Spectral regularization methods for statistical inverse learning problems

G. Blanchard

Universtität Potsdam

van Dantzig seminar, 23/06/2016

Joint work with N. Mücke (U. Potsdam)

1. General regularization and kernel methods
2. Inverse learning/regression and relation to kernels
3. Rates for linear spectral regularization methods
4. Beyond the regular spectrum case

INTRODUCTION: RANDOM DESIGN REGRESSION

- Consider the familiar regression setting on a random design,

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

  where $(X_i, Y_i)_{1 \leq i \leq n}$ is an i.i.d. sample from P_{XY} on the space $\mathcal{X} \times \mathbb{R}$,
- with $\mathbb{E}\left[\varepsilon_i | X_i\right] = 0$.
- For an estimator $\widehat{f}$ we consider the **prediction error** function,

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

  which we want to be as small as possible (in expectation or with large probability).
- We can also be interested in squared **reconstruction error**

$$\left\| \widehat{f} - f^* \right\|_{\mathcal{H}}^2$$

  where $\mathcal{H}$ is a certain Hilbert norm of interest for the user.

LINEAR CASE

- Very classical is the linear case: $\mathcal{X} = \mathbb{R}^p$, $f^*(x) = \langle x, \beta^* \rangle$, and in usual matrix form (X_i^t form the lines of the design matrix $\mathbf{X}$)

$$\mathbf{Y} = \mathbf{X}\beta^* + \varepsilon$$

- ordinary least squares solution is

$$\widehat{\beta}_{OLS} = (\mathbf{X}^t\mathbf{X})^\dagger \mathbf{X}^t\mathbf{Y}.$$

- Prediction error corresponds to $\mathbb{E}\left[\left\langle \beta^* - \widehat{\beta}, X \right\rangle^2\right]$
- Reconstruction error corresponds to $\left\|\beta^* - \widehat{\beta}\right\|^2$.

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"** through $\widetilde{x} = \Phi(x)$
- typical examples (say with $\mathcal{X} = [0,1]$) are

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

$$\widetilde{x} = \Phi(x) = (1, \cos(2\pi x), \sin(2\pi x), \cos(3\pi x), \sin(3\pi x), \ldots) \in \mathbb{R}^{2p+1}.$$

- **Problem:** large number of parameters to estimate require regularization to avoid overfitting.

REGULARIZATION METHODS

- Main idea of regularization is to replace $(\mathbf{X}^t\mathbf{X})^\dagger$ by an approximate inverse, for instance
- **Ridge regression/Tikhonov**:

$$\widehat{\beta}_{Ridge(\lambda)} = (\mathbf{X}^t\mathbf{X} + \lambda I_p)^{-1}\mathbf{X}^t\mathbf{Y}$$

- **PCA projection/spectral cut-off**: restrict $\mathbf{X}^t\mathbf{X}$ on its k first eigenvectors

$$\widehat{\beta}_{PCA(k)} = (\mathbf{X}^t\mathbf{X})^\dagger_{|k}\mathbf{X}^t\mathbf{Y}$$

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

$$\begin{aligned}\widehat{\beta}_{LW(k)} &= \widehat{\beta}_{LW(k-1)} + \mathbf{X}^t(\mathbf{Y} - \mathbf{X}\widehat{\beta}_{LW(k-1)}) \\ &= \sum_{i=0}^{k}(I - \mathbf{X}^t\mathbf{X})^k\mathbf{X}^t\mathbf{Y}\,,\end{aligned}$$

(assuming $\left\|\mathbf{X}^t\mathbf{X}\right\|_{op} \leq 1$).

General form spectral linearization

- **General form** regularization method:

$$\widehat{\beta}_{Spec(\zeta,\lambda)} = \zeta_\lambda(\mathbf{X}^t\mathbf{X})\mathbf{X}^t\mathbf{Y}$$

  for somme well-chosen function $\zeta_\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ acting on the spectrum and "approximating" the function $x \mapsto 1/x$.
- $\lambda > 0$: regularization parameter; $\lambda \to 0 \Leftrightarrow$ less regularization
- Notation of functional calculus, i.e.

$$\mathbf{X}^t\mathbf{X} = Q^T \mathrm{diag}(\lambda_1, \ldots, \lambda_p) Q \to \zeta(\mathbf{X}^t\mathbf{X}) := Q^T \mathrm{diag}(\zeta(\lambda_1), \ldots, \zeta(\lambda_p)) Q$$

- Many well-known from the inverse problem literature
- Examples:
  - **Tikhonov**: $\zeta_\lambda(t) = (t+\lambda)^{-1}$
  - **Spectral cut-off**: $\zeta_\lambda(t) = t^{-1}\mathbf{1}\{t \geq \lambda\}$
  - **Landweber iteration**: $\zeta_k(t) = \sum_{i=0}^{k}(1-t)^i$.

COEFFICIENT EXPANSION

- A useful trick of functional calculus is the "**shift rule**":

$$\zeta(\mathbf{X}^t\mathbf{X})\mathbf{X}^t = \mathbf{X}^t\zeta(\mathbf{X}\mathbf{X}^t)\,.$$

- **Interpretation**:

$$\widehat{\beta}_{Spec(\zeta,\lambda)} = \zeta(\mathbf{X}^t\mathbf{X})\mathbf{X}^t\mathbf{Y} = \mathbf{X}^t\zeta(\mathbf{X}\mathbf{X}^t)\mathbf{Y} = \sum_{i=1}^{n}\widehat{\alpha}_i X_i\,,$$

  with

$$\widehat{\alpha}_i = \zeta(G)\mathbf{Y}\,,$$

  and $G = \mathbf{X}\mathbf{X}^t$ is the (n,n) Gram matrix of $(X_1,\ldots,X_n)$.
- This representation is more economical if $p \gg n$.

THE "KERNELIZATION" ANSATZ

- Let Φ be a feature mapping into a (possibly infinite dimensional) Hilbert feature space $\mathcal{H}$.
- Representing $\widetilde{x} = \Phi(x) \in \mathcal{H}$ explicitly is cumbersome/impossible in practice, but if we can compute quickly the **kernel**

$$K(x, x') := \left\langle \widetilde{x}, \widetilde{x'} \right\rangle = \langle \Phi(x), \Phi(x') \rangle \,,$$

  then **kernel Gram matrix** $\widetilde{G}_{ij} = \left\langle \widetilde{x}_i, \widetilde{x}_j \right\rangle = K(x_i, x_j)$ is accessible.
- We can hence directly "kernelize" any classical regularization technique using the implicit representation

$$\widehat{\beta}_{Spec(\zeta,\lambda)} = \sum_{i=1}^{n} \widehat{\alpha}_i \widetilde{X}_i \,, \qquad \widehat{\alpha}_i = \zeta(\widetilde{G})\mathbf{Y} \,,$$

- the value of $f(x) = \left\langle \widehat{\beta}, \widetilde{x} \right\rangle$ can then be computed for any x:

$$f(x) = \sum_{i=1}^{n} \widehat{\alpha}_i K(X_i, x) \,.$$

REPRODUCING KERNEL METHODS

- If $\mathcal{H}$ is a Hilbert feature space, it is useful to identify it as a **space of real functions** on $\mathcal{X}$ of the form $f(x) = \langle w, \Phi(x) \rangle$. The canonical feature mapping is then $\Phi(x) = K(x, .)$ and the "reproducing kernel" property reads

$$f(x) = \langle f, \Phi(x) \rangle = \langle f, K(x, .) \rangle \ .$$

- Classical kernels on $\mathbb{R}^d$ include
  - Gaussian Kernel $K(x, y) = \exp - \|x - y\|^2 / 2\sigma^2$
  - Polynomial Kernel $K(x, y) = (1 + \langle x, y \rangle)^p$
  - Spline kernels, Matérn kernel, inverse quadratic kernel...
- Success of reproducing kernel methods since early 00's is due to their **versatility** and **ease of use**: beyond vector spaces, kernels have been constructed on various non-euclidean data (text, genome, graphs, probability distributions...)
- One of the tenets of "learning theory" is a **distribution-free point of view**; in particular the sampling distribution (of the X_is) is unknown to the user and could be very general.

SETTING: "INVERSE LEARNING" PROBLEM

- We refer to "inverse learning" (or inverse regression) for an inverse problem where we have **noisy** observations at **random design points**:

$$(X_i, Y_i)_{i=1,\dots,n} \text{ i.i.d. } : \qquad Y_i = (Af^*)(X_i) + \varepsilon_i \,. \qquad \text{(ILP)}$$

- the goal is to recover $f^* \in \mathcal{H}_1$.
- early works on closely related subjects: from the splines literature in the 80's (e.g. O'Sullivan '90)

Main assumption for Inverse Learning

Model: $Y_i = (Af^*)(X_i) + \varepsilon_i\,, i = 1, \ldots, n,$ where $A : \mathcal{H}_1 \to \mathcal{H}_2$. (ILP)

Observe:

- $\mathcal{H}_2$ should be a space of real-values functions on $\mathcal{X}$.
- the geometrical structure of the "measurement errors" will be dictated by the statistical properties of the sampling scheme – no need to assume or consider any a priori Hilbert structure on $\mathcal{H}_2$
- crucial stuctural assumption is the following:

Assumption

The family of evaluation functionals (S_x), $x \in \mathcal{X}$, defined by

$$\begin{aligned} S_x : \mathcal{H}_1 &\longrightarrow \mathbb{R} \\ f &\longmapsto (S_x)(f) := (Af)(x) \end{aligned}$$

is uniformly bounded, i.e., there exists $\kappa < \infty$ such that for any $x \in \mathcal{X}$

$$|S_x(f)| \leq \kappa \|f\|_{\mathcal{H}_1} .$$

GEOMETRY OF INVERSE LEARNING

- The inverse learning under the previous assumption was essentially considered by Caponnetto et al. (2006).
- Riesz's theorem implies the existence for any $x \in \mathcal{X}$ of $F_x \in \mathcal{H}_1$:

$$\forall f \in \mathcal{H}_1 \ : \qquad (Af)(x) = \langle f, F_x \rangle$$

- $K(x,y) := \langle F_x, F_y \rangle$ defines a positive semidefinite kernel on $\mathcal{X}$ with associated reproducing kernel Hilbert space (RKHS) denoted $\mathcal{H}_K$.
- as a pure function space, $\mathcal{H}_K$ coincides with $Im(A)$.
- assuming A injective, A is in fact an **isometric isomorphism** between $\mathcal{H}_1$ and $\mathcal{H}_K$.

GEOMETRY OF INVERSE LEARNING

- Main assumption implies that as a function space, $Im(A)$ is endowed with a natural RKHS structure with a kernel K bounded by κ.
- Furthermore this RKHS $\mathcal{H}_K$ is isometric to $\mathcal{H}_1$ (through A^{-1}).
- Therefore, the inverse learning problem is formally equivalent to the kernel learning problem

$$Y_i = h^*(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n$$

where $h^* \in \mathcal{H}_K$, and we measure the quality of an estimator $\widehat{h} \in \mathcal{H}_K$ via the RKHS norm $\left\| \widehat{h} - h^* \right\|_{\mathcal{H}_K}$

- Indeed, if we put $\widehat{f} := A^{-1}\widehat{h}$, then

$$\left\| \widehat{f} - f^* \right\|_{\mathcal{H}_1} = \left\| A(\widehat{f} - f^*) \right\|_{\mathcal{H}_K} = \left\| \widehat{h} - h^* \right\|_{\mathcal{H}_K}$$

SETTING, REFORMULATED

- We are actually back to the familiar regression setting on a random design,

$$Y_i = h^*(X_i) + \varepsilon_i \,,$$

  where $(X_i, Y_i)_{1 \leq i \leq n}$ is an i.i.d. sample from $\mathbb{P}_{XY}$ on the space $\mathcal{X} \times \mathbb{R}$,
- with $\mathbb{E}\left[\varepsilon_i | X_i\right] = 0$.
- Noise assumptions:

$$\textbf{(BernsteinNoise)} \qquad \mathbb{E}\left[\varepsilon_i^p | X_i\right] \leq \frac{1}{2} p! M^p, \quad p \geq 2$$

- h^* is assumed to lie in a (known) RKHS $\mathcal{H}_K$ with bounded kernel K.
- The criterion for measuring the quality of an estimator $\widehat{h}$ is the **RKHS norm**

$$\left\| \widehat{h} - h^* \right\|_{\mathcal{H}_K} .$$

EMPIRICAL AND POPULATION OPERATORS

- Define the (random) **empirical evaluation operator**

$$T_n : h \in \mathcal{H} \mapsto (h(X_1), \ldots, h(X_n)) \in \mathbb{R}^n \qquad \textbf{(analogue of } \widetilde{\mathbf{X}}\textbf{)}$$

  and its population counterpart the inclusion operator

$$T : h \in \mathcal{H} \mapsto h \in L_2(\mathcal{X}, \mathbb{P}_X);$$

- the (random) **empirical kernel integral operator**

$$T_n^* : (v_1, \ldots, v_n) \in \mathbb{R}^n \mapsto \frac{1}{n} \sum_{i=1}^{n} K(X_i, .) v_i \in \mathcal{H} \qquad \textbf{(analogue of } \widetilde{\mathbf{X}}^t/n\textbf{)}$$

  and its population counterpart, the **kernel integral operator**

$$T^* : f \in L_2(\mathcal{X}, \mathbb{P}_X) \mapsto T^*(f) = \int f(x) k(x, .) d\mathbb{P}_X(x) \in \mathcal{H}.$$

- finally, define the empirical covariance operator $S_n = T_n^* T_n$ (analogue of $\frac{1}{n}\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}$) and its population counterpart $S = T^* T$ (**analogue of** $\mathbb{E}\left[\frac{1}{n}\widetilde{\mathbf{X}}^t\widetilde{\mathbf{X}}\right] = \mathbb{E}\left[XX^t\right]$, uncentered covariance)
- Main intuition: S_n is a (random) approximation of S.

SPECTRAL REGULARIZATION IN KERNEL SPACE

- Linear spectral regularization in kernel space is written

$$\widehat{h}_\zeta = \zeta(S_n) T_n^* \mathbf{Y}$$

- recall

$$\zeta(S_n) T_n^* = \zeta(T_n^* T_n) T_n^* = T_n^* \zeta(T_n T_n^*) = T_n^* \zeta(K_n)\,,$$

where $K_n = T_n T_n^* : \mathbb{R}^n \to \mathbb{R}^n$ is the (normalized) **kernel Gram matrix**,

$$K_n(i,j) = \frac{1}{n} K(X_i, X_j)\,.$$

- equivalently:

$$\widehat{h}_\zeta = \sum_{i=1}^n \widehat{\alpha}_{\zeta,i} K(X_i, .)$$

with

$$\widehat{\alpha}_\zeta = \frac{1}{n} \zeta\left(K_n\right) \mathbf{Y}.$$

Structural Assumptions

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

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

or equivalently seen as a Sobolev-type regularity set

$$\mathbf{SC}(r, R): h^* \in \left\{ h \in \mathcal{H} : \sum_{i\geq 1} \lambda_i^{-2r} h_i^2 \leq R^2 \right\},$$

where h_i are the coefficients of h in the eigenbasis of S.

- **Ill-posedness:**

$$\mathbf{IP}^+(s, \beta): \quad \lambda_i \leq \beta i^{-\frac{1}{s}}$$

and

$$\mathbf{IP}^-(s, \beta'): \quad \lambda_i \geq \beta' i^{-\frac{1}{s}}$$

ERROR/RISK MEASURE

- We are measuring the error (risk) of an estimator $\widehat{h}$ in the family of norms

$$\left\| S^{\theta}(\widehat{h} - h^*) \right\|_{\mathcal{H}_K} \qquad (\theta \in [0, \frac{1}{2}])$$

- Note $\theta = 0$: reconstruction error in $\mathcal{H}_1$; $\theta = 1/2$: prediction error, since

$$\left\| S^{\frac{1}{2}}(\widehat{h} - h^*) \right\|_{\mathcal{H}_K} = \left\| \widehat{h} - h^* \right\|_{L^2(\mathbb{P}_X)} .$$

PREVIOUS RESULTS

Error	[1]	[2]	[3]	[4]
$\left\|\widehat{h} - h^*\right\|_{L^2(\mathbb{P}_X)}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{2r+1}{2r+2}}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{2r+1}{2r+2}}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{(2r+1)}{2r+1+s}}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{(2r+1)}{2r+1+s}}$
$\left\|\widehat{h} - h^*\right\|_{\mathcal{H}_K}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{r}{r+1}}$	$\left(\frac{1}{\sqrt{n}}\right)^{\frac{r}{r+1}}$	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 $\left\|\widehat{h} - h^*\right\|_{L^2(\mathbb{P}_X)}$ [2].

Compare to results known for regularization methods under White Noise model: Mair and Ruymgaart (1996), Nussbaum and Pereverzev (1999), Bissantz, Hohage, Munk and Ruymgaart (2007).
See also: recent preprint of Dicker, Foster, Hsu (2016)

ASSUMPTIONS ON REGULARIZATION FUNCTION

From now on we assume $\kappa = 1$ for simplicity. Standard assmptions 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:*

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

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

Upper 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* **(BernsteinNoise)**. *Define*

$$\widehat{h}_{\lambda_n}^{(n)} = \zeta_{\lambda_n}(S_n) T_n^* \mathbf{Y}$$

using a regularization family (ζ_λ) *satisfying the standard assumptions with qualification* $q \geq r + \theta$, *and the parameter choice rule*

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

it holds for any $\theta \in [0, \frac{1}{2}], \eta \in (0, 1), p \geq 1$*:*

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}(r,R,s,\beta)} \mathbb{E}^{\otimes n} \left(\left\| S^\theta (h^* - \widehat{h}_{\lambda_n}^{(n)}) \right\|_{\mathcal{H}_K}^p \right)^{\frac{1}{p}} \Big/ R \left(\frac{\sigma^2}{R^2 n}\right)^{\frac{(r+\theta)}{2r+1+s}} \leq C.$$

COMMENTS

- it follows that the convergence rate obtained is of order

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

- the "constant" C 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* **(BernsteinNoise)**. *(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":

$$\inf_{k\geq 1} \frac{\lambda_{2k}}{\lambda_k} > 0$$

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* **(BernsteinNoise)**. *(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

COMMENTS

- obtained rates are minimax (but not adaptive) in the parameters $R, n, \sigma \ldots$
- ... provided $\mathbf{(IP^-)}(s,\beta) \cap \mathbf{(IP^+)}(s,\alpha)$ is not empty.

STATISTICAL ERROR CONTROL

Error controls were introduced and used by Caponnetto and De Vito (2007), Caponnetto (2007), using Bernstein's inequality for Hilbert space-valued variables (see Pinelis and Sakhanenko; Yurinski).

Theorem (Caponetto, De Vito)

Define

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

then under assumption **(BernsteinNoise)** *we have the following:*

$$\mathbb{P}\left[\left\| (S+\lambda)^{-\frac{1}{2}} (T_n^* \mathbf{Y} - S_n h^*) \right\| \leq 2M \left(\sqrt{\frac{\mathcal{N}(\lambda)}{n}} + \frac{2}{\sqrt{\lambda} n} \right) \log \frac{6}{\delta} \right] \geq 1 - \delta .$$

Also, the following holds:

$$\mathbb{P}\left[\left\| (S+\lambda)^{-\frac{1}{2}} (S_n - S) \right\|_{HS} \leq 2 \left(\sqrt{\frac{\mathcal{N}(\lambda)}{n}} + \frac{2}{\sqrt{\lambda} n} \right) \log \frac{6}{\delta} \right] \geq 1 - \delta .$$

1. General regularization and kernel methods
2. Inverse learning/regression and relation to kernels
3. Rates for linear spectral regularization methods
4. Beyond the regular spectrum case

LIMITATIONS

- In the case of spectrum $\lambda_i \asymp i^{-1/s}$, we have shown that general regularization methods (with sufficient qualification) attain minimax rates over source conditions regularity sets.
- Remember λ_i are eigenvalues of kernel integral operator

$$T^* f = \int f(x) k(x, .) d\mathbb{P}_X(x) \,,$$

  hence depend on kernel **and** of sampling distribution!
- The assumption on a sharp power decay of the spectrum seems too strong, especially in the "distribution-free" philosophy:
  - decay rates such as $\lambda_i \asymp i^{-b} (\log i)^c (\log \log i)^d$?
  - spectrum with long plateaus separated by relative gaps?
  - multiscale behavior, shifting or switching between different polynomial-type regimes?

General spectrum: assumptions

Consider the following weaker assumption on the spectrum:

For any j sufficiently large and some $\nu_* \geq \nu^* > 1$,

$$\mathrm{OR}^<(\nu^*) \qquad \frac{\lambda_{2j}}{\lambda_j} \leq 2^{-\nu^*} ;$$

$$\mathrm{OR}^>(\nu_*) \qquad \frac{\lambda_{2j}}{\lambda_j} \geq 2^{-\nu_*} .$$

- Related to the notion of one-sided $\mathcal{O}$-regular variation
- Allows for a much broader range of behavior of the spectra
- Assumption $\mathrm{OR}^>(\nu_*)$ still implies that the spectrum is lower bounded by a power function: exponential decay of spectrum is not covered.

- Introduce: $\mathcal{F}(t) := \#\{j \in \mathbb{N} : \lambda_j \geq t\}\,, \qquad \mathcal{G}(t) := \dfrac{t^{2r+1}}{\mathcal{F}(t)}$
- Put

$$a_n := R\left(\mathcal{G}^{\leftarrow}\left(\frac{\sigma^2}{R^2 n}\right)\right)^{r+\theta}\,,$$

Theorem

Assume r, R, ν_*, ν^* are fixed positive constants and let $\mathcal{P}(\mathbb{P}_X, r, R)$ denote the set of distributions on $\mathcal{X} \times \mathcal{Y}$ with marginal $\mathbb{P}_X$ and satisfying **(SC)**(r, R) *and* **(BernsteinNoise)**.
If $\mathbb{P}_X$ satisfies $OR^{>}(\nu_*)$, then a_n is a lower minimax rate of convergence for the norm $\left\|S^{\theta}(\cdot)\right\|$.
If $\mathbb{P}_X$ satisfies $OR^{<}(\nu^*)$, the rate a_n is attained by an estimator based on any regularization function of qualification $q \geq r$ for the parameter choice

$$\lambda_n = \mathcal{G}^{\leftarrow}\left(\frac{\sigma^2}{R^2 n}\right)\,.$$

*(**NB:** ν_*, ν^* only influence multiplicative constants in front of rate)*

OVERVIEW:

- inverse problem setting under random i.i.d. design scheme
- "learning setting": unknown sampling distribution, related discretization error
- for source condition: Hölder of order r ;
- for ill-posedness: polynomial decay of eigenvalues of order s .
- Same regularization parameter works both in reconstruction error and prediction error.
- Minimax rates (incl. correct dependence on R, σ) are attained by general regularization methods (also Conjugate Gradient)
- rates of the form (for $\theta \in [0, \frac{1}{2}]$):

$$\left\| S^{\theta}(h^* - \widehat{h}) \right\|_{\mathcal{H}_K} \leq O\left(n^{-\frac{(r+\theta)}{2r+1+s}} \right) .$$

- matches "classical" rates in the white noise model (=sequence model) with $\sigma^{-2} \leftrightarrow n$.
- matching upper/lower bounds beyond polynomial spectrum decay

CONCLUSION/PERSPECTIVES

- We filled gaps in the existing picture for inverse learning methods.
- 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)
  - Hold-out principle: only valid for direct problem? But optimal parameter does not depend on risk norm: hope for validity in inverse case.

THANK YOU FOR YOUR ATTENTION!


F. Bauer, S. Pereverzev, and L. Rosasco.
On regularization algorithms in learning theory.
J. Complexity, 23(1):52–72, 2007.


N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart.
Convergence rates of general regularization methods for statistical inverse problems and applications.
SIAM J. Numer. Analysis, 45(6):2610–2636, 2007.


E. De Vito, L. Rosasco, and A. Caponnetto.
Discretization error analysis for Tikhonov regularization.
Analysis and Applications, 4(1):81–99, 2006.


S. Smale and D. Zhou.
Learning theory estimates via integral operators and their approximation.
Constructive Approximation, 26(2):153–172, 2007.


A. Caponnetto and Y. Yao.
Cross-validation based Adaptation for Regularization Operators in Learning
Analysis and Applications, 8(2):161–183 2010.


L. Dicker, D. Foster and D. Hsu
Kernel methods and regularization techniques for nonparametric regression: Minimax optimality and adaptation
ArXiv, 2016.