Optimal Rates for Regularized Least-Squares Algorithm

Caponnetto, De Vito

Arthur Gretton's notes

February 18, 2013

Problem setup

We want to minimize the squared error

$$\mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \|f(x) - y\|_{\mathcal{Y}}^2 \, d\rho(x,y),$$

for some Hilbert spaces $\mathcal{X}$, $\mathcal{Y}$. If there were no constraints on f, the best solution would be:

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x).$$

Problem setup

We want to minimize the squared error

$$\mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \|f(x) - y\|_{\mathcal{Y}}^2 \, d\rho(x, y),$$

for some Hilbert spaces $\mathcal{X}$, $\mathcal{Y}$. If there were no constraints on f, the best solution would be:

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x).$$

In practice, f is in a hypothesis class $\mathcal{H}$.

A learning algorithm is universally consistent if it takes data $\mathbf{z} := ((x_1, y_1), \ldots, (x_\ell, y_\ell))$, returns $f_{\mathbf{z}} \in \mathcal{H}$ such that

$$\lim_{\ell\to\infty} \mathbb{P}\left[\mathcal{E}(f_{\mathbf{z}}) - \inf_{f\in\mathcal{H}}[f] > \epsilon\right] = 0 \quad \forall\epsilon > 0$$

(meaning: only as good as best function in $\mathcal{H}$)

Motivating example: $\mathcal{Y} = \mathbb{R}^n$

We propose to solve this with a vector-valued RKHS
Motivating example: kernel ridge regression to $\mathcal{Y} = \mathbb{R}^n$.
We write elements of $\mathcal{H}$ as vectors of scalar-valued RKHS functions,

$$f(\cdot) := \left[\begin{array}{ccc} f_1(\cdot) & \dots & f_n(\cdot) \end{array} \right],$$

with inner product

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \langle f_i, g_i \rangle_{\mathcal{H}_i} .$$

We write $K(x,t)$ as an $n \times n$ diagonal matrix,

$$K(x,t) = \mathrm{diag} \left[\begin{array}{ccc} k_1(x,t) & \dots & k_n(x,t) \end{array} \right].$$

$$K_x = \mathrm{diag} \left[\begin{array}{ccc} k_1(x,\cdot) & \dots & k_n(x,\cdot) \end{array} \right]$$

The two essential RKHS properties: $\mathcal{Y} = \mathbb{R}^n$

Property 1: Reproducing property

$$\langle K_x y, f \rangle_{\mathcal{H}} = \langle y, f(x) \rangle_{\mathcal{Y}} .$$

This holds, since

$$\langle K_x y, f \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \langle y_i k_i(x_i, \cdot), f_i(\cdot) \rangle_{\mathcal{H}_i}$$

Property 2: reprooducing property between kernels:

$$\begin{aligned} \langle y, K(x,t) z \rangle_{\mathcal{Y}} &= y^\top \mathrm{diag}\left[\begin{array}{ccc} k_1(x,t) & \ldots & k_n(x,t) \end{array}\right] z \\ &= \sum_{i=1}^{n} \langle y_i k_i(x, \cdot), z_i k_i(t, \cdot) \rangle_{\mathcal{H}_i} \\ &= \langle K_x y, K_t z \rangle_{\mathcal{H}} . \end{aligned}$$

A non-diagonal case, $\mathcal{Y} = \mathbb{R}^n$

- $j \in \{1, \dots, m\}$
- D_j an $r \times r$ diagonal matrix of scalar valued kernels.
- A_j an $r \times n$ matrix

A valid $\mathbb{R}^n$ valued kernel is

$$K(x, t) = \sum_{j=1}^{m} A_j^\top D_j A_j.$$

A non-diagonal case, $\mathcal{Y} = \mathbb{R}^n$

- $j \in \{1, \ldots, m\}$
- D_j an $r \times r$ diagonal matrix of scalar valued kernels.
- A_j an $r \times n$ matrix

A valid $\mathbb{R}^n$ valued kernel is

$$K(x,t) = \sum_{j=1}^{m} A_j^\top D_j A_j.$$

In general case (infinite dimensional):

$$\begin{array}{rl} K_x & \mathcal{Y} \to \mathcal{H} \subset \mathcal{Y}^{\mathcal{X}} \\ K(x,u) & \mathcal{Y} \to \mathcal{Y} \end{array}$$

Least squares regression

Empirical problem setting: minimize

$$E(f) = \sum_{j=1}^{\ell} \|y_j - f(x_j)\|_{\mathcal{Y}}^2 + \lambda \|f\|_{\mathcal{H}}^2 .$$

The unique minimizer of the above takes the form:

$$f_{\mathbf{z}}^{\lambda} = \sum_{j=1}^{\ell} K_{x_j} c_j$$

where $c_j \in \mathcal{Y}$ are the solutions to

$$\sum_{i=1}^{\ell} \left(K(x_j, x_i) + \mu\delta_{ij}\right) c_i = y_j$$

Example: tensor product RKHS (not in paper!)

Infinite dimensional case: the previous diagonal example generalizes straightforwardly

- Recall tensor product definition:

$$[y \otimes x]u = y\langle x, u\rangle_{\mathcal{X}}.$$

- We define

$$K_x y = y \otimes x$$

The map K_x is a linear operator from $\mathcal{Y}$ to $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ as required: given $u \in \mathcal{X}$, it is defined as

$$\begin{aligned}[K_x y]\,(u) &= [y \otimes x]u \\ &\quad y\langle x, u\rangle_{\mathcal{X}} \in \mathcal{Y}.\end{aligned}$$

Example: tensor product RKHS (not in paper!)

- We define
$$K(x,u) := I_{\mathcal{Y}} \langle x, u\rangle_{\mathcal{X}}.$$
This is a map from $\mathcal{Y} \to \mathcal{Y}$ as required.
- We want to ensure the relation
$$\langle K_x y, K_u v\rangle_{\mathcal{H}} = \langle v, K(x,u)y\rangle_{\mathcal{Y}}.$$
This just the standard Hilbert-Schmidt inner product, matrix analogue is $\mathrm{tr}(A^\top B)$:
$$\begin{aligned}\langle K_x y, K_u v\rangle_{\mathcal{H}} &= \langle y\otimes x, v \otimes u\rangle_{\mathrm{HS}}\\ &= \langle v, [y\otimes x]\, u\rangle_{\mathcal{Y}}\\ &= \left\langle v, \underbrace{I_{\mathcal{G}} \langle x,u\rangle_{\mathcal{X}}}_{K(x,u)} y\right\rangle_{\mathcal{Y}}\\ &= \langle x,u\rangle_{\mathcal{X}} \langle y,v\rangle_{\mathcal{Y}}.\end{aligned}$$

Form of result

The "upper rate" obtained is:

$$\lim_{\tau\to\infty} \limsup_{\ell\to\infty} \sup_{\rho\in\mathcal{P}} \mathbb{P}_{\mathbf{z}\sim\rho^\ell} \underbrace{\left[\mathcal{E}\left[f_{\mathbf{z}}^{\lambda_\ell}\right] - \inf_{f\in\mathcal{H}} \mathcal{E}[f] > \tau\alpha_\ell\right]}_{\text{(a)}} = 0$$

for $f_{\mathbf{z}}^{\lambda_\ell}$ obtained via least squares regression, where

1. Term (a) is event "error of $f_{\mathbf{z}}^{\lambda_\ell}$ compared to best $f \in \mathcal{H}$ is worse than $\tau\alpha_\ell$ for given ρ, ℓ.
2. $\sup_{\rho\in\mathcal{P}}$ is "hardest" probability in the family, *given* ℓ
3. $\limsup_{\ell\to\infty}$ is the limiting upper bound. Eg. for different ℓ, a different ρ might be the hardest.
4. $\lim_{\tau\to\infty}$ since there is a constant τ in front of α_ℓ, which we don't want to figure out. I.e. for some "sufficiently large value of τ" (and hence all τ above it) the limit is zero.

Assumptions on distribution (1)

Family of probabilities is $\mathcal{P}(b, c)$. Here $1 \leq b < \infty$, $1 \leq c \leq 2$.
Assumptions:

1. y has finite variance,

$$\int \|y\|_{\mathcal{Y}}^2 d\rho(x, y) < \infty,$$

2. A noise assumption is satisfied: noise must be bounded, Gaussian, or sub-Gaussian. Technical condition:

– there are two positive constants Σ, M such that

$$\int_Y \left(e^{\frac{\|y - f_{\mathcal{H}}(x)\|_Y}{M}} - \frac{\|y - f_{\mathcal{H}}(x)\|_Y}{M} - 1 \right) d\rho(y|x) \leq \frac{\Sigma^2}{2M^2} \tag{9}$$

for ρ_X-almost all $x \in X$.

Assumptions on the distribution (2)

Define covariance operator T on random variable X,

$$T_x = K_x K_x^* \in \mathcal{L}(\mathcal{H}) \qquad T := \int_{\mathcal{X}} T_x d\rho_X(x).$$

Given the singular value decomposition (where N can be $+\infty$),

$$T := \sum_{n=1}^{N} t_n \langle \cdot, e_n \rangle_{\mathcal{H}} e_n.$$

1. Assume $N = +\infty$, then $\exists\, \alpha, \beta > 0$ such that

$$\alpha \leq n^b t_n \leq \beta$$

   (effective dimension of $\mathcal{H}$ wrt ρ_x)
2. The infimum $\inf_{f\in\mathcal{H}}[f]$ is attained at $f_{\mathcal{H}}$ satisfying, for $\|g\|^2_{\mathcal{H}} \leq R < \infty$,

$$f_{\mathcal{H}} = T^{(c-1)/2} g$$

   (complexity of regression function)

The bound

Theorem 1. *Given* $1 < b \leq +\infty$ *and* $1 \leq c \leq 2$, *let*

$$\lambda_\ell = \begin{cases} \left(\frac{1}{\ell}\right)^{\frac{b}{bc+1}} & b < +\infty \quad c > 1 \\ \left(\frac{\log \ell}{\ell}\right)^{\frac{b}{b+1}} & b < +\infty \quad c = 1 \\ \left(\frac{1}{\ell}\right)^{\frac{1}{2}} & b = +\infty \end{cases} \tag{19}$$

and

$$a_\ell = \begin{cases} \left(\frac{1}{\ell}\right)^{\frac{bc}{bc+1}} & b < +\infty \quad c > 1 \\ \left(\frac{\log \ell}{\ell}\right)^{\frac{b}{b+1}} & b < +\infty \quad c = 1 \\ \frac{1}{\ell} & b = +\infty \end{cases} \tag{20}$$

then

$$\lim_{\tau \to \infty} \limsup_{\ell \to \infty} \sup_{\rho \in \mathcal{P}(b,c)} \mathbb{P}_{\mathbf{z} \sim \rho^\ell} \left[\mathcal{E}[f_{\mathbf{z}}^{\lambda_\ell}] - \mathcal{E}[f_{\mathcal{H}}] > \tau a_\ell \right] = 0 \tag{21}$$

The bound being used

$\mathcal{K}$ is real separable Hilbert space. ξ is random variable on $\mathcal{K}$. Assume there exists positive constants L,σ such that

$$\mathbb{E}\left(\|\xi - \mathbb{E}\xi\|_{\mathcal{K}}^m\right) \leq \frac{1}{2} m! \sigma^2 L^{m-2} \qquad \forall m \geq 2.$$

Then

$$\mathbb{P}\left[\left\|\frac{1}{\ell}\sum_{i=1}^{\ell}\xi_i - \mathbb{E}\xi\right\|_{\mathcal{K}} \leq 2\left(\frac{L}{\ell} + \frac{\sigma}{\sqrt{\ell}}\right)\log\frac{2}{\eta}\right] \geq 1 - \eta.$$

True when:

$$\begin{array}{lcl} \|\xi(\omega)\|_{\mathcal{K}} & \leq & \frac{L}{2} \quad \text{a.}s \\ \mathbb{E}[\|\xi\|_{\mathcal{K}}^2] & \leq & \sigma^2. \end{array}$$

How is the proof done?

Define $f_{\mathcal{H}}$ as the argument of the infimum (i.e., assume it is attained). Then

$$\mathcal{E}\left[f_{\mathbf{z}}^{\lambda_\ell}\right] - \mathcal{E}[f_{\mathcal{H}}] \leq A(\lambda) + S_1(\lambda, \mathbf{z}) + S_2(\lambda, \mathbf{z})$$

where:

- $A(\lambda) := \mathcal{E}(f^\lambda) - \mathcal{E}(f_{\mathcal{H}})$, and f^λ is the population *regularized* solution ($A(\lambda)$ is bias term)
- $S_1(\lambda, \mathbf{z}) = \left\| \sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}(g_{\mathbf{z}} - T_{\mathbf{x}} f_{\mathcal{H}}) \right\|_{\mathcal{H}}^2$, converges via bound under noise assumption on $\rho(y|x)$.
- $S_2(\lambda, \mathbf{z}) = \left\| \sqrt{T}(T_{\mathbf{x}} + \lambda)^{-1}(T - T_{\mathbf{x}})(f^\lambda - f_{\mathcal{H}}) \right\|_{\mathcal{H}}^2$, converges since we can prove mean and variance requirement for bound.

Why "optimal"?

When $\mathcal{Y}$ is finite dimensional, the upper bound is matched by a minimax lower rate: the "best you can do":

$$\lim_{\tau\to 0}\liminf_{\ell\to+\infty}\inf_{f_\ell}\sup_{\rho\in\mathcal{P}}\mathbb{P}_{\mathbf{z}\sim\rho^\ell}\left[\mathcal{E}[f_{\mathbf{z}}^\ell]-\inf_{f\in\mathcal{H}}\mathcal{E}[f]>\tau a_\ell\right]>0,$$

Boundedness

Definition (Operator norm)

The operator norm of a linear operator $A : \mathcal{F} \to \mathcal{G}$ is defined as

$$\|A\| = \sup_{f \in \mathcal{F}} \frac{\|Af\|_{\mathcal{G}}}{\|f\|_{\mathcal{F}}}$$

If $\|A\| < \infty$, A is called a **bounded linear operator**.

Boundedness

Definition (Operator norm)

The operator norm of a linear operator $A : \mathcal{F} \to \mathcal{G}$ is defined as

$$\|A\| = \sup_{f \in \mathcal{F}} \frac{\|Af\|_{\mathcal{G}}}{\|f\|_{\mathcal{F}}}$$

If $\|A\| < \infty$, A is called a **bounded linear operator**.

$\|A\|$ is the smallest number λ such that the inequality $\|Af\|_{\mathcal{G}} \leq \lambda \|f\|_{\mathcal{F}}$ holds for *every* $f \in \mathcal{F}$.

Boundedness

Definition (Operator norm)

The operator norm of a linear operator $A : \mathcal{F} \to \mathcal{G}$ is defined as

$$\|A\| = \sup_{f \in \mathcal{F}} \frac{\|Af\|_{\mathcal{G}}}{\|f\|_{\mathcal{F}}}$$

If $\|A\| < \infty$, A is called a **bounded linear operator**.

$\|A\|$ is the smallest number λ such that the inequality $\|Af\|_{\mathcal{G}} \leq \lambda \|f\|_{\mathcal{F}}$ holds for *every* $f \in \mathcal{F}$.

bounded operator $\neq$ bounded function

Generalization of the parallelogram law

The following is a generalization of the parallelogram law:

$$\|x\|^2 + \|y\|^2 + \|z\|^2 + \|x+y+z\|^2 = \|x+y\|^2 + \|y+z\|^2 + \|z+x\|^2 \quad (1)$$

Then apply $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$ and the parallelogram relation for the remaining two norms on the left, to get

$$\|x\|^2 + \|y\|^2 + \|z\|^2 + \|x+y+z\|^2 \leq 4\|x\|^2 + 4\|y\|^2 + 4\|z\|^2$$

and hence

$$\|x+y+z\|^2 \leq 3\left(\|x\|^2 + \|y\|^2 + \|z\|^2\right).$$

Proof: generalization of parallelogram law

To now prove (1): start with the standard parallelogram identity,

$$\|x+y\|^2 - \|x\|^2 = \|y\|^2 + 2\langle x, y\rangle.$$

Then defining $x_4 = x_1$,

$$\begin{aligned}\sum_{i=1}^{3}\|x_i + x_{i+1}\|^2 - \sum_{i=1}^{3}\|x_i\|^2 &= \sum_{i=1}^{3}\|x_i\|^2 + \sum_{i=1}^{3}2\langle x_i, x_{i+1}\rangle \\ &= \left\|\sum_{i=1}^{3}x_i\right\|^2\end{aligned}$$