{2}}(\widehat{f}_\lambda^{(1)} - \widehat{f}_\lambda^{(2)}) \right\|_{\mathcal{H}}^2 \\ &= \frac{1}{2n}\sum_{i=1}^{2n}(\widehat{f}_\lambda^{(1)} - \widehat{f}_\lambda^{(2)})^2(x_i) + \lambda \left\| \widehat{f}_\lambda^{(1)} - \widehat{f}_\lambda^{(2)}) \right\|_{\mathcal{H}}^2,\end{aligned}$$

and (rescaling so that the estimate is approximately unbiased)

$$\widehat{\sigma}^2 := \frac{\Delta^2}{\sum_{i,j=1}^2 \|A_{ij}\|_{HS}^2},$$

where $A_{ij} = (\widehat{\Sigma}^{(i)} + \lambda)^{\frac{1}{2}} \zeta_\lambda(\widehat{\Sigma}^{(j)})(\widehat{\Sigma}^{(j)})^{\frac{1}{2}}$.

# Estimation of the variance $\sigma^2$

**Theorem**

*If $\lambda \geq \widehat{\lambda}_{min}$, where*

$$\widehat{\lambda}_{min} = \min\left\{\lambda > 0 : \lambda \geq 100(\widehat{\mathcal{N}}(\lambda) \vee \log(\eta^{-1})/2)\right\},$$

*then with probability at least $1-\eta$, it holds*

$$\widehat{\sigma}^2 \in \left[\sigma^2 \pm \left(\lambda\sigma^2 + F(\lambda)\log(\eta^{-1})\right)\right],$$

*with $F(\lambda) \to 0$ as $\lambda \to 0$.*

(Proof: based on previously established tools in this setting+Hanson-Wright inequality)

**Conclusion:** the estimator $\widehat{\sigma}^2$ is consistent, and can be used as a proxy for $\sigma^2$ in the procedure, with the same conclusions (up to changes in numerical constants, and for $n$ big enough).



## Thank you for your attention

G. Blanchard, P. Mathé, N. Mücke. Lepskii Principle in Supervised Learning. ArXiv 1902.05404
G. Blanchard, N. Mücke. Optimal Rates For Regularization Of Statistical Inverse Learning Problems. Foundations of Computational Mathematics , 2017.