Pre-asymptotic Optimality of Cross-validation in Scattered Data Approximation

Felix Bartel Ralf Hielscher

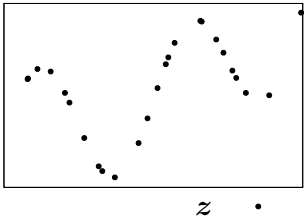
International Conference on Computational Harmonic Analysis

14th of September 2021

Mathematik!
TU Chemnitz

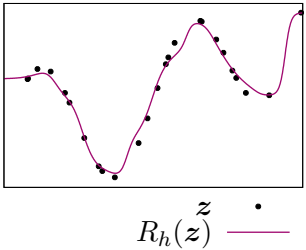
Setting: scattered data approximation

- data $\boldsymbol{z} = (\boldsymbol{x}_i, y_i)_{i=1}^n \in (\Omega \times Y)^n$
- $(\boldsymbol{x}_i, y_i)$ distributed according to ρ



Setting: scattered data approximation

- data $\boldsymbol{z} = (\boldsymbol{x}_i, y_i)_{i=1}^n \in (\Omega \times Y)^n$
- $(\boldsymbol{x}_i, y_i)$ distributed according to ρ
- reconstruction algorithm
  $R_h \colon (\Omega \times Y)^n \to Y^\Omega$



Motivation 1: estimate risk

Given: data $\boldsymbol{z}$, reconstruction model $R_h(\boldsymbol{z})$

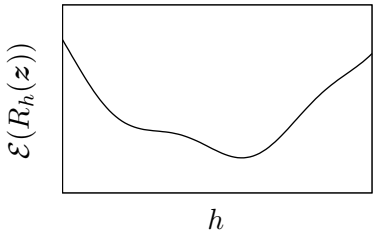
Question: How big is the L_2-reconstruction error of $R_h(\boldsymbol{z})$?

$$\mathcal{E}(R_h(\boldsymbol{z})) = \int_{\Omega \times Y} |(R_h(\boldsymbol{z}))(\boldsymbol{x}) - y|^2 \ \mathrm{d}\rho(\boldsymbol{x}, y)$$

Motivation 2: model selection

Given: data $\boldsymbol{z}$, hypothesis space $\{R_h(\boldsymbol{z}) : h \in H\}$

Question: Which reconstruction model $R_h(\boldsymbol{z})$ to choose?



Strategy

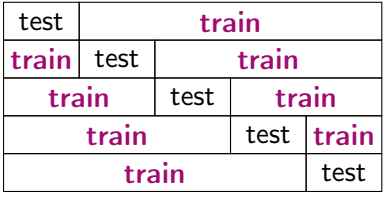
Purly data-driven method to approximate the risk

- validation set

Strategy

Purly data-driven method to approximate the risk

- validation set
- "Probably the simplest and most widly used method for estimation prediction error is cross-validation." [Hastie 2001]



Leave-one-out cross-validation [Stone '74]

1. calculate the reconstructions $R_h(\boldsymbol{z}_{-i})$ from the function values $\boldsymbol{z}_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$
2. evaluate the residual of $R_h(\boldsymbol{z}_{-i})$ in the i-th node $\boldsymbol{x}_i$

$$|R_h(\boldsymbol{z}_{-i})(\boldsymbol{x}_i) - y_i|$$

3. calculate the mean value with respect to all nodes

$$\mathrm{CV}(\boldsymbol{z}, \lambda) := \frac{1}{n} \sum_{i=1}^{n} |R_h(\boldsymbol{z}_{-i})(\boldsymbol{x}_i) - y_i|^2$$

Bound by Golub, Heath, and Whaba from '79

For PLSE with $1/n \operatorname{trace} \boldsymbol{H}_\lambda < 1$ they showed

$$\frac{|\mathbb{E}(\mathcal{E}(R_\lambda(\boldsymbol{Z})) - \mathbb{E}(\mathrm{CV}(\boldsymbol{Z}, \lambda))|}{\mathbb{E}(\mathrm{CV}(\boldsymbol{Z}, \lambda))}$$
$$\leq \left(\frac{2}{n} \operatorname{trace} \boldsymbol{H}_\lambda + \frac{(\operatorname{trace} \boldsymbol{H}_\lambda)^2}{\operatorname{trace} \boldsymbol{H}_\lambda^2} \right) \frac{1}{(1 - 1/n \operatorname{trace} \boldsymbol{H}_\lambda)^2}.$$

Bound by Li from '87

Let

$$h^+ = \arg\min_h \mathcal{E}(R_h(\boldsymbol{z})) \quad \text{and} \quad h^\star = \arg\min_h \mathrm{CV}(\boldsymbol{z}, h).$$

Then under certain assumptions

$$\frac{\mathcal{E}(R_{h^\star}(\boldsymbol{z}))}{\mathcal{E}(R_{h^+}(\boldsymbol{z}))} \to 1 \quad \text{for} \quad n \to \infty.$$

Theorem (Hoeffding '63)

Let

- $Z_1, \ldots, Z_n$ *independent rv's with values in* $[0,1]$ *and*
- $m = \mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n} Z_i\right\}$.

Then for $\varepsilon > 0$

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - m\right| > \varepsilon\right\} \leq 2\exp\left(-2n\varepsilon^2\right).$$

Concentration inequalities

A function $f\colon \Omega^n \to \mathbb{R}$ is **$\boldsymbol{c}$-bounded** on $\Xi \subset \Omega^n$ for $\boldsymbol{c} = (c_1, \ldots, c_n)$, iff

$$|f(z_1, \ldots, z_n) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \leq c_i$$

for all $(z_1, \ldots, z_n), (z_1', \ldots, z_n') \in \Xi$, and $1 \leq i \leq n$.

Theorem (McDiarmid '89)

Let

- $Z_1, \ldots, Z_n$ *be independent rv's with values in* Ω,
- $f\colon \Omega^n \to \mathbb{R}$ *be* $\boldsymbol{c}$*-bounded on* Ω^n, *and*
- $m = \mathbb{E}\{f(Z_1, \ldots, Z_n)\}$.

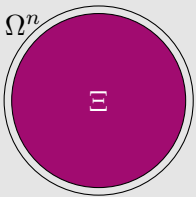
Then for $\varepsilon > 0$

$$\mathbb{P}\{|f(Z_1, \ldots, Z_n) - m| > \varepsilon\} \leq 2\exp\left(-\frac{2\varepsilon^2}{\|\boldsymbol{c}\|_2^2}\right).$$

Theorem (Extension of McDiarmid by Combes '15)

Let

- $Z_1, \dots, Z_n$ *be independent rv's with values in* Ω,
- $f\colon \Omega^n \to \mathbb{R}$ *be* $\boldsymbol{c}$*-bounded on* $\Xi \subset \Omega^n$, *and*
- $m = \mathbb{E}\{f(Z_1, \dots, Z_n) | (Z_1, \dots, Z_n) \in \Xi\}$, *and*
- $\gamma = 1 - \mathbb{P}\{(Z_1, \dots, Z_n) \in \Xi\}$.



Then for $\varepsilon > \gamma\|\boldsymbol{c}\|_1$

$$\mathbb{P}\left\{|f(Z_1, \dots, Z_n) - m| > \varepsilon\right\} \leq 2\gamma + 2\exp\left(-\frac{2(\varepsilon - \gamma\|\boldsymbol{c}\|_1)^2}{\|\boldsymbol{c}\|_2^2}\right).$$

General framework

Define $\Xi = \Xi(C_1, C_2)$ as the set of $\boldsymbol{z} = (z_1, \ldots, z_n)$ fulfilling

1. a uniform bound on the reconstruction error

$$\|R_h(\boldsymbol{z}_{-i}) - f\|_\infty \leq C_1$$

for $1 \leq i \leq n$.

General framework

Define $\Xi = \Xi(C_1, C_2)$ as the set of $\boldsymbol{z} = (z_1, \ldots, z_n)$ fulfilling

1. a uniform bound on the reconstruction error

$$\|R_h(\boldsymbol{z}_{-i}) - f\|_\infty \leq C_1$$

2. $\boldsymbol{z}_{-i} \mapsto R_h(\boldsymbol{z}_{-i})(\boldsymbol{x})$ $\boldsymbol{c}$-bounded for all $\boldsymbol{x} \in \Omega$ on Ξ with $\boldsymbol{c} = \mathbb{1}C_2$

for $1 \leq i \leq n$.

General framework

Risk functional $\boldsymbol{z} \mapsto \mathcal{E}(R_h(\boldsymbol{z}_{-i}))$ is $\boldsymbol{c}$-bounded on Ξ with $\boldsymbol{c} = 4C_1^2\mathbb{1}$.

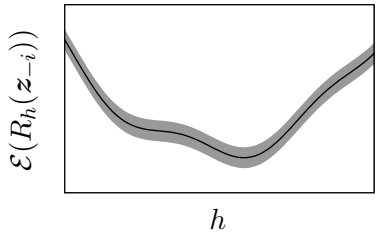
Cross-validation score $\boldsymbol{z} \mapsto \mathrm{CV}(\boldsymbol{z}, h)$ is $\boldsymbol{c}$-bounded on Ξ with $\boldsymbol{c} = 5C_1^2\mathbb{1}$.

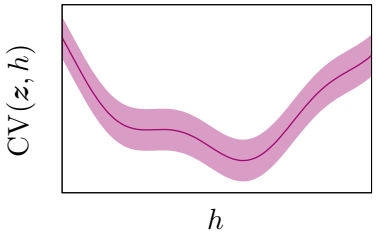
General framework

Risk functional $\boldsymbol{z} \mapsto \mathcal{E}(R_h(\boldsymbol{z}_{-i}))$ is $\boldsymbol{c}$-bounded on Ξ with $\boldsymbol{c} = 4C_1^2\mathbb{1}$.

Cross-validation score $\boldsymbol{z} \mapsto \mathrm{CV}(\boldsymbol{z}, h)$ is $\boldsymbol{c}$-bounded on Ξ with $\boldsymbol{c} = 5C_1^2\mathbb{1}$.

Now we apply Combes extension of McDiarmid.

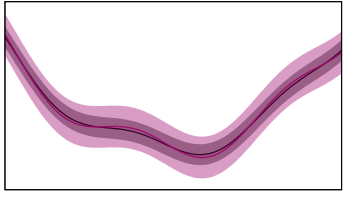




Lemma

For $\boldsymbol{Z}'$ representing $n-1$ samples and $\boldsymbol{Z}$ representing n samples element-wise distributed according to ρ we have

$$\mathbb{E}_{\boldsymbol{Z}'}\left\{\mathcal{E}(R_h(\boldsymbol{Z}'))\right\} = \mathbb{E}_{\boldsymbol{Z}}\left\{\mathrm{CV}(\boldsymbol{Z}, h)\right\}.$$



h

$\mathcal{E}(R_h(\boldsymbol{z}_{-i}))$
$\mathrm{CV}(\boldsymbol{z}, h)$

Theorem (B., Hielscher '21)

Let $n \geq 3$,

- $\boldsymbol{Z}$ *be element-wise distributed according to* ρ *with values in* $(\Omega \times Y)^n$, *and*
- $\gamma = 1 - \mathbb{P}\{\boldsymbol{Z} \in \Xi\}$.

Then for $\delta > 0$ *we have with probability larger than* $1 - 2(\gamma + \delta)$

$$
\begin{aligned}
&|\mathrm{CV}(\boldsymbol{Z}, h) - \mathcal{E}(R_h(\boldsymbol{Z}_{-1}))| \\
&\leq \max\left\{2\gamma(M + \|f\|_\infty)^2, 12\sqrt{n}C_1^2(\sqrt{2n}\gamma + \sqrt{-\log\delta})\right\}
\end{aligned}
$$

Shepard's model

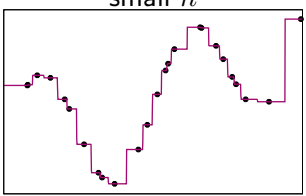
Shepard's model for $\Omega = \mathbb{T}$, $Y = \mathbb{R}$, and $\boldsymbol{z} = (x_i, f(x_i))_{i=1}^n$

$$R_h(\boldsymbol{z}) = \frac{\sum_{i=1}^m K_h(\cdot, x_i) f(x_i)}{\sum_{i=1}^m K_h(\cdot, x_i)}$$

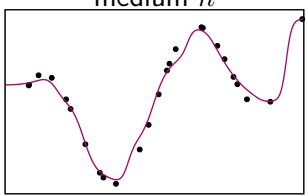
with

$$K_h(x_1, x_2) = \max\{0, 1 - h|x_1 - x_2|\}$$

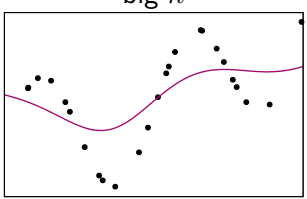
small h



medium h



big h



Shepard's model

We need estimates for

- $\|(R_h(\cdot))(\cdot)\|_\infty$,
- $\|R_h(\boldsymbol{z}) - f\|_\infty$ for $\boldsymbol{z} \in \Xi$, and
- $\gamma = 1 - \mathbb{P}\{\boldsymbol{Z} \in \Xi\}$.

Shepard's model

We need estimates for

- $\|(R_h(\cdot))(\cdot)\|_\infty \leq \|f\|_\infty$,
- $\|R_h(\boldsymbol{z}) - f\|_\infty$ for $\boldsymbol{z} \in \Xi$, and
- $\gamma = 1 - \mathbb{P}\{\boldsymbol{Z} \in \Xi\}$.

Shepard's model

We need estimates for

- $\|(R_h(\cdot))(\cdot)\|_\infty \leq \|f\|_\infty$,
- $\|R_h(\boldsymbol{z}) - f\|_\infty$ for $\boldsymbol{z} \in \Xi$, and
- $\gamma = 1 - \mathbb{P}\{\boldsymbol{Z} \in \Xi\}$.

Theorem (B., Hielscher '21)

Let $x_1, \ldots, x_n$ be uniformly random, $K_h(x, \cdot)$ be supported on $[x - 1/h, x + 1/h]$, and f be Lipschitz with constant L. Then $\|R_h(\boldsymbol{z}) - f\|_\infty \leq L/h$ with probability

$$1 - \gamma \geq \sum_{k=0}^{\lfloor h \rfloor} (-1)^{k+1} \binom{n}{k} \left(1 - \frac{k}{2h}\right)^{n-1} \approx 1 - \exp\left(-n \exp\left(-\frac{n}{h}\right)\right).$$

Theorem (B., Hielscher '21)

Let

- f *be Lipschitz with constant* L,
- $\boldsymbol{z} = (x_i, f(x_i))_{i=1}^n$ *be samples with* $x_1, \dots, x_n \in \mathbb{T}$ *uniformly distributed,*
- $K_h(x, \cdot)$ *be supported on* $[x - 1/h, x + 1/h]$, *and*
- $\gamma \approx \exp(-n \exp(-n/h))$.

Then we have for $\delta > 0$ *with probability larger than* $1 - 2(\gamma + \delta)$

$$
\begin{aligned}
&|\mathrm{CV}(\boldsymbol{Z}, h) - \mathcal{E}(R_h(\boldsymbol{Z}_{-1}))| \\
&\leq \max\left\{ 4\gamma \|f\|_\infty^2, \frac{12\sqrt{n} L^2}{h^2} (\sqrt{2n}\gamma + \sqrt{-\log \delta}) \right\}
\end{aligned}
$$

Györfi, Kohler, Krzyżak, and Walk in '02

Using

- binary kernels $K_h : \Omega \times \Omega \to \{0, 1\}$
- Ξ all possible samples

they showed

$$\gamma = 0, \quad C_1 = 2\|f\|_\infty, \quad C_2 \sim \frac{1}{n}$$

and obtained with probability $1 - 2\delta$

$$|\mathrm{CV}(\boldsymbol{Z}, h) - \mathcal{E}(R_h(\boldsymbol{Z}_{-1}))| \lesssim n^{-1/2}.$$

Györfi, Kohler, Krzyżak, and Walk in '02

Using

- binary kernels $K_h : \Omega \times \Omega \to \{0, 1\}$
- Ξ all possible samples

they showed

$$\gamma = 0, \quad C_1 = 2\|f\|_\infty, \quad C_2 \sim \frac{1}{n}$$

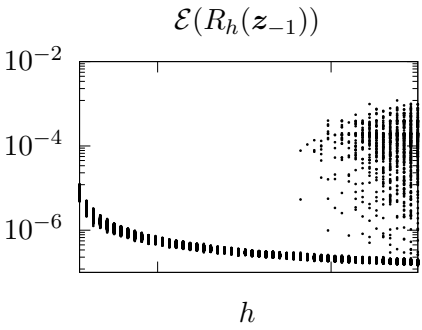
and obtained with probability $1 - 2\delta$

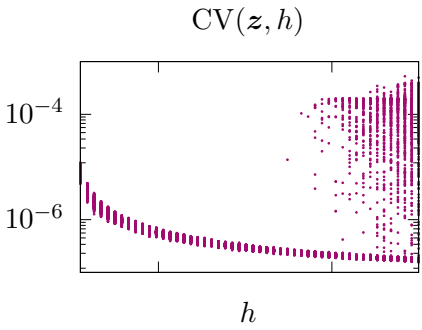
$$|\mathrm{CV}(\boldsymbol{Z}, h) - \mathcal{E}(R_h(\boldsymbol{Z}_{-1}))| \lesssim n^{-1/2}.$$

Assuming $h \sim n$, we obtained with probability $1 - 2(\delta + \exp(-n \exp(-n/h)))$

$$|\mathrm{CV}(\boldsymbol{Z}, h) - \mathcal{E}(R_h(\boldsymbol{Z}_{-1}))| \lesssim n^{-3/2}.$$

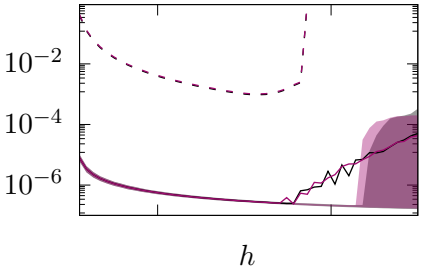
Numerical example





Numerical example

$\mathrm{CV}(\boldsymbol{z}, h)$ and $\mathcal{E}(R_h(\boldsymbol{z}_{-1}))$



$|\,\mathrm{CV}(\boldsymbol{z}, h) - \mathcal{E}(R_h(\boldsymbol{z}_{-1}))|$