Convergence rates of spectral methods for statistical inverse learning problems

G. Blanchard

Universerstität Potsdam

UCL/Gatsby unit, 04/11/2015

Joint work with N. Mücke (U. Potsdam); N. Krämer (U. München)

1. The “inverse learning” setting

2. Rates for linear spectral regularization methods

3. Rates for conjugate gradient regularization

1. The “inverse learning” setting

2. Rates for linear spectral regularization methods

3. Rates for conjugate gradient regularization

DETERMINISTIC AND STATISTICAL INVERSE PROBLEMS

- Let A be a bounded operator between Hilbert spaces $\mathcal{H}_1 \to \mathcal{H}_2$ (assumed known)
- Classical (deterministic) inverse problem: observe

$$y^\sigma = Af^* + \sigma\eta\,, \tag{IP}$$

  under the assumption $\|\eta\| \leq 1$.
- Note: the $\mathcal{H}_2$-norm measures the observation error; the $\mathcal{H}_1$-norm measures the reconstruction error.
- Classical deterministic theory: see Engl, Hanke and Neubauer (2000).

DETERMINISTIC AND STATISTICAL INVERSE PROBLEMS

- Inverse problem

$$y^\sigma = Af^* + \sigma\eta\,. \tag{IP}$$

- What if noise is random? Classical statistical inverse problem model: η is a Gaussian white noise process on $\mathcal{H}_2$.
- Note: in this case (IP) is not an equation between elements in $\mathcal{H}_2$, but is to be interpreted as process on $\mathcal{H}_2$.
- Under Hölder source condition of order r and polynomial ill-posedness (eigenvalue decay) of order $1/s$, sharp minimax rates are known in this setting:

$$\left\|(A^*A)^\theta(\widehat{f} - f^*)\right\|_{\mathcal{H}_1} \asymp O\left(\sigma^{\frac{2(r+\theta)}{2r+1+s}}\right) \asymp O\left(\sigma^{\frac{2(\nu+b\theta)}{2\nu+b+1}}\right)\,,$$

for $\theta \in [0, \frac{1}{2}]$ ($\theta = 0$: inverse problem; $\theta = \frac{1}{2}$: direct problem.) (Alternate parametrization: $b := 1/s,\ \nu := rb$ "intrinsic regularity".)

LINEAR SPECTRAL REGULARIZATION METHODS

- Inverse problem (deterministic or statistical) where A is known.
- First consider the so-called "normal equation":

$$A^* y^\sigma = (A^* A) f^* + \sigma(A^* \eta)\,.$$

- Linear spectral methods: let $\zeta_\lambda(x) : \mathbb{R}_+ \to \mathbb{R}_+$ be a real function of 1 real variable which is an "approximation of $1/x$" and $\lambda > 0$ a tunig parameter.
- Define

$$\widehat{f}_\lambda = \zeta_\lambda(A^* A) A^* y^\sigma$$

- Examples: Tikhonov $\zeta_\lambda(x) = (x+\lambda)^{-1}$, spectral cut-off $\zeta_\lambda(x) = x^{-1}\mathbf{1}\{x \geq \lambda\}$, Landweber iteration polynomials, ν-methods . . .
- Under general conditions on ζ_λ, optimal/mimimax rates can be attained by such methods (Deterministic: Engl et al. , 2000; Stochastic noise: Bissantz et al, 2007)

Statistical learning

- "Learning" usually refers to the following setting:

$$(X_i, Y_i)_{i=1,\ldots,n} \quad \text{i.i.d.} \;\sim \mathbb{P}_{XY} \text{ on } \mathcal{X} \times \mathcal{Y}$$

  where $\mathcal{Y} \subset \mathbb{R}$,

- Goal: estimate some functional related to the dependency between X and Y,
- for instance (nonparametric) least squares regression: estimate

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

  and measure the quality of an estimator $\widehat{f}$ via

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

SETTING: "INVERSE LEARNING" PROBLEM

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

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

- 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 – we do not need to assume or consider any a priori Hilbert structure on $\mathcal{H}_2$
- the crucial stuctural assumption we make 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 setting was essentially introduced 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

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

1 The “inverse learning” setting

2 Rates for linear spectral regularization methods

3 Rates for conjugate gradient regularization

Empirical and population operators

- Define the (random) empirical evaluation operator

$$T_n : h \in \mathcal{H} \mapsto (h(X_1), \ldots, h(X_n)) \in \mathbb{R}^n$$

 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

$$T_n^* : (v_1, \ldots, v_n) \in \mathbb{R}^n \mapsto \frac{1}{n} \sum_{i=1}^{n} K(X_i, .) v_i \in \mathcal{H}$$

 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 $S_n = T_n^* T_n$ and its population counterpart $S = T^* T$.
- observe that S_n, S are both opertors $\mathcal{H}_K \to \mathcal{H}_K$; the intuition is that S_n is a (random) approximation of S.

- Recall the model with $h^* \in \mathcal{H}_K$:

$$Y_i = h^*(X_i) + \varepsilon_i \qquad \text{i.e.} \qquad \mathbf{Y} = T_n h^* + \varepsilon \,,$$

  where $\mathbf{Y} := (Y_1, \ldots, Y_n)$.

- Associated “normal equation”:

$$Z = T_n^* \mathbf{Y} = T_n^* T_n h^* + T_n^* \varepsilon = S_n h^* + T_n^* \varepsilon$$

- Idea (Rosasco, Caponnetto, De Vito, Odone): use methods from inverse problems literature
- Observe that there is also an error on the operator
- Use concentration principles to bound $\| T_n^* \varepsilon \|$ and $\| S_n - S \|$.

LINEAR SPECTRAL REGULARIZATION METHODS

- Linear spectral methods:

  $$\widehat{h}_\zeta = \zeta(S_n)Z$$

  for somme well-chosen function $\zeta : \mathbb{R} \to \mathbb{R}$ acting on the spectrum and "approximating" the function $x \mapsto x^{-1}$.

- Examples: Tikhonov $\zeta_\lambda(t) = (t+\lambda)^{-1}$, spectral cut-off $\zeta_\lambda(t) = t^{-1}\mathbf{1}\{t \geq \lambda\}$, Landweber iteration polynomials, ν-methods ...

Spectral regularization in kernel space

- Linear spectral regularization in kernel space is written

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

- notice

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

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

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

- equivalently:

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

  with

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

STRUCTURAL ASSUMPTIONS

- Two parameters determine attainable convergence rates:
- (Hölder) Source condition for the signal: for $r > 0$, define

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

  (can be generalized to “extended source conditions”, see e.g. Mathé and Pereverzev 2003)

- Ill-posedness: if $(\lambda_i)_{i \ge 1}$ is the sequence of positive eigenvalues of S in nonincreasing order, then define

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

  and

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

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 $\theta = 0$: inverse problem; $\theta = 1/2$: direct problem, 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

[1]: Smale and Zhou (2007)
[2]: Bauer, Pereverzev, Rosasco (2007)
[3]: Caponnetto, De Vito (2007)
[4]: Caponnetto (2006)

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

Matching lower bound: only for $\left\Vert \widehat{h} - h^* \right\Vert_{L^2(\mathbb{P}_X)}$ [2].

Compare to results known for regularization methods under Gaussian White Noise model: Mair and Ruymgaart (1996), Nussbaum and Pereverzev (1999), Bissantz, Hohage, Munk and Ruymgaart (2007).

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 $M' < \infty$ such that

$$\sup_{0<\lambda\leq 1} \sup_{0<t\leq 1} \lambda |\zeta_\lambda(t)| \leq \frac{M'}{,}$$

(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 (Mücke, Blanchard)

Assume r, R, b, 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*

$$\widehat{h}_{\lambda_n}^{(n)} = \zeta_{\lambda_n}(S_n) Z^{(n)}$$

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

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

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

$$\sup_{P \in \mathcal{P}(r,R,s,\beta)} P^{\otimes n} \left(\left\| S^\theta (h^* - \widehat{h}_{\lambda_n}^{(n)}) \right\|_{\mathcal{H}_K} > C(\log \eta^{-1}) R \left(\frac{\sigma^2}{R^2 n} \right)^{-\frac{(r+\theta)}{2r+1+s}} \right) \leq \eta.$$

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!
- 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 (Mücke, Blanchard)

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 (Mücke, Blanchard)

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 **(IP$^-$)**$(s, \beta) \cap$ **(IP$^+$)**(s, α) 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

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

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. The “inverse learning” setting

2. Rates for linear spectral regularization methods

3. Rates for conjugate gradient regularization

Partial Least Squares regularization

Consider first the classical linear regression setting

$$\mathbf{Y} = \mathbf{X}\omega + \varepsilon\,,$$

where $\mathbf{Y} := (Y_1, \ldots, Y_n)$; $\mathbf{X} := (X_1, \ldots, X_n)^t$; $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$.

- Algorithmic description of Partial Least Squares:
- find direction v_1 s.t.

$$v_1 = \operatorname*{Arg\,Max}_{v \in \mathbb{R}^d} \frac{\widehat{\mathrm{Cov}}(\langle X, v\rangle\,, Y)}{\|v\|} = \operatorname*{Arg\,Max}_{v \in \mathbb{R}^d} \frac{\mathbf{Y}^t\mathbf{X}v}{\|v\|} \propto \mathbf{X}^t\mathbf{Y}$$

- project $\mathbf{Y}$ orthogonally on $\mathbf{X}v$ yielding $\mathbf{Y}_1$
- iterate the procedure on the residual $\mathbf{Y} - \mathbf{Y}_1$
- The fit at step m is $\sum_{i=1}^m \mathbf{Y}_i$.
- Regularization is obtained by early stopping.

PLS and Conjugate Gradient

- An equivalent definition of PLS:

$$\omega_m = \operatorname*{Arg\,Min}_{\omega \in \mathcal{K}_m(\mathbf{X}\mathbf{X}^t, \mathbf{X}^t\mathbf{Y})} \left\| \mathbf{Y} - \mathbf{X}\omega \right\|^2$$

where

$$\mathcal{K}_m(A, z) = \operatorname{span}\left\{ z, Az, \ldots, A^{m-1}z \right\}$$

is a Krylov space of order m.

PLS and Conjugate Gradient

- An equivalent definition of PLS:

$$\omega_m = \operatorname*{Arg\,Min}_{\omega \in \mathcal{K}_m(\mathbf{X}\mathbf{X}^t, \mathbf{X}^t\mathbf{Y})} \left\| \mathbf{Y} - \mathbf{X}\omega \right\|^2$$

  where

$$\mathcal{K}_m(A, z) = \operatorname{span}\left\{ z, Az, \ldots, A^{m-1}z \right\}$$

  is a Krylov space of order m.

- This definition is equivalent to m steps of the conjugate gradient algorithm applied to iteratively solve the linear equation

$$A\omega = \mathbf{X}^t\mathbf{X}\omega = \mathbf{X}^t\mathbf{Y} = z$$

PLS and Conjugate Gradient

- An equivalent definition of PLS:

$$\omega_m = \operatorname*{Arg\,Min}_{\omega \in \mathcal{K}_m(\mathbf{X}\mathbf{X}^t, \mathbf{X}^t\mathbf{Y})} \left\| \mathbf{Y} - \mathbf{X}\omega \right\|^2$$

  where

$$\mathcal{K}_m(A, z) = \operatorname{span}\left\{ z, Az, \ldots, A^{m-1}z \right\}$$

  is a Krylov space of order m.

- This definition is equivalent to m steps of the conjugate gradient algorithm applied to iteratively solve the linear equation

$$A\omega = \mathbf{X}^t\mathbf{X}\omega = \mathbf{X}^t\mathbf{Y} = z$$

- For any fixed m, the fit $\mathbf{Y}_m = \mathbf{X}\omega_m$ is a nonlinear function of $\mathbf{Y}$.

Properties of conjugate gradient

- by definition ω_m has the form

$$\omega_m = p_m(A)z = \mathbf{X}^t p_m(\mathbf{X}\mathbf{X}^t)\mathbf{Y},$$

where p_m is a polynomial of degree $\leq m-1$.

- of particular interest are the residual polynomials

$$r_m(t) = 1 - tp_m(t)\,; \qquad \|\mathbf{Y} - \mathbf{Y}_m\| = \left\| r_m(\mathbf{X}\mathbf{X}^t)\mathbf{Y} \right\|$$

- the polynomials r_m form a family of **orthogonal polynomials** for the inner product

$$\langle p, q \rangle = \left\langle p(\mathbf{X}\mathbf{X}^t)\mathbf{Y}, \mathbf{X}\mathbf{X}^t q(\mathbf{X}\mathbf{X}^t)\mathbf{Y} \right\rangle$$

and with the normalization $r_m(0) = 1$.

- the polynomials r_m follow an order 2 recurrence relation of the type

$$r_{m+1}(t) = a_m t r_m(t) + b_m r_m(t) + c_m r_{m-1}(t)$$

($\rightarrow$ simple implementation)

ALGORITHM FOR CG/PLS

```
Initialize: $\omega_0 = 0$; $r_0 = \mathbf{X}^t\mathbf{Y}$; $g_0 = r_0$
for $m = 0, \ldots, (m_{\max} - 1)$ do
    $\alpha_m = \left\|\mathbf{X}r_m\right\|^2 / \left\|\mathbf{X}^t\mathbf{X}g_m\right\|^2$
    $\omega_{m+1} = \omega_m + \alpha_m g_m$ (update)
    $r_{m+1} = r_m - \alpha_m \mathbf{X}^t\mathbf{X}g_m$ (residuals)
    $\beta_m = \left\|\mathbf{X}r_{m+1}\right\|^2 / \left\|\mathbf{X}r_m\right\|^2$
    $g_{m+1} = r_{m+1} + \beta_m g_m$ (next direction)
end for
Return: approximate solution $\omega_{m_{\max}}$
```

Kernel-CG regularization
(≈ Kernel Partial Least Squares)

- Define the m-th iterate of CG as

$$\widehat{h}_{CG(m)} = \operatorname*{Arg\,Min}_{h \in \mathcal{K}_m(S_n, T_n^* \mathbf{Y})} \left\| T_n^* \mathbf{Y} - h \right\|_{\mathcal{H}} ,$$

where $\mathcal{K}_m$ denotes Krylov space:

$$\mathcal{K}_m(A, z) = \operatorname{span}\left\{ z, Az, \ldots, A^{m-1} z \right\}$$

- equivalently:

$$\alpha_{CG(m)} = \operatorname*{Arg\,Min}_{\alpha \in \mathcal{K}_m(K_n, \mathbf{Y})} \left\| K_n^{\frac{1}{2}} \left(\mathbf{Y} - K_n \alpha \right) \right\|^2$$

and

$$\widehat{h}_{CG(m)} = \sum_{i=1}^{n} \alpha_{CG(m),i} K(X_i, .) \, .$$

Rates for CG

Consider the following stopping rule for some fixed τ

$$\widehat{m} := \min\left\{ m \geq 0 : \left\| T_n^*(T_n \widehat{h}_{CG(m)} - \mathbf{Y}) \right\| \leq \tau \left(\frac{1}{n} \log^2 \frac{6}{\delta} \right)^{\frac{r+1}{2r+1+s}} \right\} . \qquad (1)$$

Theorem (Blanchard, Krämer)

Assume **(BernsteinNoise)***,* **SC(**r, R**)***,* **IP(**s, β**)** *hold; let* $\theta \in [0, \frac{1}{2})$*. Then for* τ *large enough, with probability larger than* $1 - \delta$ *:*

$$\left\| S^\theta (\widehat{h}_{CG(\widehat{m})} - h^*) \right\|_{\mathcal{H}_k} \leq c(r, R, s, \beta, \tau) \left(\frac{1}{n} \log^2 \frac{6}{\delta} \right)^{\frac{r+\theta}{2r+1+s}} .$$

Technical tools: again, concentration of the error in appropriate norm, and suitable reworking of the arguments of Nemirovskii (1980) for deterministic CG.

OUTER RATES

- It it natural (for the prediction problem) to assume extension of source condition for $h^* \notin \mathcal{H}$ (now assuming $h^* \in L^2(\mathbb{P}_X)$)

$$\mathbf{SC}_{\mathbf{outer}}(r,R): \quad \left\| B^{-(r+\frac{1}{2})} h^* \right\|_{L^2} \leq R \qquad \text{(for } B := TT^*\text{)}$$

  to include the possible range $r \in (-\frac{1}{2}, 0]$.
- For such “outer” source conditions , even for Kernel ridge regression and for the direct (=prediction) problem, there are no known results without additional assumptions to reach the optimal rate $\mathcal{O}\left(n^{-\frac{r+\frac{1}{2}}{2r+1+s}}\right)$.
- Mendelson and Neeman (2009) make assumptions on the sup norm of the eigenfunctions of the integral operator
- Caponnetto (2006) assumes additional unlabeled examples $X_{n+1}, \ldots, X_{\widetilde{n}}$ are available, with

$$\frac{\widetilde{n}}{n} \sim \mathcal{O}\left(n^{\frac{(1-2r-s)_+}{2r+1+s}}\right)$$

Construction with unlabeled data

- assume $\widehat{n}$ i.i.d. X-examples are available, out of which n are labeled.
- extend the n vector $\mathbf{Y}$ to a $\widetilde{n}$-vector

$$\widetilde{\mathbf{Y}} = \frac{\widetilde{n}}{n}\,(Y_1, \ldots, Y_n, 0, \ldots, 0)$$

- perform the same algorithm as before on $\mathbf{X}, \widetilde{\mathbf{Y}}$.
- notice in particular that

$$T^*_{\widetilde{n}}\widetilde{\mathbf{Y}} = T^*_n\mathbf{Y}.$$

Construction with unlabeled data

- assume $\widehat{n}$ i.i.d. X-examples are available, out of which n are labeled.
- extend the n vector $\mathbf{Y}$ to a $\widetilde{n}$-vector

$$\widetilde{\mathbf{Y}} = \frac{\widetilde{n}}{n}\,(Y_1, \ldots, Y_n, 0, \ldots, 0)$$

- perform the same algorithm as before on $\mathbf{X}, \widetilde{\mathbf{Y}}$.
- notice in particular that

$$T^*_{\widetilde{n}}\widetilde{\mathbf{Y}} = T^*_n\mathbf{Y}.$$

- Recall:

$$\widehat{h}_{CG1(m)} = \operatorname*{Arg\,Min}_{h \in \mathcal{K}_m(\widetilde{S}_n, T^*_n\mathbf{Y})} \left\| T_n\mathbf{Y} - h \right\|_{\mathcal{H}}$$

- equivalently:

$$\alpha = \operatorname*{Arg\,Min}_{\omega \in \mathcal{K}_m(\widetilde{K}_n, \widetilde{\mathbf{Y}})} \left\| \widetilde{K}_n^{\frac{1}{2}} \left(\widetilde{\mathbf{Y}} - \widetilde{K}_n\alpha \right) \right\|^2$$

OUTER RATES FOR CG REGULARIZATION

Consider the following stopping rule for some fixed $\tau > \frac{3}{2}$,

$$\widehat{m} := \min\left\{ m \geq 0 : \left\| T_n^*(T_n \widehat{h}_{CG(m)} - \mathbf{Y}) \right\| \leq \tau M \left(\frac{4\beta}{n} \log^2 \frac{6}{\delta} \right)^{\frac{r+1}{2r+1+s}} \right\} . \tag{2}$$

Furthermore assume

$$\textbf{(BoundedY)} : \quad |Y| \leq M \qquad \text{a.s.}$$

Theorem

Assume **(BoundedY)***,* $\mathbf{SC}_{\mathbf{outer}}(r, R)$*,* $\mathbf{IP}^+(s, \beta)$*, and* $r \in (-\min(s, \frac{1}{2}), 0)$ *.*

Assume unlabeled data is available with $\frac{\widetilde{n}}{n} \geq \left(\frac{16\beta^2}{n} \log^2 \frac{6}{\delta} \right)^{-\frac{(-2r)_+}{2r+1+s}}$ *. Then for* $\theta \in [0, r + \frac{1}{2})$*, with probability larger than* $1 - \delta$ *:*

$$\left\| B^{-\theta}(Th_{\widehat{m}} - h^*) \right\|_{L^2} \leq c(r, \tau)(M + R) \left(\frac{16\beta^2}{n} \log^2 \frac{6}{\delta} \right)^{\frac{r + \frac{1}{2} - \theta}{2r+1+s}} .$$

OVERVIEW:

- inverse problem setting under random i.i.d. design scheme (“learning setting”),
- for source condition: Hölder of order r ;
- for ill-posedness: polynomial decay of eigenvalues of order s ;
- rates of the form (for $\theta \in [0, \frac{1}{2}]$):

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

- rates established for general linear spectral methods, as well as CG.
- matching lower bound.
- matches “classical” rates in the white noise model (=sequence model) with $\sigma^{-2} \leftrightarrow n$.
- extension to “outer rates” ($r \in (-\frac{1}{2}, 0)$) if additional **unlabeled** data available.

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.