Statistical inverse learning and regularization by projection

Tapio Helin
tapio.helin@lut.fi
School of Engineering Science
LUT University

November 29, 2022

Setting the stage

Consider a linear inverse problem

$$g = Af,$$

where $A\colon \mathcal{H} \to H_k$ is a one-to-one linear operator without continuous inverse.

Assumptions:

- H_k is a reproducing kernel Hilbert space (RKHS) induced by the kernel $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ such that $H_k \subset L^2(\mathcal{D}, \nu)$,
- ν is a (design) probability measure on $\mathcal{D} \subset \mathbb{R}^d$ and
- $\mathcal{H}$ be a separable real Hilbert space.

TH, *LS approximations in linear statistical inverse learning problems*, arXiv:2211.12121

Setting the stage

Consider a linear inverse problem

$$g = Af,$$

where $A\colon \mathcal{H} \to H_k \subset L^2(\mathcal{D}, \nu)$ is a one-to-one linear operator without continuous inverse.

Statistical inverse learning problem:

- $(x_n)_{i=1}^N \subset \mathcal{D}$ be i.i.d. in ν and
- noisy observations $\mathbf{y}^\delta = (y_n^\delta)_{n=1}^N \in \mathbb{R}^N$ such that

$$y_n = g^\dagger(x_n) + \delta\epsilon_n, \quad n = 1, \ldots, N,$$

where $g^\dagger = Af^\dagger$ and $\epsilon_n \sim \mathcal{N}(0, 1)$

Find $f^\dagger \in \mathcal{H}$!

Some background

- Regularization by projection has an extensive literature
- Mathé–Pereverzev (2001): optimal discretization in Hilbert scales for the statistical inverse problem
- Blanchard–Mücke (2018): minimax optimal rates for spectral regularization methods
- Since 2018, extensions to non-linear and Hilbert scales (Mathé, Rastogi and others) and convex penalties (Burger, **TH** et al)

What to expect (as meta-theorems)

Suppose f_α is some probabilistic estimator of $f^\dagger$. Assume ν has suitable properties and $f^\dagger$ satisfies a source condition.

Theorem (Probabilistic bound)

Let $0 < \eta < 1$ *satisfy* $\log(1/\eta) \leq \sqrt{N}\alpha^r$. *Then*

$$\left\| f_\alpha - f^\dagger \right\|_{\mathcal{H}} \lesssim \alpha^s + \log\left(\frac{1}{\eta}\right) \cdot \frac{\delta}{\alpha^t \sqrt{N}}$$

with probability greater than $1 - \eta$.

Notice that the condition on η is equivalent to $\eta \geq \exp(-\sqrt{N}\alpha^r)$.

What to expect (as meta-theorems)

Recall that $\mathbb{E}X = \int_0^\infty \mathbb{P}(X > z)dz$ for positive X.

Interpolation: If

$$\begin{aligned} \mathbb{P}(X > a - b\log\eta) &\leq \eta \quad \text{for } \eta > \eta_0 \quad \text{and} \\ \mathbb{P}(X > a' - b'\log\eta) &\leq \eta \quad \text{for } \eta \in [0,1] \end{aligned}$$

then

$$\mathbb{E}X^p \lesssim a^p + b^p + \eta_0\left[(a')^p + (-b'\log\eta_0)^p\right]$$

Theorem (Bound in expectation)

$$\mathbb{E}\|f^\dagger - f_\alpha\|_{\mathcal{H}}^p \lesssim m^{-ps} + \frac{\delta^p m^{p\gamma}}{N^{\frac{p}{2}}} + l.o.t..$$

Least-squares estimators

Assumption

Let V_m, $m \geq 1$, be finite-dimensional subspaces of $\mathcal{H}$ such that

- $\dim V_m = m$,
- $V_m \subset V_{m+1}$ *and*
- $\overline{\cup_{m=1}^{\infty} V_m} = \mathcal{H}$.

Define the ML estimator on V_m:

$$f_{m,N} = \arg\min_{f \in V_m} \left\| S_X A f - \mathbf{y}^{\delta} \right\|_N^2 ,$$

where $\| \cdot \|_N$ induced by $\langle \mathbf{x}, \mathbf{z} \rangle_N = \frac{1}{N} \sum_{n=1}^{N} x_n z_n$ with $\mathbf{x}, \mathbf{z} \in \mathbb{R}^N$.

Normal operator

Consider the normal sampling operator

$$B_X = A^* S_X^* S_X A = A^* \left(\frac{1}{N} \sum_{n=1}^{N} K_{x_n} \otimes K_{x_n} \right) A : \mathcal{H} \to \mathcal{H}$$

What is the limit as $N \uparrow \infty$? We denote

$$A_\nu = \iota A : \mathcal{H} \to L^2(\mathcal{D}, \nu),$$

where $\iota : H_k \to L^2(\mathcal{D}, \nu)$ is the canonical injection map, and introduce

$$B_\nu := A_\nu^* A_\nu = A^* \left(\int_{\mathcal{D}} K_x \otimes K_x \nu(dx) \right) A : \mathcal{H} \to \mathcal{H}$$

Fundamental concentration result

Set $L_x := A^* K_x \in \mathcal{H}$ for $x \in \mathcal{D}$. Recall

$$B_\nu = \int_{\mathcal{D}} L_x \otimes L_x \nu(dx) \quad \text{and} \quad B_X = \frac{1}{N} \sum_{n=1}^{N} L_{x_n} \otimes L_{x_n}.$$

Corollary (Blanchard–Mücke, Prop. 5.5)

Suppose $B_\nu : \mathcal{H} \to \mathcal{H}$ is a Hilbert–Schmidt operator and $\|B_\nu\| \leq 1$. For any sample size $N > 0$ and $0 < \eta < 1$ it holds that

$$\|B_\nu - B_X\|_{\mathrm{HS}} \leq 6 \log\left(\frac{2}{\eta}\right) \frac{1}{\sqrt{N}}$$

with probability greater than $1 - \eta$.

Source condition and smoothness

Source condition: we assume $f^\dagger$ in

$$\Theta(s) = \{f \in \mathcal{H} \mid \|(I - P_m)f\|_{\mathcal{H}} \leq R_0(m+1)^{-s} \text{ for all } m \geq 0\} \subset \mathcal{H},$$

where $s, R_0 > 0$ and P_m is an orthogonal projection to $V_m \subset \mathcal{H}$, $m \geq 1$ (use convention $P_0 = 0$).

Smoothness: Let $\mathcal{P}(\mathcal{D})$ denote all probability measures on domain $\mathcal{D} \subset \mathbb{R}^d$ and introduce

$$\begin{aligned} \mathcal{P}^{>}(t) &= \left\{\nu \in \mathcal{P}(\mathcal{D}) \,\middle|\, \lambda_{min}(P_m B_\nu P_m) \geq C m^{-t} \quad \forall m \in \mathbb{N}\right\} \\ \mathcal{P}^{\times} &= \left\{\nu \in \mathcal{P}(\mathcal{D}) \,\middle|\, \left\|(P_m B_\nu P_m)^\dagger B_\nu (I - P_m)\right\| \leq C \quad \forall m \in \mathbb{N}\right\}, \end{aligned}$$

where $\lambda_{min}(P_m B_\nu P_m)$ = smallest eigenvalue.

Probabilistic concentration

Theorem

Suppose $\nu \in \mathcal{P}^{>}(t) \cap \mathcal{P}^{\times}$ *and* $f^{\dagger} \in \Theta(s)$ *for some constants* $s, t > 0$. *Let* $0 < \eta < 1$ *satisfy*

$$\log\left(\frac{8}{\eta}\right) \leq \frac{1}{12}\sqrt{N}\lambda_{\min}(P_m B_\nu P_m).$$

Then

$$\left\| f_{m,N} - f^{\dagger} \right\|_{\mathcal{H}} \lesssim m^{-s} + \log\left(\frac{8}{\eta}\right) \cdot \delta \left(\frac{m^t}{N} + \frac{m^{\frac{t+1}{2}}}{\sqrt{N}} \right)$$

with probability greater than $1 - \eta$.

Proof schematics

- Decompose error:

$$\begin{aligned} f_{m,N} - f^{\dagger} &= (P_m B_X P_m)^{\dagger} (S_X A)^{*} \mathbf{y}^{\delta} - f^{\dagger} \\ &= \left((P_m B_X P_m)^{\dagger} B_X - I \right) f^{\dagger} + \delta (P_m B_X P_m)^{\dagger} (S_X A)^{*} \boldsymbol{\epsilon} \\ &=: I_1 + I_2, \end{aligned}$$

  where I_1 is bias/approximation error and I_2 is variance.
- Find probabilistic bounds for I_1 and I_2 and combine

Brief insights: Bound on bias

Lemma (Modified concentration)

Let $0 < \eta < 1$ *satisfy* $\log\left(\frac{2}{\eta}\right) \leq \frac{1}{12}\sqrt{N}\lambda_m$. *With probability greater than* $1 - \eta$, *it holds that* $\|B_X - B_\nu\|_{\mathrm{HS}} \leq \frac{1}{2}\lambda_m$.

Proposition

Suppose $\nu \in \mathcal{P}^\times$ *and* $f^\dagger \in \Theta(s)$. *Let* $0 < \eta < 1$ *satisfy* $\log\left(\frac{8}{\eta}\right) \leq \frac{1}{12}\sqrt{N}\lambda_m$. *Then* $\|I_1\|_{\mathcal{H}} \lesssim m^{-s}$ *with probability greater than* $1 - \eta/4$.

Follows from

$$I_1 = \left((P_m B_X P_m)^\dagger B_X - I\right) f^\dagger = \Big[\underbrace{(P_m B_X P_m)^\dagger B_X}_{\text{lemma+assump.}} + I\Big](I - P_m) f^\dagger$$

Brief insights: Bound on variance

Proposition

Suppose $\nu \in \mathcal{P}^{<}(t, D_1) \cap \mathcal{P}^{\times}(D_2)$ *and let* $0 < \eta < 1$ *satisfy* $\log\left(\frac{8}{\eta}\right) \leq \frac{1}{12}\sqrt{N}\lambda_m$. *Then*

$$\|I_2\|_{\mathcal{H}} \lesssim \delta \log\left(\frac{8}{\eta}\right)\left(\frac{m^t}{N} + \frac{m^{\frac{t+1}{2}}}{\sqrt{N}}\right),$$

with probability greater than $1 - \frac{3}{4}\eta$.

Idea: decompose I_2 into three terms

$$I_2 = \delta \underbrace{(P_m B_X P_m)^{-\frac{1}{2}}}_{=:K_1} \cdot \underbrace{(P_m B_X P_m)^{-\frac{1}{2}}(P_m B_\nu P_m)^{\frac{1}{2}}}_{=:K_2} \cdot \underbrace{(P_m B_\nu P_m)^{-\frac{1}{2}} A^* S_X^* \epsilon}_{=:K_3}.$$

Probabilistic concentration: revisited

Theorem

Suppose $\nu \in \mathcal{P}^{>}(t) \cap \mathcal{P}^{\times}$ and $f^{\dagger} \in \Theta(s)$ for some constants $s, t > 0$. Let $0 < \eta < 1$ satisfy $\log\left(\frac{8}{\eta}\right) \leq \frac{1}{12}\sqrt{N}\lambda_{\min}(P_m B_\nu P_m)$. Then

$$\left\| f_{m,N} - f^{\dagger} \right\|_{\mathcal{H}} \lesssim m^{-s} + \log\left(\frac{8}{\eta}\right) \cdot \delta \left(\frac{m^t}{N} + \frac{m^{\frac{t+1}{2}}}{\sqrt{N}} \right)$$

with probability greater than $1 - \eta$.

Proof. If we have independent events E_1 and E_2 such that $\mathbb{P}(E_1) \geq 1 - \frac{\eta}{4}$ and $\mathbb{P}(E_2) \geq 1 - \frac{3\eta}{4}$, respectively, then

$$\mathbb{P}(E_1 \cap E_2) = \left(1 - \frac{\eta}{4}\right)\left(1 - \frac{3\eta}{4}\right) = 1 - \eta + \frac{3\eta^2}{16} \geq 1 - \eta.$$

How to derive expectations?

We define our nonlinear estimator according to

$$g_{m,N}^R = T_R(f_{m,N}), \qquad T_R(f) = \begin{cases} f & \text{if } \|f\| \le R, \\ 0, & \text{otherwise} \end{cases}$$

where R is set below and will depend on m and δ.

Idea:

$$\mathbb{E}\|f^\dagger - g_{m,N}^R\|_{\mathcal{H}}^p$$
$$\lesssim \int_{\Omega_+ \cap \Omega_R} \|f^\dagger - f_{m,N}\|_{\mathcal{H}}^p \, \mathbb{P}(\mathrm{d}\omega) + R^p(\mathbb{P}(\Omega_+ \cap \Omega_R^c) + \mathbb{P}(\Omega_-)),$$

where

- $\Omega_R = \{\|f_{m,N}\|_{\mathcal{H}} \le R\}$,
- $\Omega_+ := \left\{\omega \in \Omega : \|B_X - B_\nu\|_{\mathrm{HS}} \le \frac{1}{2}\lambda_m\right\}$
- $\Omega_- = \Omega_+^c$.

Concentration in expectation

Theorem

Suppose $\nu \in \mathcal{P}^{>}(t) \cap \mathcal{P}^{\times}$ *and* $f^\dagger \in \Theta(s)$ *for* $2s - t + 1 > 0$*. For the parameter choice*

$$m = \left(\frac{\delta}{\sqrt{N}}\right)^{-\frac{2}{2s+t+1}}$$

and $R = R(m, \delta) \propto \delta / \lambda_{min}(P_m B_\nu P_m)$*, it holds that*

$$\left(\mathbb{E}_{\boldsymbol{\nu}_N} \left\| g_{m,N}^R - f^\dagger \right\|_{\mathcal{H}}^p \right)^{\frac{1}{p}} \lesssim \left(\frac{\delta}{\sqrt{N}}\right)^{\frac{2s}{2s+t+1}} =: a_{N,\delta}$$

where $\boldsymbol{\nu}_N = \otimes_{n=1}^N \nu$*.*

.

Finally: Minimax optimality

Corollary

Let $s, t, R_0 > 0$, $2s - t + 1 > 0$, and
$\mathcal{P}' = \left\{ \nu \in \mathcal{P} \;\middle|\; \nu \in \mathcal{P}^{>}(t) \cap \mathcal{P}^{\times} \cap \mathcal{P}^{<}(t) \right\}$ *and* $\Theta' = \Theta(s)$. *Then* $g^R_{m,N}$ *with parameter choice rules on previous slide is strong minimax optimal in* L^p *for all* $p > 0$ *over the class of admissible models specified by* Θ' *and* $\mathcal{P}'$.

That is: the rate $a_{N,\delta}$ is also strong minimax lower rate of convergence such that

$$\inf_{f^\dagger \in \Theta(s)} \liminf_{n \to \infty} \inf_{\hat{f}} \sup_{\nu \in \mathcal{P}'} \frac{\left(\mathbb{E}_{\nu_N} \left\| \hat{f} - f^\dagger \right\|^p_{\mathcal{H}} \right)^{\frac{1}{p}}}{a_{N,R_0,\delta}} > 0,$$

where the infimum is taken over all estimators (measurable mappings) $\hat{f} : \mathcal{D}^N \times \mathbb{R}^N \to \mathcal{H}$.

Outlook

- Generalize to nonlinear problems (obviously)
- Consider subspaces induced by the (random) data; what if conditions such as

$$\lambda_{min}(P_m B_\nu P_m) \geq Cm^{-t}$$

  are satisfied with given probability (think Krylov spaces, power method approximation to spectrum, data-driven projections etc).
- Sparse dictionaries etc.

TH, *LS approximations in linear statistical inverse learning problems*, arXiv:2211.12121