The nonparametric least squares estimator under covariate shift

Johannes Schmidt-Hieber

talk is based on the paper

Bernoulli **30**(3), 2024, 1845–1877
https://doi.org/10.3150/23-BEJ1655

Local convergence rates of the nonparametric least squares estimator with applications to transfer learning

JOHANNES SCHMIDT-HIEBER[a] and PETR ZAMOLODTCHIKOV[b]

nonparametric least squares

statistical model: we observe n i.i.d. pairs $(X_1, Y_1), \ldots, (X_n, Y_n) \in [0,1] \times \mathbb{R}$, with

$$Y_i = f_0(X_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

nonparametric least squares for function class $\mathcal{F}$

$$\widehat{f}_n \in \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} \big(Y_i - f(X_i)\big)^2$$

covariate shift

statistical model: we observe n i.i.d. pairs $(X_1, Y_1), \ldots, (X_n, Y_n) \in [0,1] \times \mathbb{R}$, with

$$Y_i = f_0(X_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

denote the marginal distribution of the X_i by P_X

covariate shift: The distribution of the X_i might be different during test time,

$$P_X \to Q_X$$

notation: we assume that P_X and Q_X have densities that are denoted by p and q

LSE under covariate shift

how well does nonparametric least squares estimator perform under covariate shift?

- **statistical learning theory** yields bounds for the risk

$$\mathbb{E}\Big[\int_0^1 \big(\widehat{f}_n(x) - f_0(x)\big)^2 p(x)\,dx\Big]$$

- this is very natural, given that the training loss is

$$\sum_{i=1}^{n} \big(Y_i - f(X_i)\big)^2$$

- to understand behavior under covariate shift, we need bounds for

$$\mathbb{E}\Big[\int_0^1 \big(\widehat{f}_n(x) - f_0(x)\big)^2 q(x)\,dx\Big]$$

Lipschitz class

- here we study the LSE for the class Lip(1) consisting of all univariate Lipschitz functions

$$|f(x) - f(y)| \leq |x - y|.$$

- without covariate shift, the convergence rate of the prediction risk is $n^{-2/3}$ with n the sample size

Question: What is the rate of the prediction risk under the target density q,

$$\mathbb{E}\Big[\int_0^1 \big(\widehat{f}_n(x) - f_0(x)\big)^2 q(x)\, dx\Big].$$

motivation for Lip(1) is relation to ReLU networks and mathematical tractability

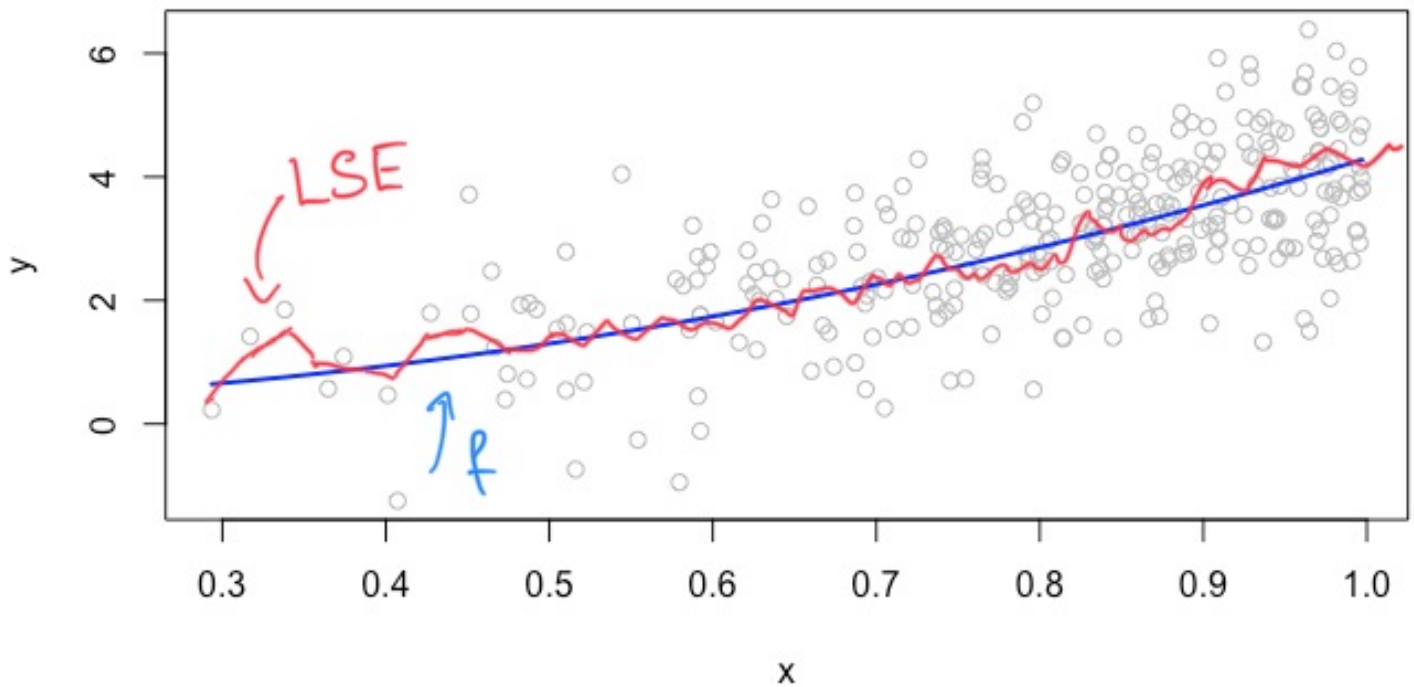
connection to NNs

Denote by $\mathrm{ReLU}_N(1)$ the function class of all shallow ReLU networks with N hidden nodes that are moreover 1-Lipschitz. If $N \geq n-1$, then,

$$\operatorname*{argmin}_{f \in \mathrm{ReLU}_N(1)} \sum_{i=1}^{n} \big(Y_i - f(X_i)\big)^2 \subseteq \operatorname*{argmin}_{f \in \mathrm{Lip}(1)} \sum_{i=1}^{n} \big(Y_i - f(X_i)\big)^2.$$

suggests that this could also describe behavior of neural network fits

local convergence rates



local convergence rate

- P_X source design distribution
- $t_n(x)$ unique solution of

$$t_n(x)^2 \, P_X\big(\big[x - t_n(x), x + t_n(x)\big]\big) = \frac{\log n}{n}$$

- assume doubling property: there exists a constant D such that for any x and any $\eta > 0$,

$$P_X(x - 2\eta, x + 2\eta) \leq D \, P_X(x - \eta, x + \eta)$$

Theorem: Let $\delta > 0$. If $\widehat{f}_n$ denotes the LSE taken over the class of 1-Lipschitz functions Lip(1), then, for a sufficiently large constant K,

$$\sup_{f_0 \in \mathrm{Lip}(1-\delta)} P_{f_0}\left(\sup_{x \in [0,1]} \frac{|\widehat{f}_n(x) - f_0(x)|}{t_n(x)} > K \right) \to 0 \quad \text{as } n \to \infty.$$

we also derived a corresponding minimax lower bound, proving that t_n is the optimal rate

standard statistical learning approach

If

$$\widehat{f}_n \in \operatorname*{argmin}_{f \in \mathrm{Lip}(1)} \sum_{i=1}^{n} \big(Y_i - f(X_i)\big)^2$$

and if true regression function $f_0 \in \mathrm{Lip}(1)$, then,

$$\frac{1}{n}\sum_{i=1}^{n} \big(\widehat{f}_n(X_i) - f_0(X_i)\big)^2 \leq \frac{2}{n}\sum_{i=1}^{n} \varepsilon_i \big(\widehat{f}_n(X_i) - f_0(X_i)\big).$$

Taking expectation, l.h.s. can be related to prediction error

$$\mathbb{E}\Big[\int_0^1 \big(\widehat{f}_n(x) - f_0(x)\big)^2 p(x)\, dx\Big]$$

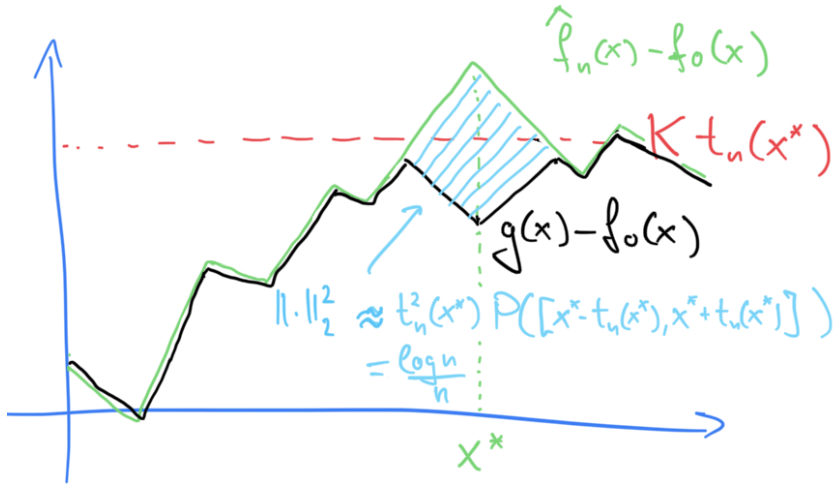
no possibility to extend this to derive local rates

high-level proof idea

- proof by contradiction
- assume there exists $0 \leq x^* \leq 1$ with

$$\widehat{f}_n(x^*) - f_0(x^*) > K t_n(x^*)$$

  for a large K
- construct a 1-Lipschitz function g as follows:



- contradicts the fact that $\widehat{f}_n$ is LSE

proving existence of such a g is hard

on the convergence rate

local convergence rate is given by equation

$$t_n(x)^2 \, P_X\big(\big[x - t_n(x), x + t_n(x)\big]\big) = \frac{\log n}{n}$$

with n the sample size

- higher density $\Leftrightarrow$ smaller local rate t_n
- if density is bounded away from zero,

$$t_n(x) = \left(\frac{\log n}{n}\right)^{1/3}$$

- very different behavior can occur if density is near zero
- to make this visible it makes sense to think of sequences $P_X \to P_X^n \Leftrightarrow p_n$

although LSE optimizes a global loss, it achieves the (optimal) local rate $t_n \rightsquigarrow$ no reweighting of loss necessary, cannot be obtained by kernel smoothing with fixed bandwidth

application to covariate shift

If q is the density under the target distribution, then w.h.p.,

$$\int_0^1 \left(\widehat{f}_n(x) - f_0(x)\right)^2 q(x)\,dx \le K^2 \int_0^1 t_n(x)^2 q(x)\,dx$$

Example: Let $\beta \in (0, 2]$. If $\inf_{x\in[0,1]} p_n(x) \ge n^{-\beta/(3+\beta)} \log n$, and p_n is β-smooth then

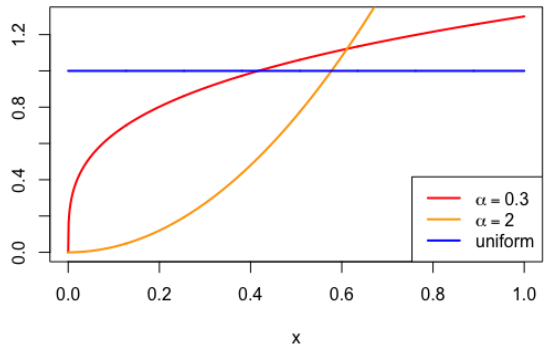
$$t_n(x)^2 \asymp \left(\frac{\log n}{n p_n(x)}\right)^{2/3},$$

and w.h.p.

$$\int_0^1 \left(\widehat{f}_n(x) - f_0(x)\right)^2 q(x)\,dx \lesssim \left(\frac{\log n}{n}\right)^{2/3} \int_0^1 \frac{q(x)}{p_n(x)^{2/3}}\,dx$$

decreasing densities

Let $\alpha > 0$. For source density $p(x) = (\alpha+1)x^\alpha$ and uniform target density q,



$$t_n(x) \asymp \left(\frac{\log n}{n}\right)^{1/(\alpha+3)} \wedge \left(\frac{\log n}{n x^\alpha}\right)^{1/3}$$

and

$$\int_0^1 \left(\widehat{f}_n(x) - f_0(x)\right)^2 q(x)\, dx \lesssim \left(\frac{\log n}{n}\right)^{3/(3+\alpha)} \vee \left(\frac{\log n}{n}\right)^{2/3}$$

change in the rate occurs for $\alpha = 3/2$

Pathak, Ma, Wainwright '22 show that for the Nadaraya-Watson estimator the change occurs already for $\alpha = 1$

covariate shift with two samples

suppose we also observe m iid samples from the target distribution

we observe $(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m}) \in [0,1] \times \mathbb{R}$ with

$$\begin{aligned} X_i &\sim P_X, && \text{for } i = 1, \ldots, n, \\ X_i &\sim Q_X, && \text{for } i = n+1, \ldots, n+m, \\ Y_i &= f_0(X_i) + \varepsilon_i, && \text{for } i = 1, \ldots, n+m, \end{aligned}$$

local rate associated to the subsamples

$$t_n^P(x)^2 \, P_X\big([x - t_n^P(x), x + t_n^P(x)]\big) = \frac{\log n}{n}$$

$$t_m^Q(x)^2 \, Q_X\big([x - t_m^Q(x), x + t_m^Q(x)]\big) = \frac{\log m}{m}$$

combined estimator

on the full sample we consider a **combined estimator** $\widehat{f}_{n,m}$ that copies the value of the LSE with the smaller local rate

$$\widehat{f}_{n,m}(x) := \widehat{f}_n^P(x)\mathbf{1}\left(\widehat{t}_n^P(x) \leq \widehat{t}_m^Q(x)\right) + \widehat{f}_m^Q(x)\mathbf{1}\left(\widehat{t}_n^P(x) > \widehat{t}_m^Q(x)\right)$$

with

- $\widehat{f}_n^P$ and $\widehat{f}_n^Q$ the LSEs over Lip(1) functions under the source and target subsample
- $\widehat{t}_n^P(x)$ and $\widehat{t}_n^Q(x)$ the empirical versions of the local rates (can be shown to be consistent estimators as $n, m \to \infty$)

all the information this estimator needs from the source sample is the LSE $\widehat{f}_n^P$ and some covariates $X_i \sim P_X$ to estimate $\widehat{t}_n^P(x)$

convergence rate

$$\widehat{f}_{n,m}(x) := \widehat{f}_n^P(x)\mathbf{1}\left(\widehat{t}_n^P(x) \le \widehat{t}_m^Q(x)\right) + \widehat{f}_m^Q(x)\mathbf{1}\left(\widehat{t}_n^P(x) > \widehat{t}_m^Q(x)\right)$$

Theorem: For a sufficiently large constant K,

$$\sup_{f_0 \in \mathrm{Lip}(1-\delta)} \mathbb{P}_{f_0}^{n,m}\left(\sup_{x\in[0,1]} \frac{|\widehat{f}_{n,m}(x) - f_0(x)|}{t_n^P(x) \wedge t_m^Q(x)} > K\right) \to 0 \quad \text{as } n, m \to \infty.$$

we also showed that the local convergence rate $t_n^P(x) \wedge t_m^Q(x)$ is **minimax optimal** in this model

- shows that additional sample can only improve the convergence rate if $t_m^Q(x) \ll t_n^P(x)$ somewhere

example

(everything up to log-terms)

- $p(x) = (\alpha + 1)x^{\alpha+1}$,
- Q uniform distribution
- $\alpha > 3/2$ (regime where LSE rate is $\ll n^{-2/3}$)
- if $n^{3/(3+\alpha)} \ll m \leq n$, then, w.h.p.

$$\int_0^1 \left(\widehat{f}_{n,m}(x) - f_0(x)\right)^2 q(x)\, dx \lesssim m^{-2/3}\left(\frac{m}{n}\right)^{1/\alpha}$$

- if $m \asymp n$, the rate is

$$\lesssim n^{-2/3}$$

- if $m \lesssim n^{3/(3+\alpha)}$, the rate is

$$\lesssim n^{-3/(3+\alpha)}$$

and the target sample does not improve the rate

open problems

a) **arbitrary smoothness β and covariate/input dimension d:**
local rate should be determined by equation

$$t_n(x)^2 \mathbb{P}_X\big(y : |x-y|_\infty \leq t_n(x)^{1/\beta}\big) = \frac{\log n}{n}$$

issues:

- construction seems difficult to generalize to $\beta > 1$, as for two functions also the pointwise maximum (and minimum) need to be contained in the class
- properties of the LSE for $\beta < d/2$?

b) **gradient descent iterates:**
what can we say about local convergence of say shallow neural networks?

c) $\boldsymbol{\delta = 0}$ **?**

We take the LSE over all 1-Lipschitz functions, but assume that the true regression function is at mot $(1 - \delta)$-Lipschitz for $\delta > 0$. What happens for $\delta = 0$?

conclusion

- least squares estimator (LSE) taken over Lipschitz functions minimizes global criterion, but achieves the optimal local rate $t_n(x)$
- proof technique is non-standard
- convergence result can be used to derive rates for the generalization error of the LSE under covariate shift
- if we also have m observations from the target distribution, we constructed an estimator based on the individual LSEs that achieves the optimal rate $t_n^P(x) \wedge t_m^Q(x)$
- many extensions possible

more details in the paper