Some applications of concentration inequalities to machine learning

from Hilbert space geometry to computational efficiency

G. Blanchard

Universtität Potsdam

Stochastic processes and statistical machine learning I, 14/02/2018

Based on joint work with: N. Mücke, O. Zadorozhnyi (U. Potsdam), N. Krämer (Staburo)

Mathematical model for statistical learning

- (Simplified) goal of a machine learning problem: predict a value $Y \in \mathcal{Y}$ (the "label") from observed data $X \in \mathcal{X}$ (the "input").
- Find a **prediction function** $f(X)$ as close to Y as possible.
  (In a sense to be specified)
- Data (X, Y) are modeled as **random**.
- In this talk: Y is real-valued (regression, $\mathcal{Y} = \mathbb{R}$).

Risk (=expected prediction error)

- Prediction will never be 100% perfect: we define a quantitative notion of error, and the **risk** as its expected value.

- Squared prediction risk (for Y real-valued):

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

- We want to find f so that $\mathcal{E}(f)$ is as small as possible.

Regression

- Under the square prediction risk the optimal prediction function is

$$f^*(x) = \mathbb{E}[Y|X = x],$$

the model is equivalently written as

$$Y_i = f^*(X_i) + \xi_i,$$

with $\mathbb{E}[\xi_i|X_i] = 0$ (ξ_i = "noise").

- **Note:** in this model, the **excess risk** of a predictor f with respect to the optimal f^* is

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

"Learning" from data

- We do not access exactly to $\mathcal{E}(f) := \mathbb{E}\left[(f(X) - Y)^2\right]$ (theoretical quantity)
- But we have observed data in a database: $(X_i, Y_i)_{i=1,\dots,n}$
  "Training data"
- $(X_i, Y_i)_{i=1,\dots,n}$ independent, identically distributed (i.i.d.) from $\mathbb{P}_{XY}$
- We can hope to approach $\mathbb{E}\left[(f(X) - Y)^2\right]$ by the averaged error on the database:

$$\widehat{\mathcal{E}}(f) := \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2$$

("empirical error").

Linear regression

- The linear case: $\mathcal{X} = \mathbb{R}^p$, $f^*(x) = f_{\beta_*}(x) = \langle x, \beta_* \rangle$.
- In usual matrix form:

$$Y = X\beta_* + \xi.$$

  - X_i^T form the lines of the (n, p) design matrix X
  - $Y = (Y_1, \ldots, Y_n)^T$
  - $\xi = (\xi_1, \ldots, \xi_n)^T$

- "Reconstruction" error corresponds to $\left\|\beta^* - \widehat{\beta}\right\|^2$.
- Prediction error corresponds to

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

  where $\Sigma := \mathbb{E}\left[XX^T\right]$.

THE FOUNDING FATHERS OF MACHINE LEARNING?

THE FOUNDING FATHERS OF MACHINE LEARNING?

A.M. Legendre

C.F. Gauß

THE FOUNDING FATHERS OF MACHINE LEARNING?

A.M. Legendre

C.F. Gauß

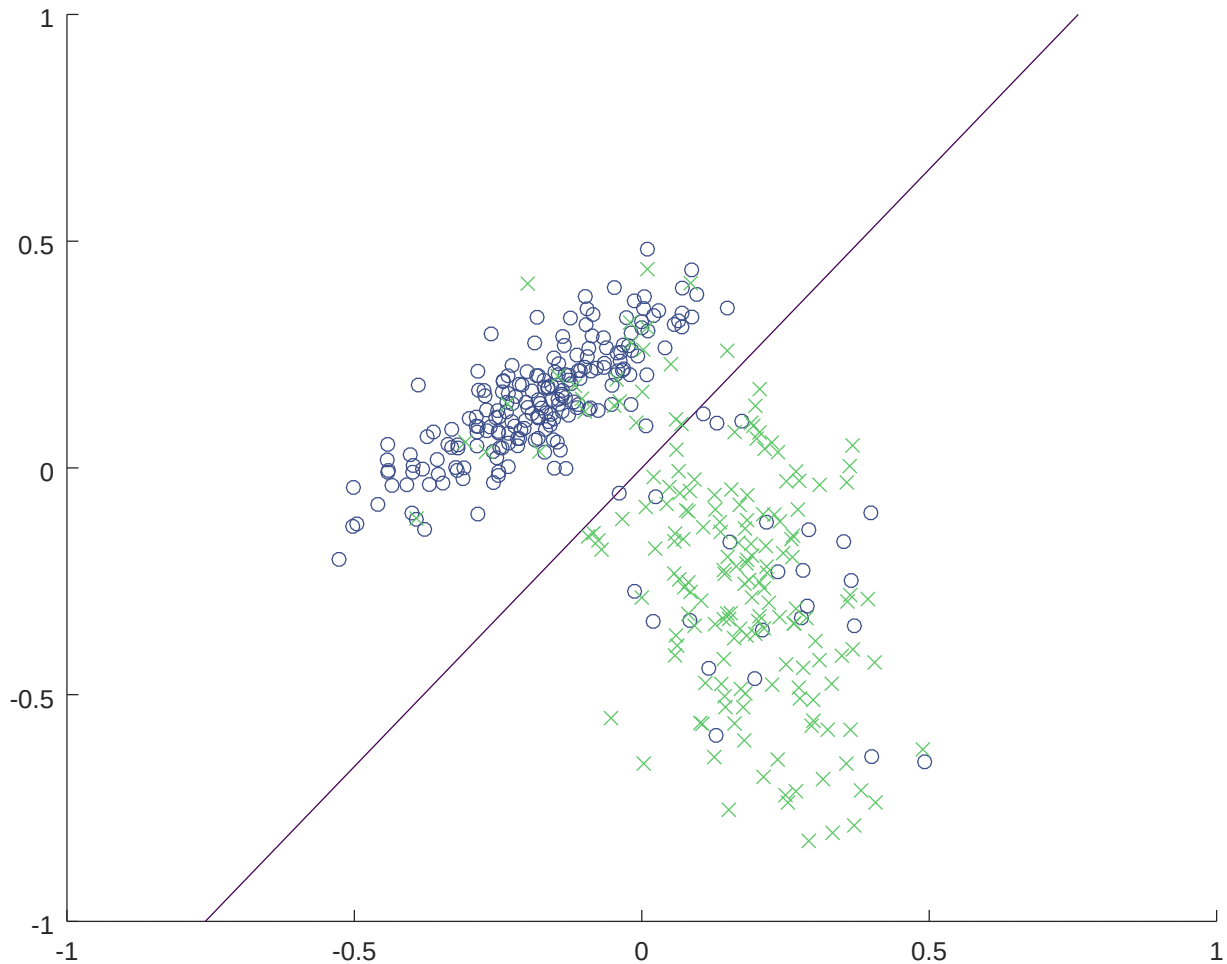
The "ordinary" least squares (OLS) solution:

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

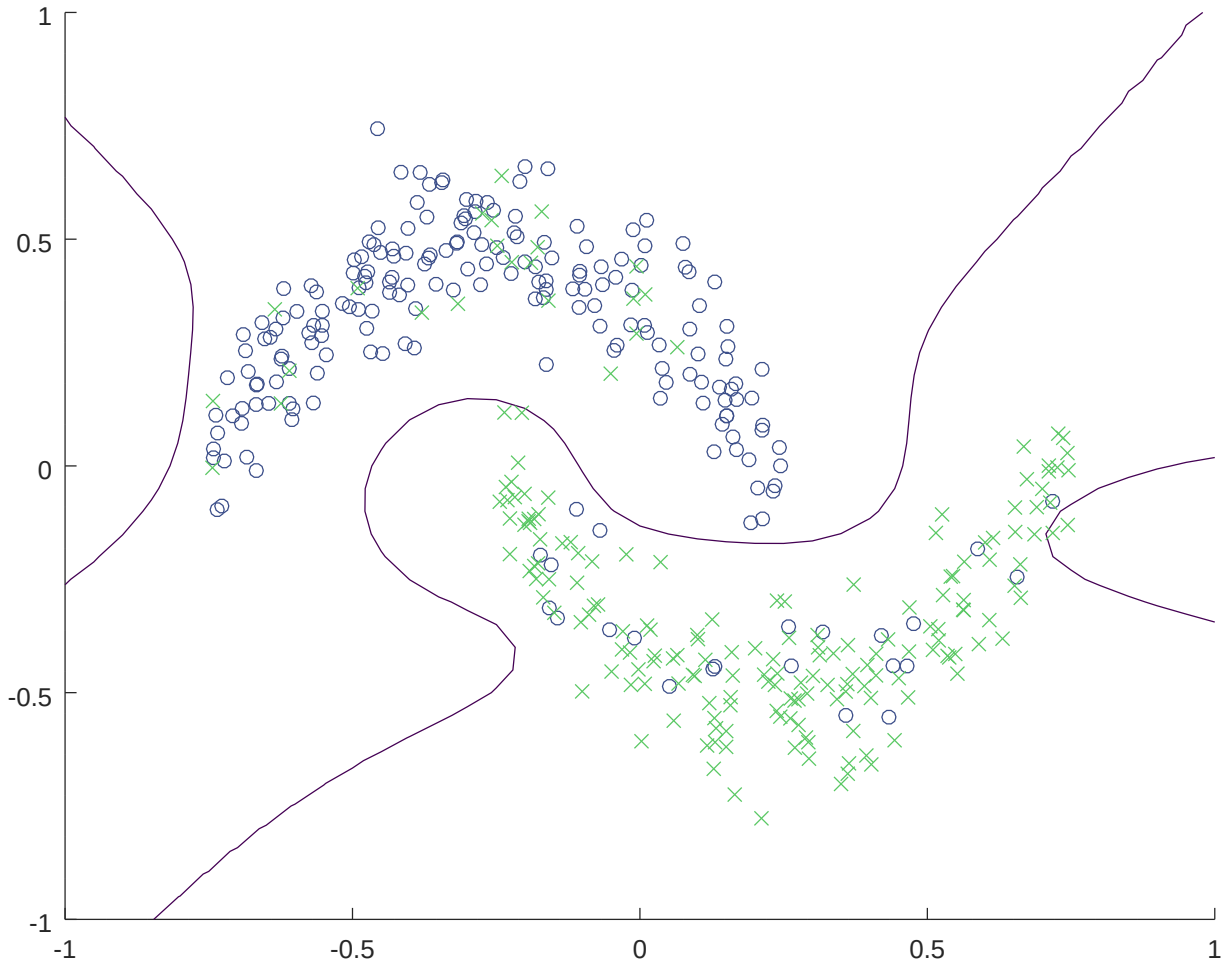
WHY LINEAR REGRESSION?

(c) Randall Munroe, `xkcd.com`

FROM LINEAR TO NONLINEAR



FROM LINEAR TO NONLINEAR



Extending the scope of linear regression

- Common strategy to model more complex functions:
  map input variable $x \in \mathcal{X}$ to a so-called "feature space" $\widetilde{\mathcal{X}}$ through $\widetilde{x} = \Phi(x) \in \widetilde{\mathcal{X}} = \mathbb{R}^D$.

- Typical examples (say with $\mathcal{X} = [0, 1]$):

$$\widetilde{x} = \Phi(x) = (1, x, x^2, \ldots, x^p) \in \mathbb{R}^{p+1};$$

$$\widetilde{x} = \Phi(x) = (1, \cos(2\pi x), \sin(2\pi x), \ldots, \cos(p\pi x), \sin(p\pi x)) \in \mathbb{R}^{2p+1}.$$

- More generally: feature space is a Hilbert space $\mathcal{H}$:
  - Functional Data Analysis: input x is already a function (e.g. idealized time series).
  - Reproducing Kernel methods: popular and versatile in machine learning.

Convergence of OLS

- We want to understand the behavior of $\widehat{\beta}_\lambda$, when the data size n grows large. Will we be close to the optimal prediction β_*?
- 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 $n \to \infty$:

$$\widehat{\Sigma} := \frac{1}{n} X^T X = \frac{1}{n} \sum_{i=1}^{n} \underbrace{X_i X_i^T}_{=:Z_i'} \longrightarrow \mathbb{E}\left[X_1 X_1^T\right] =: \Sigma;$$

$$\widehat{\gamma} := \frac{1}{n} X^T Y = \frac{1}{n} \sum_{i=1}^{n} \underbrace{X_i Y_i}_{=:Z_i} \longrightarrow \mathbb{E}[X_1 Y_1] = \Sigma\beta^* =: \gamma;$$

- Hence $\widehat{\beta} = \widehat{\Sigma}^{-1}\widehat{\gamma} \to \Sigma^{-1}\gamma = \beta^*$. (Assuming Σ invertible.)

From OLS to Hilbert-space regression

- For ordinary linear regression with $\mathcal{X} = \mathbb{R}^p$ (fixed p, $n \to \infty$):
  - LLN implies $\widehat{\beta}_{OLS} (= \widehat{\Sigma}^{-1}\widehat{\gamma}) \to \beta_* (= \Sigma^{-1}\gamma)$;
  - CLT+Delta Method imply asymptotic normality and convergence in $\mathcal{O}(n^{-\frac{1}{2}})$.

- **How to generalize to $\widetilde{\mathcal{X}} = \mathcal{H}$?**

- **Main issue:** $\Sigma = \mathbb{E}\left[XX^T\right]$ does not have a continuous inverse.
  ($\to$ ill-posed problem)

- **Roadmap:**
  1. Need to consider a suitable approximation $\zeta(\widehat{\Sigma})$ of Σ^{-1} **(regularization)**.
  2. Use a **nonasymptotic** version of vectorial LLN/CLT to control $\|\gamma - \widehat{\gamma}\|$ and $\left\|\Sigma - \widehat{\Sigma}\right\|$.
  3. Use (deterministic) **functional calculus** to get a handle on $\left\|\beta - \widehat{\beta}\right\|$ (reconstruction) or $\left\|\Sigma^{1/2}(\beta - \widehat{\beta})\right\|$ (prediction).

Vectorial Bernstein's inequality




- Result of **Pinelis and Shakanenko (1985):** if $Z_1, \ldots, Z_n$ are independent identically distributed vectors in a Euclidean or Hilbert space such that:
  - $\|Z_i\| \leq B$;
  - $\mathbb{E}\left[\|Z_i - \mathbb{E}[Z_i]\|^2\right] \leq \sigma^2$.
- Then it holds:

$$\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_i]\right\| \leq 2t\left(\frac{B}{n} + \frac{\sigma}{\sqrt{n}}\right),$$

  with probability larger than $1 - 2e^{-t}$.
- **Note**: works in **any dimension** p – even in a Hilbert space ($p = \infty$)!
- **Note**: also holds if $\|Z_i\|$ is unbounded but satisfies Bernstein-type moment conditions

Statistical error control

Error controls were introduced and used by Caponnetto and De Vito (2007), Caponnetto (2007), as a consequence of the Pinelis-Shakanenko inequality.

Theorem (Caponetto, De Vito)

Assume $\|X\| \leq 1$, $|Y| \leq M$ *and* $\mathrm{Var}[Y|X] \leq \sigma^2$ *a.s.*
Let $\lambda > 0$ *be fixed and define*

$$\mathcal{N}(\lambda) = \mathrm{Tr}\left((\Sigma + \lambda)^{-1}\Sigma \right),$$

then with probability at least $1 - 12e^{-t}$*:*

$$\left\| (\Sigma + \lambda)^{-\frac{1}{2}} (\widehat{\gamma} - \gamma) \right\| \leq 2t \left(\sigma \sqrt{\frac{\mathcal{N}(\lambda)}{n}} + \frac{2M}{\sqrt{\lambda} n} \right),$$

and

$$\left\| (\Sigma + \lambda)^{-\frac{1}{2}} (\widehat{\Sigma} - \Sigma) \right\|_{HS} \leq 2t \left(\sqrt{\frac{\mathcal{N}(\lambda)}{n}} + \frac{2}{\sqrt{\lambda} n} \right).$$

Effective dimension

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

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

resp.

$$\mathbf{IP}^{>}(s,\alpha'): \quad \mu_i \geq \alpha' i^{-\frac{1}{s}}.$$

- This implies quantitative estimates of the "effective dimension" entering in the concentration bound,

$$\mathcal{N}(\lambda) = \mathrm{Tr}(\ (\Sigma+\lambda)^{-1}\Sigma\ ) \underset{\gtrsim}{\overset{\lesssim}{\sim}} \lambda^{-s}$$

Regularization methods

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

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

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

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

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

$$\begin{aligned}\widehat{\beta}_{LW(k)} &= \widehat{\beta}_{LW(k-1)} + \frac{1}{n}\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}_{LW(k-1)}) \\ &= \sum_{i=0}^{k}(I - \widehat{\Sigma})^k\widehat{\gamma}\,,\end{aligned}$$

(assuming $\left\|\widehat{\Sigma}\right\|_{op} \leq 1$).

General form spectral linearization

- **General form** regularization method:

$$\widehat{\beta}_{Spec(\zeta,\lambda)} = \zeta_\lambda(\widehat{\Sigma})\widehat{\gamma}$$

  for somme 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 functional calculus, i.e.

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

- 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

From now on we assume $\kappa = 1$ for simplicity. 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

- Denote $(\mu_i)_{i \geq 1}$ the sequence of positive eigenvalues of Σ in nonincreasing order.

- **Source condition** for the signal: for $r > 0$, define

$$\mathbf{SC}(r, R): \quad \beta^* = \Sigma^r h_0 \text{ for some } h_0 \text{ with } \|h_0\| \leq R$$

or equivalently, as a Sobolev-type regularity

$$\mathbf{SC}(r, R): \quad \beta^* \in \left\{ \beta \in \mathcal{H} : \sum_{i \geq 1} \mu_i^{-2r} \beta_i^2 \leq R^2 \right\},$$

where β_i are the coefficients of h in the eigenbasis of Σ.

Convergence analysis

- Recall linear model $\boldsymbol{Y} = \boldsymbol{X}\beta_* + \boldsymbol{\xi}$, regularized estimator $\widehat{\beta}_\lambda = \zeta_\lambda(\widehat{\Sigma})\boldsymbol{X}^T\boldsymbol{Y}/n$.
- Induces decomposition

$$\widehat{\beta}_\lambda - \beta_* = \underbrace{(\zeta_\lambda(\widehat{\Sigma})\widehat{\Sigma} - \boldsymbol{I})\beta_*}_{\text{Approximation term}} + \underbrace{\zeta_\lambda(\widehat{\Sigma})\boldsymbol{X}^T\boldsymbol{\xi}/n}_{\text{Noise term}}$$

Convergence analysis

- Recall linear model $Y = X\beta_* + \xi$, regularized estimator $\widehat{\beta}_\lambda = \zeta_\lambda(\widehat{\Sigma})X^T Y/n$.
- Induces decomposition

$$\widehat{\beta}_\lambda - \beta_* = \underbrace{(\zeta_\lambda(\widehat{\Sigma})\widehat{\Sigma} - I)\beta_*}_{\text{Approximation term}} + \underbrace{\zeta_\lambda(\widehat{\Sigma})X^T\xi/n}_{\text{Noise term}}$$

- Noise Term: has zero expectation, and

$$\|\zeta_\lambda(\widehat{\Sigma})X^t\xi/n\| \leq \|\zeta_\lambda(\widehat{\Sigma})\|_{op}\|X^t\xi/n\| \leq \lambda^{-1}\left\|\frac{1}{n}\sum_{i=1}^n X_i\xi_i\right\|.$$

- Approximation Term (using **source condition**):

$$(\zeta_\lambda(\widehat{\Sigma})\widehat{\Sigma} - I)\beta_* = r_\lambda(\widehat{\Sigma})\Sigma^r h_0.$$

- If we can "replace" Σ^r by $\widehat{\Sigma}^r$ above (using concentration+operator perturbation inequalities), we can use **qualification** assumption to bound

$$\left\|r_\lambda(\widehat{\Sigma})\widehat{\Sigma}^r h_0\right\| \lesssim \lambda^r R.$$

Upper bound on rates

Theorem

Assume r, R, s, α *are fixed positive constants and assume* $\mathbb{P}_{XY}$ *satisfies* **(IP$^+$)**(s, α), **(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 (ζ_λ) *satisfying the standard assumptions with qualification* $q \geq r + \frac{1}{2}$, *and the parameter choice rule*

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

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

$$\limsup_{n\to\infty} \mathbb{E}^{\otimes n}\left(\left\|\beta_* - \widehat{\beta}_n\right\|^p\right)^{1/p} \Big/ R\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{r}{2r+1+s}} \leq C_{\blacktriangle};$$

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

Comments

- It follows that the convergence rate obtained is of order

$$C_{\blacktriangle} R \left(\frac{\sigma^2}{R^2 n} \right)^{\frac{(r+\theta)}{2r+1+s}}$$

  (with $\theta = 0$ resp. $1/2$ for reconstruction resp. prediction risk).

- The "constant" $C_{\blacktriangle}$ depends on the various parameters entering in the assumptions, but **not** on n, R, σ, M!

- The result applies to all linear spectral regularization methods but assuming a precise tuning of the regularization constant λ as a function of the assumed regularization parameters of the target – **not adaptive.**

"Weak" lower bound on rates

Theorem

Assume r, R, s, β are fixed positive constants and let $\mathcal{P}'(r, R, s, \beta)$ denote the set of distributions on $\mathcal{X} \times \mathcal{Y}$ satisfying **(IP$^-$)**(s, β), **(SC)**(r, R) *and Bernstein moments conditions for the noise. (We assume this set to be non empty!) Then*

$$\limsup_{n\to\infty} \inf_{\widehat{h}} \sup_{P\in\mathcal{P}'(r,R,s,\beta)} P^{\otimes n}\left(\left\| S^{\theta}(h^* - \widehat{h}) \right\|_{\mathcal{H}_K} > CR\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{(r+\theta)}{2r+1+s}} \right) > 0$$

Proof: Fano's lemma technique

"STRONG" LOWER BOUND ON RATES

Assume additionally "no big jumps in eigenvalues":

$$\inf_{k \geq 1} \frac{\mu_{2k}}{\mu_k} > 0$$

Theorem

Assume r, R, s, β are fixed positive constants and let $\mathcal{P}'(r, R, s, \beta)$ denote the set of distributions on $\mathcal{X} \times \mathcal{Y}$ satisfying **(IP$^-$)**(s, β), **(SC)**(r, R) *and Bernstein moment conditions for the noise. (We assume this set to be non empty!) Then*

$$\liminf_{n \to \infty} \inf_{\widehat{h}} \sup_{P \in \mathcal{P}'(r,R,s,\beta)} P^{\otimes n} \left(\left\| S^\theta (h^* - \widehat{h}) \right\|_{\mathcal{H}_K} > CR \left(\frac{\sigma^2}{R^2 n} \right)^{\frac{(r+\theta)}{2r+1+s}} \right) > 0$$

Proof: Fano's lemma technique

Previous results

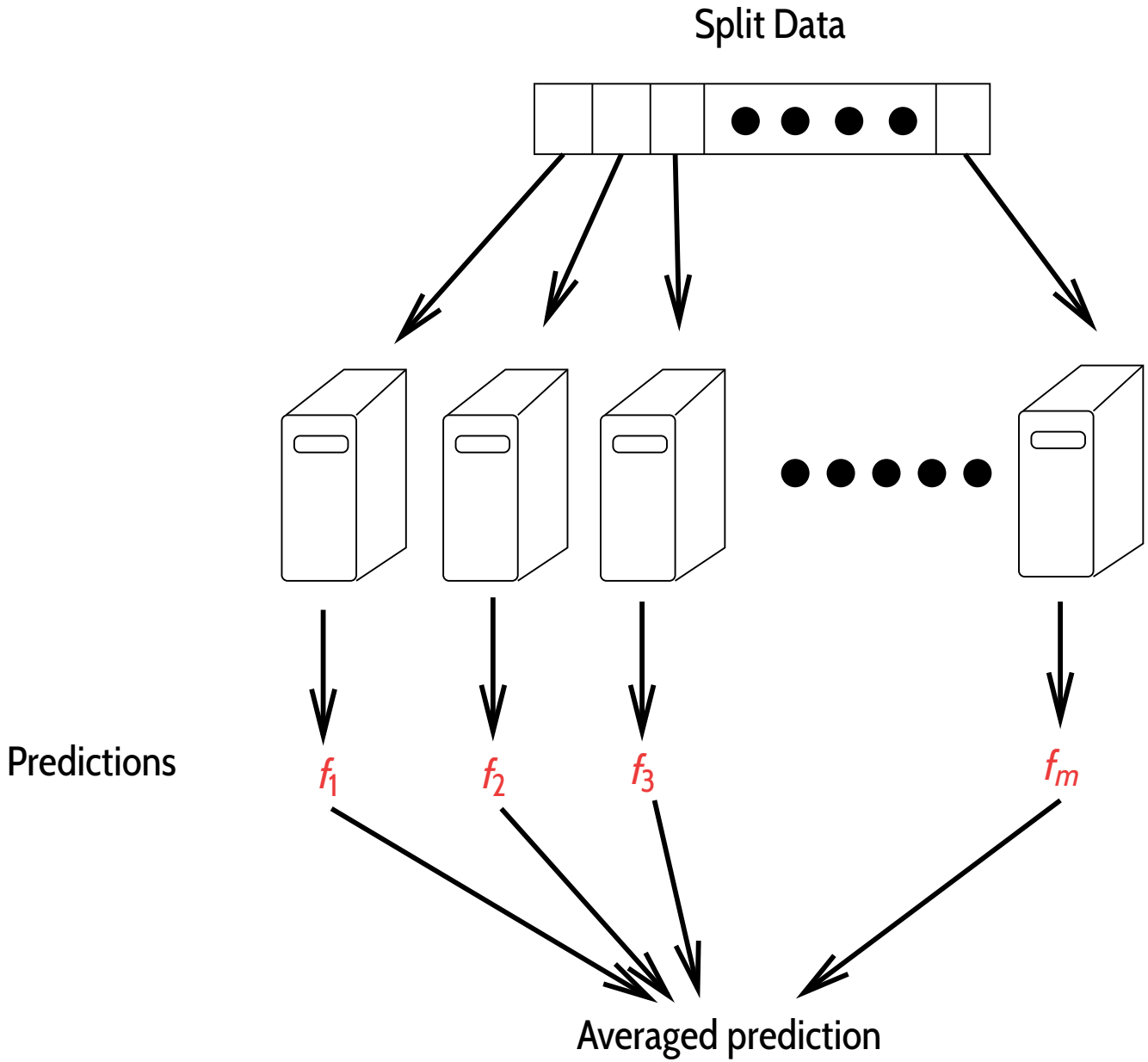
Error	[1]	[2]	[3]	[4]
$\left\| f_{\widehat{\beta}} - f_{\beta_*} \right\|_{2,X}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{2r+1}{2r+2}}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{2r+1}{2r+2}}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{(2r+1)}{2r+1+s}}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{(2r+1)}{2r+1+s}}$
$\left\| \widehat{\beta} - \beta^* \right\|$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{r}{r+1}}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{r}{r+1}}$	N/A	N/A
Assumptions (q: qualification)	$r \leq \frac{1}{2}$	$r \leq q - \frac{1}{2}$	$r \leq \frac{1}{2}$	$0 \leq r \leq q - \frac{1}{2}$ +unlabeled data if $2r + s < 1$
Method	Tikhonov	General	Tikhonov	General

[1]: Smale and Zhou (2007)
[2]: Bauer, Pereverzev, Rosasco (2007)
[3]: Caponnetto, De Vito (2007)
[4]: Caponnetto and Yao (2010)

Matching lower bound: only for $\left\| f_{\widehat{\beta}} - f_{\beta_*} \right\|_{2,X}$ [2].

Gaining computational efficiency

The divide-and-average paradigm



Gaining computational efficiency

The divide-and-average paradigm

- **Divide and average:**
  - Divide sample $(X_i, Y_i)_{1 \leq i \leq N}$ into m equal-size subsamples
  - Apply learning method $\widehat{\beta}_\lambda$ on each subsample (this can be distributed over m independent machines)
  - Take the simple average of the obtained estimators

- Use the **same** regularization parameter λ_n as the optimal one without parallelization

- Rough intuition:
  - The "bias" (approximation error) using a subsample should be of the same order as when using the whole sample
  - The "variance" (estimation error) is higher on a subsample, but gets reduced by averaging

Divide-and-average analysis

- Suppose the computational complexity is of order n^3.
- If we can distribute the load over m parallel computers each treating a sample of size n/m, the overall complexity will be of order

$$m.(n/m)^3 = n^3/m^2,$$

  a gain of a factor m^2!
- Theory question: how can we choose m as big as possible such that **statistical optimality** (for convergence rates) is preserved?
- Answer is obtained again by using vectorial concentration tools (separately on each machine, then for the final averaging step, which is also an i.i.d. average!)

Divide-and-average result

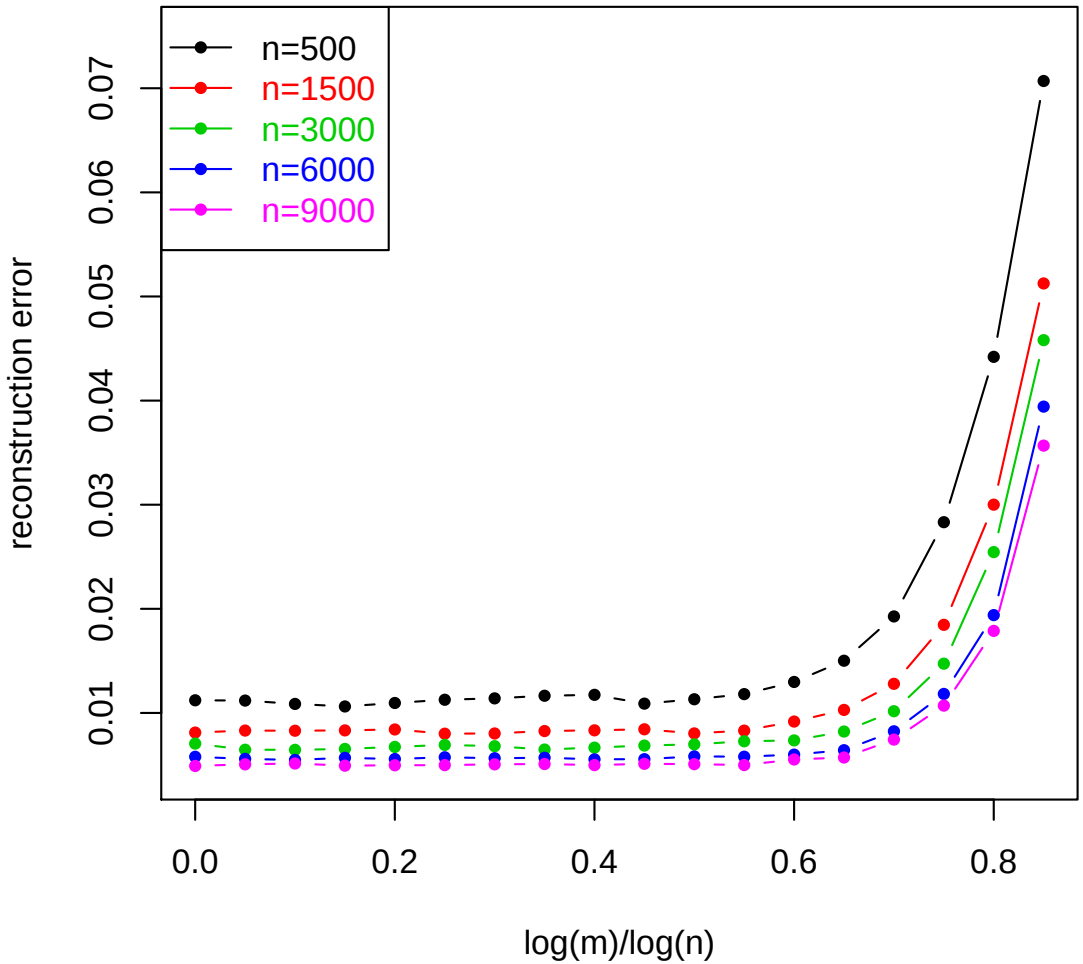
Theorem

Under the same assumptions as in the previous theorem, using divide-and-average over $m = n^{\alpha}$ machines and with the same choice of regularization parameter λ_n as before results in the same asymptotic bound (in all p-norms) on the convergence rate as for a single machine, provided

$$\alpha \leq \frac{2\min(r,1)}{2r+1+s}$$

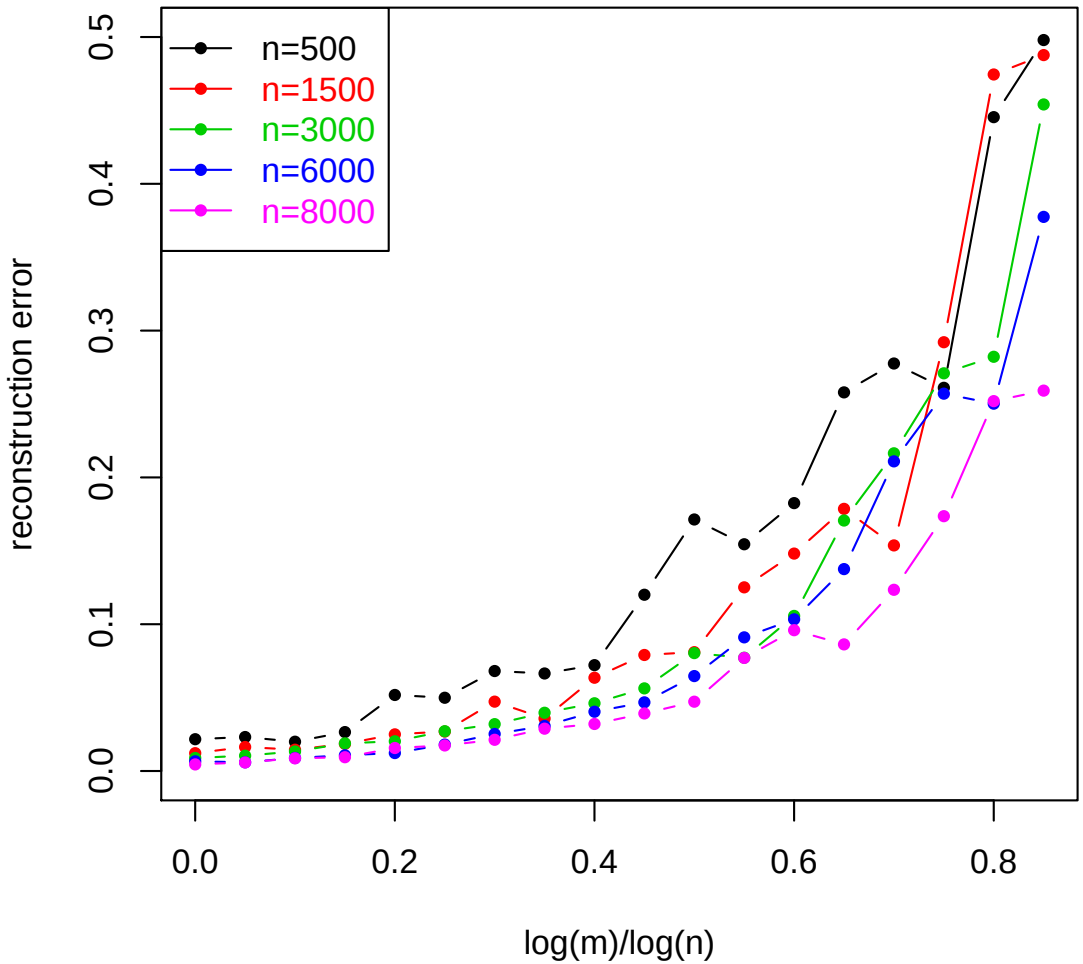
- Approximation term: has nonzero expectation. No help from averaging, need to be small for all machines. For this choose regularization parameter λ_n as in single-machine case.
  - The "replace Σ^r by $\widehat{\Sigma}^r$" step is the bottleneck giving rise to the limitation on α.
- Noise term: has zero expectation. Averaging over independent subsamples reduces variability!
  - Control moments via the single machine analysis. For moments of average use **vector-valued Rosenthal's inequality** due to Pinelis.

SIMULATION: ROUGH SIGNAL



$s = \frac{1}{2}, r = \frac{3}{4}$, theoretical sufficient : $\alpha \leq \frac{1}{2}$.

SIMULATION: SUPERSMOOTH SIGNAL



$s = \frac{1}{2}$, $r = \infty$, theory: no parallelization optimality guarantee

Non-i.i.d. data

- The convergence analysis is decomposed into:
  - a **probabilistic** part: vectorial concentration inequality in Hilbert space
  - a **deterministic** part: under the event of large probability where deviations are controlled, use (deterministic) operator perturbation tools to get estimates

- **What to do if the data is not i.i.d.**? If we can find a replacement for Pinelis and Shakanenko's vectorial Bernstein inequality, we can follow through with the rest of the analysis.

Vectorial Bernstein's under weak dependence

- if $Z_1, \ldots, Z_n$ are independent identically distributed random vectors such that:
  - $\|Z_i\| \leq B$;
  - $\mathbb{E}\left[\|Z_i - \mathbb{E}[Z_i]\|^2\right] \leq \sigma^2$.

- Then it holds:

$$\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i - \mathbb{E}[Z_i] \right\| \leq 2t \left(\frac{B}{n} + \frac{\sigma}{\sqrt{n}} \right),$$

with probability larger than $1 - 2e^{-t}$.

Vectorial Bernstein's under weak dependence

- if $Z_1, \dots, Z_n$ are *independent identically distributed* random vectors such that:
  - $\|Z_i\| \leq B$;
  - $\mathbb{E}\left[\|Z_i - \mathbb{E}[Z_i]\|^2\right] \leq \sigma^2$.
- Weak dependence assumption:

$$\Phi(k) := \sup\left\{\left\|E[\varphi(Z_{i+k})|(Z_j)_{j\leq i}] - E[\varphi(Z_{i+k})]\right\|_\infty \mid \varphi \in \mathcal{C}, i \geq 1\right\},$$

where $\mathcal{C} := \left\{x \mapsto \|x\|^2; x \mapsto \langle w, x\rangle, \|w\| \leq 1\right\}$.

See: Maume-Deschamps (2006), Dedecker et al. (2007), Dedecker and Merlevede (2015)

- Then it holds:

$$\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z_i]\right\| \leq 2Ct\left(\frac{B}{\ell_n^*} + \frac{\sigma}{\sqrt{\ell_n^*}}\right),$$

with probability larger than $1 - 2e^{-t}$, for ℓ_n^* satisfying $\Phi\left(\left\lfloor \frac{n}{\ell} \right\rfloor\right) \leq \frac{B}{\ell} \vee \frac{\sigma}{\sqrt{\ell}}$.

Conclusion/Perspectives

- We filled gaps in the existing picture for linear learning methods in Hilbert space.
- The method (and convergence analysis) lend themselves well to parallelization.
- Extension to weakly dependent data.
- **Concentration + operator perturbation methods offer a nice and robust set of mathematical tools to analyze convergence.**
- Adaptivity: ideally attain optimal rates without a priori knowledge of r **nor** of s!
  - Lepski's method/balancing principle: in progress. Need a good estimator for $\mathcal{N}(\lambda)$! (Prior work on this: Caponnetto; need some sharper bound)

G. Blanchard, N. Mücke.

Optimal Rates For Regularization Of Statistical Inverse Learning Problems.

Foundations of Computational Mathematics, 2017.

G. Blanchard, N. Krämer.

Convergence rates of Kernel Conjugate Gradient for random design regression.

Analysis and Applications 14 (6): 763-794, 2016.

G. Blanchard, N. Mücke.

Parallelizing Spectral Algorithms for Kernel Learning.

arXiv:1610.07487 (hopefully soon in J. Mach. Learn. Res.)

G. Blanchard, O. Zadorozhnyi.

Concentration of weakly dependent Banach-valued sums and applications to kernel learning methods.

arXiv:1712.01934

G. Blanchard, N. Mücke.

Kernel regression, minimax rates and effective dimensionality: beyond the regular case.

arXiv:1611.03979