A New Interpolative Perspective on Shape-restricted Estimation

Kenta Takatsu, Carnegie Mellon University
Joint work with Tianyu Zhang and Arun Kuchibhotla
WNAR June, 2025

From Isotonic to Lipschitz Regression: A New Interpolative Perspective on Shape-restricted Estimation

Kenta Takatsu, Tianyu Zhang, and Arun Kumar Kuchibhotla

Department of Statistics and Data Science, Carnegie Mellon University

Abstract

This manuscript seeks to bridge two seemingly disjoint paradigms of nonparametric regression: estimation based on smoothness assumptions and shape constraints. The proposed approach is motivated by a conceptually simple observation: Every Lipschitz function is a sum of monotonic and linear functions. This principle is further generalized to the higher-order monotonicity and multivariate covariates. A family of estimators is proposed based on a sample-splitting procedure, which inherits desirable methodological, theoretical, and computational properties of shape-restricted estimators. The theoretical analysis provides convergence guarantees of the estimator under heteroscedastic and heavy-tailed errors, as well as adaptivity properties to the unknown complexity of the true regression function. The generality of the proposed decomposition framework is demonstrated through new approximation results, and extensive numerical studies validate the theoretical properties of the proposed estimation framework.

Keywords— Nonparametric regression, Model selection, Shape-restricted estimation, Constructive Approximation, Heavy-tailed data

arXiv:2307.05732

Nonparametric Regression

Observe IID real-valued R.V. $(X_1, Y_1), \ldots, (X_n, Y_n) \in [0,1] \times \mathbb{R}$ such that

$$Y_i = f_0(X_i) + \varepsilon_i \quad \text{where} \quad f_0(x) := \mathbb{E}[Y \mid X = x].$$

Nonparametric Regression

Observe IID real-valued R.V. $(X_1, Y_1), \ldots, (X_n, Y_n) \in [0,1] \times \mathbb{R}$ such that

$$Y_i = f_0(X_i) + \varepsilon_i \quad \text{where} \quad f_0(x) := \mathbb{E}[Y \mid X = x].$$

Choice 1 Assume that f_0 is smooth (e.g., once-differentiable).

Choice 2 Assume that f_0 has some shape (e.g., monotonic).

Nonparametric Regression

Observe IID real-valued R.V. $(X_1, Y_1), \ldots, (X_n, Y_n) \in [0,1] \times \mathbb{R}$ such that

$$Y_i = f_0(X_i) + \varepsilon_i \quad \text{where} \quad f_0(x) := \mathbb{E}[Y \mid X = x].$$

Choice 1 Assume that f_0 is smooth (e.g., once-differentiable).

Choice 2 Assume that f_0 has some shape (e.g., monotonic).

Without additional structure, the minimax rates are $\|\hat{f} - f_0\|^2 = O_P(n^{-2/3})$ for both choices.

Nonparametric Regression

Observe IID real-valued R.V. $(X_1, Y_1), \ldots, (X_n, Y_n) \in [0,1] \times \mathbb{R}$ such that

$$Y_i = f_0(X_i) + \varepsilon_i \quad \text{where} \quad f_0(x) := \mathbb{E}[Y|X = x].$$

Choice 1 Assume that f_0 is smooth (e.g., once-differentiable).

Choice 2 Assume that f_0 has some shape (e.g., monotonic).

Without additional structure, the minimax rates are $\|\hat{f} - f_0\|^2 = O_P(n^{-2/3})$ for both choices.

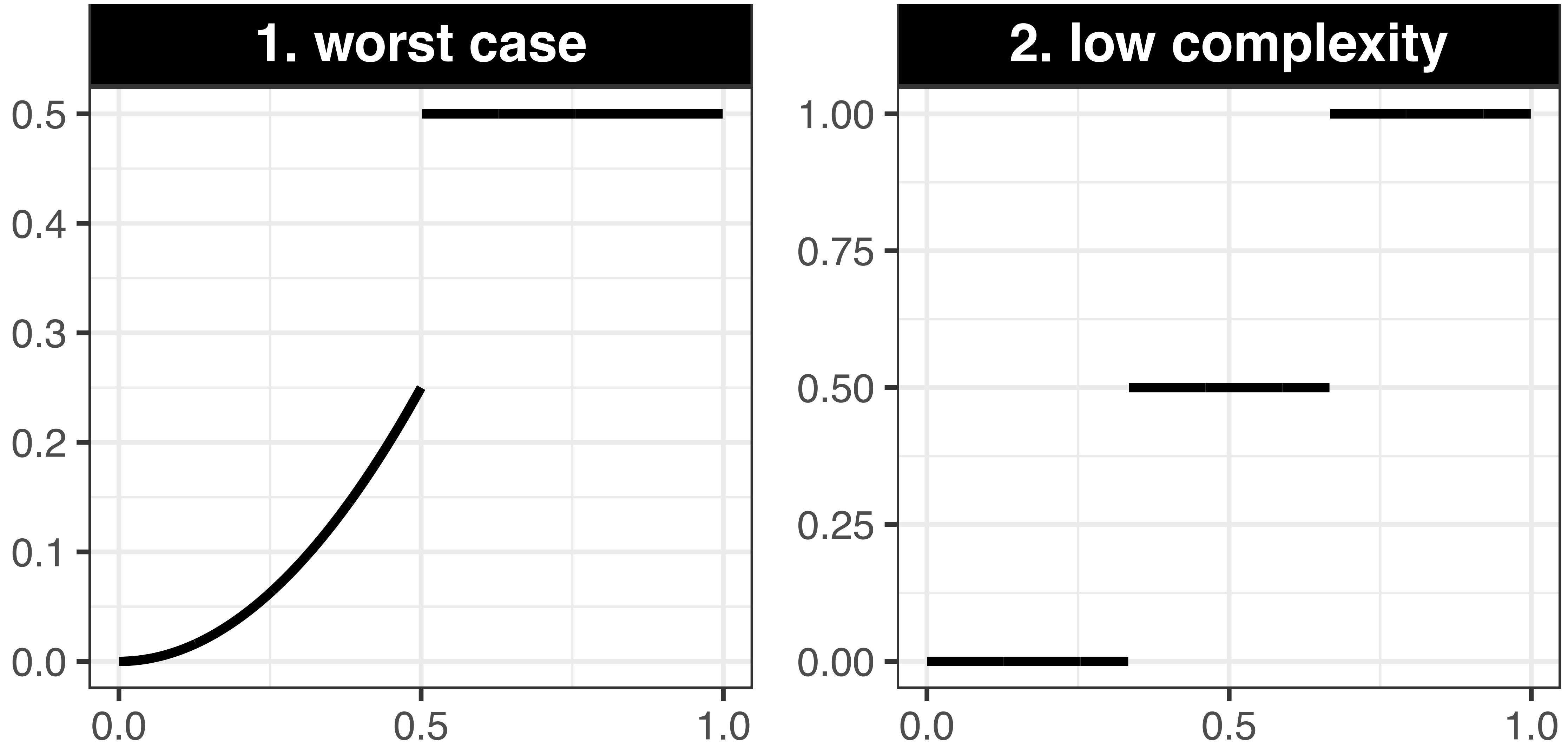
Q. Are there connections between the spaces of monotonic and once-differentiable functions?

Properties of Monotone Estimator

1. Many shape-restricted estimators (e.g., LSEs) are tuning parameter free.

2. They converge at *adaptive rates*;

$$\|\hat{f} - f_0\|^2 = O_P(n^{-2/3}) \qquad \|\hat{f} - f_0\|^2 = O_P(m/n)$$



Q. Can we identify structural links between smooth and monotone functions?

Q. Can we construct “tuning parameter free” and adaptive estimators beyond shape-restricted problems?

Q. Can we develop an optimal estimator for both smooth and shape-restricted classes simultaneously?

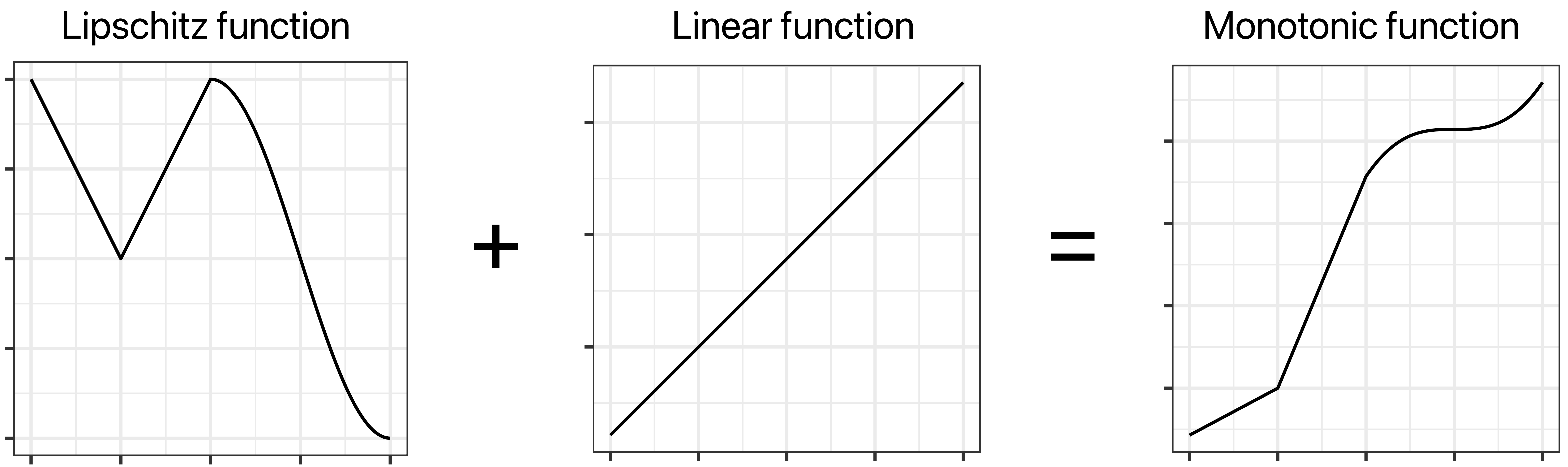
For any L-Lipschitz function f, there exists a non-decreasing function g such that:
$f(x) = g(x) - L'x$ where $L' \geq L$.

Proposition 3.1

For any L-Lipschitz function f, there exists a non-decreasing function g such that:

$$f(x) = g(x) - L'x \text{ where } L' \geq L.$$

Proposition 3.1



For any L-Lipschitz function f, there exists a non-decreasing function g such that:
$f(x) = g(x) - L'x$ where $L' \geq L$.

Proposition 3.1

Adding a linear term to a monotone estimator significantly increases its *expressiveness*.

We introduce a new decomposition class:

$$\mathcal{F}_1(L) := \left\{ f : [0,1] \mapsto \mathbb{R} \ : \ f(x) = g(x) - Lx, \exists g \text{ is non-decr} \right\}.$$

The class $\mathcal{F}_1(0)$ is equivalent to the space of monotone functions. As $k \uparrow L$, the space $\mathcal{F}_1(k)$ becomes larger and eventually includes all L-Lipschitz functions (and more).

For any L-Lipschitz function f, there exists a non-decreasing function g such that:
$f(x) = g(x) - L'x$ where $L' \geq L$.

Proposition 3.1

A Lipschitz estimator can take the form: $\hat{f}(x) = \hat{g}(x) - \hat{L}x$.

Lipschitz est. $\Rightarrow$ Monotone est. $+$ Model selection

For any L-Lipschitz function f, there exists a non-decreasing function g such that:
$f(x) = g(x) - L'x$ where $L' \geq L$.

Proposition 3.1

A Lipschitz estimator can take the form: $\hat{f}(x) = \hat{g}(x) - \hat{L}x$.

Lipschitz est. $\Rightarrow$ Monotone est. $+$ Model selection

Use sample-splitting to avoid overfitting.

This leads to two stages of least-squares.

Estimation Procedure

1. Split data into two parts D_1 and D_2. Prepare a set $\mathscr{L}$ for candidate L (e.g., $\mathscr{L} = [-\log n, \log n]$).

In practice, you do not need to specify $\mathscr{L}$.

The optimization program (e.g., optim($\cdot$) in R) will handle this.

https://github.com/Kenta426/npparam

Estimation Procedure

1. Split data into two parts D_1 and D_2. Prepare a set $\mathscr{L}$ for candidate L (e.g., $\mathscr{L} = [-\log n, \log n]$).

2. For each $L \in \mathscr{L}$, compute isotonic regression (monotone LSE) on D_1:

$$\widehat{g}_L := \underset{g:\text{monotone}}{\arg\min} \sum_{(x,y)\in D_1} ((y + Lx) - g(x))^2 \text{ and return } \hat{f}_L(x) := \widehat{g}_L(x) - Lx.$$

Step 2 can be computed using near linear time algorithm (e.g., isoreg($\cdot$) in R).

https://github.com/Kenta426/npparam

Estimation Procedure

1. Split data into two parts D_1 and D_2. Prepare a set $\mathscr{L}$ for candidate L (e.g., $\mathscr{L} = [-\log n, \log n]$).

2. For each $L \in \mathscr{L}$, compute isotonic regression (monotone LSE) on D_1:

$$\hat{g}_L := \arg\min_{g:\text{monotone}} \sum_{(x,y)\in D_1} ((y + Lx) - g(x))^2 \text{ and return } \hat{f}_L(x) := \hat{g}_L(x) - Lx.$$

3. Perform model selection over $\mathscr{L}$ on D_2, for instance,

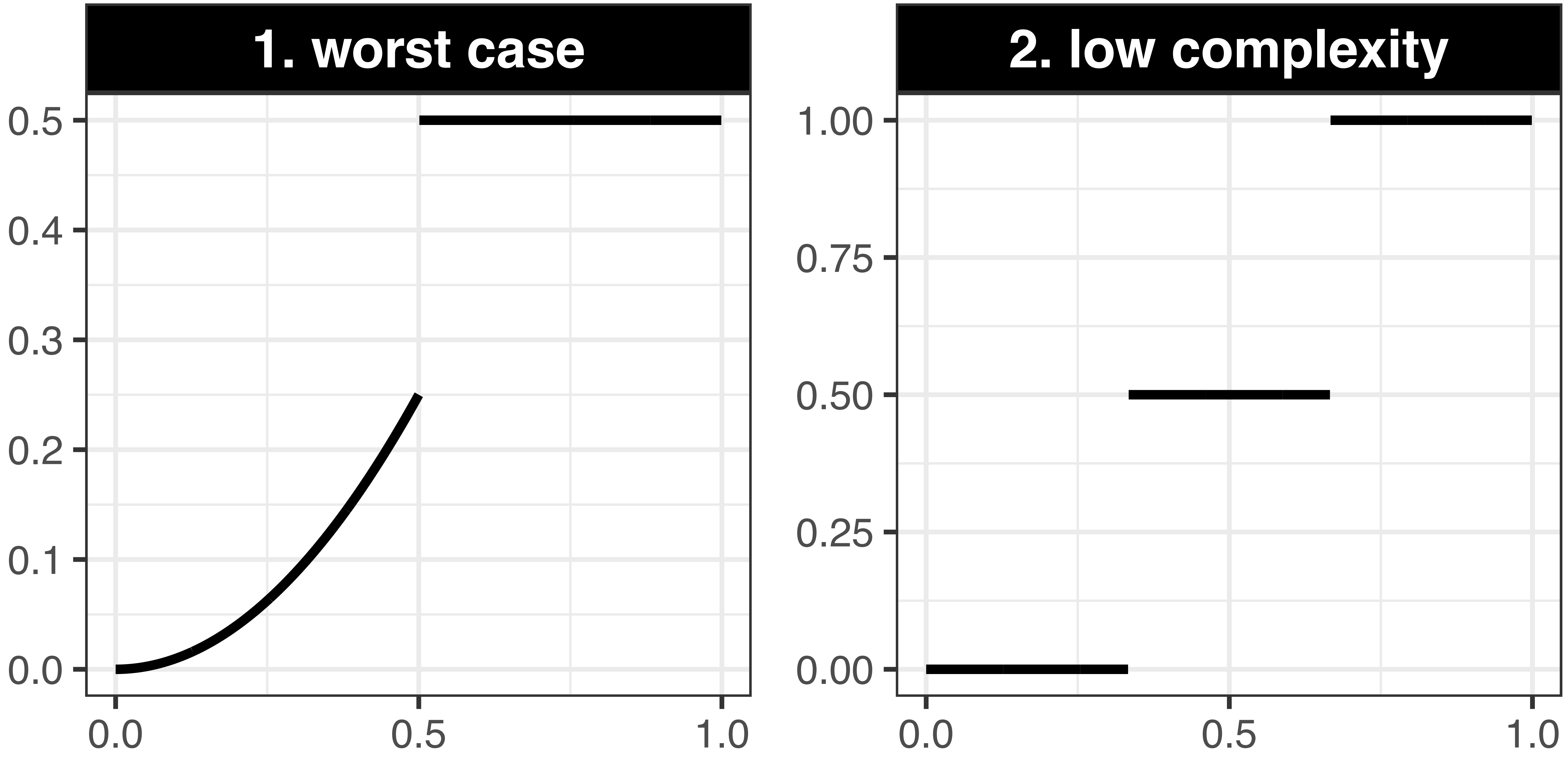
$$\hat{L} := \arg\min_{L\in\mathscr{L}} \sum_{(x,y)\in D_2} (y - \hat{f}_L(x))^2 \text{ and return } \hat{f}(x) := \hat{g}_{\hat{L}}(x) - \hat{L}x.$$

One can use other “robust” cross-validation using the median-of-means or aggregation.

https://github.com/Kenta426/npparam

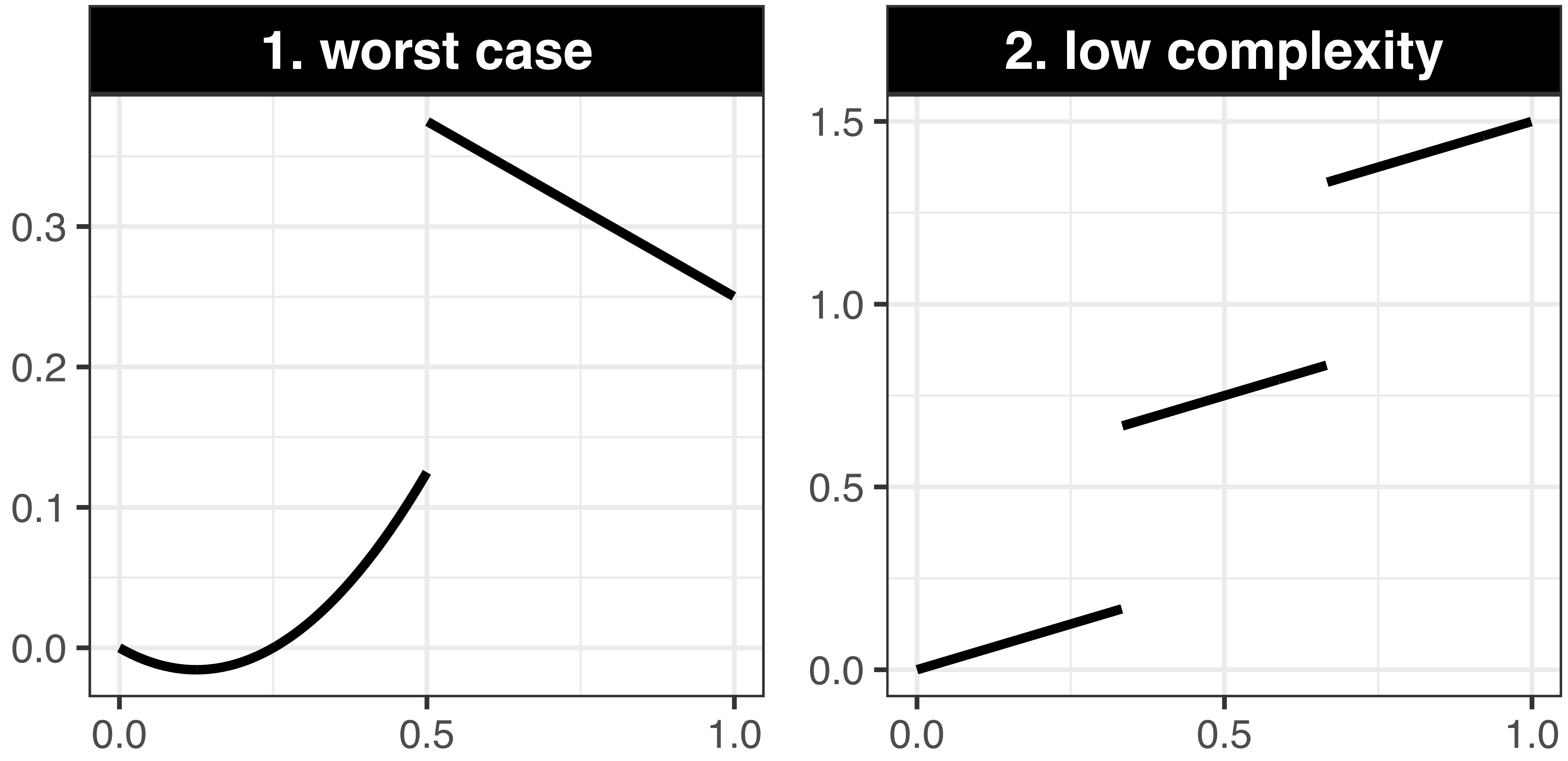
Convergence Rates of the Estimator

We expect $O_P(n^{-2/3})$ We expect $O_P(m/n)$



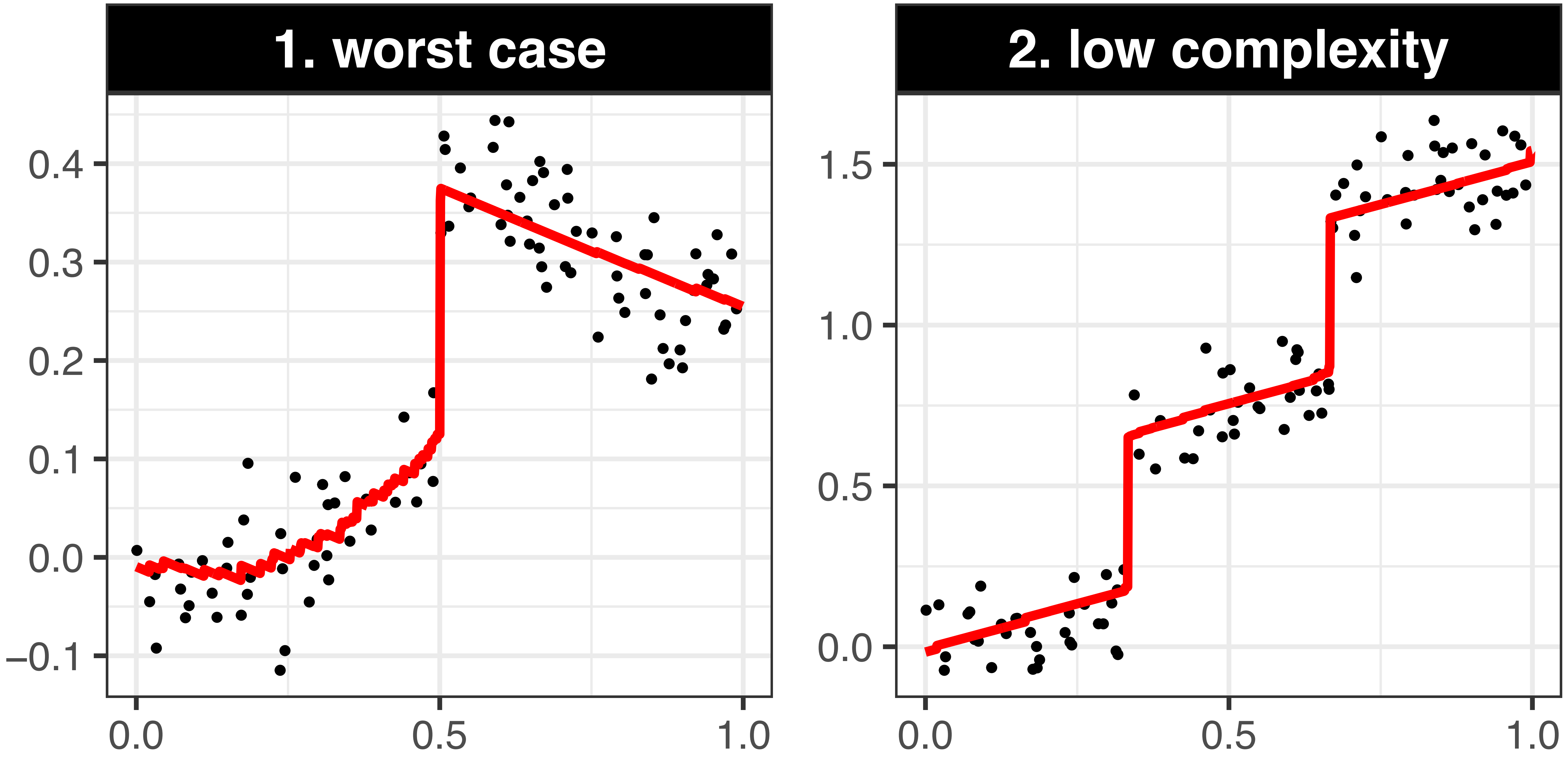
Convergence Rates of the Estimator

We expect $O_P(n^{-2/3})$ We expect $O_P(m/n)$



Convergence Rates of the Estimator

We expect $O_P(n^{-2/3})$ We expect $O_P(m/n)$



Convergence Rates of the Estimator

Lipschitz est. $\Rightarrow$ **Monotone est.** $+$ **Model Selection**

For each fixed $L \in \mathscr{L}$, fitting $y + Lx \sim x$ using monotone regression $\hat{g}_L$ is (almost) identical to the standard estimation for shape-restricted problems.

Selecting the "best" $L \in \mathscr{L}$ from a set $\{L \mapsto \hat{g}_L(x) - Lx\}$ is essentially a model selection performed over a nonparametric function space.

$$\|\hat{f} - f_0\|^2 = O_P\Big(\inf_{L \in \mathscr{L}} R^{\mathrm{mono}}_{n,L} + R^{\mathrm{CV}}_{n,\mathscr{L}} \Big)$$

Convergence Rates of the Estimator

Recall $Y_i = f_0(X_i) + \varepsilon_i$

A1 ε_i has finite q moments conditioning on X_i i.e., $\mathbb{E}[\,|\,\varepsilon_i|^q\,|\,X_i] \leq C$.

A2 ε_i is β-sub-Weibull conditioning on X_i ($\beta = 1$ corresponds to sub-exponential and $\beta = 2$ corresponds to sub-Gaussian).

Theorem 4.1-4.3

	Worst case	Low complexity
A1		
A2		

$R_{n,L}^{\text{mono}} + R_{n,L}^{\text{CV}}$

Convergence Rates of the Estimator

Recall $Y_i = f_0(X_i) + \varepsilon_i$

A1 ε_i has finite q moments conditioning on X_i i.e., $\mathbb{E}[\,|\,\varepsilon_i|^q\,|\,X_i] \leq C$.

A2 ε_i is β-sub-Weibull conditioning on X_i ($\beta = 1$ corresponds to sub-exponential and $\beta = 2$ corresponds to sub-Gaussian).

Theorem 4.1-4.3

	Worst case	Low complexity
A1		
A2	$O(n^{-2/3}) + O(n^{-1})$	$O(m/n) + O(n^{-1})$

$R^{\text{mono}}_{n,L} + R^{\text{CV}}_{n,L}$

Convergence Rates of the Estimator

Recall $Y_i = f_0(X_i) + \varepsilon_i$

A1 ε_i has finite q moments conditioning on X_i i.e., $\mathbb{E}[\,|\,\varepsilon_i|^q\,|\,X_i] \leq C$.

A2 ε_i is β-sub-Weibull conditioning on X_i ($\beta = 1$ corresponds to sub-exponential and $\beta = 2$ corresponds to sub-Gaussian).

Theorem 4.1-4.3

	Worst case	Low complexity
A1	$O(n^{-2/3}) + O(n^{-1+1/q})$	$O(m/n) + O(n^{-1+1/q})$
A2	$O(n^{-2/3}) + O(n^{-1})$	$O(m/n) + O(n^{-1})$

$R_{n,L}^{\text{mono}} + R_{n,L}^{\text{CV}}$

Convergence Rates of the Estimator

Recall $Y_i = f_0(X_i) + \varepsilon_i$

A1 ε_i has finite q moments conditioning on X_i i.e., $\mathbb{E}[\,|\,\varepsilon_i\,|^q\,|\,X_i] \leq C$.

A2 ε_i is β-sub-Weibull conditioning on X_i ($\beta = 1$ corresponds to sub-exponential and $\beta = 2$ corresponds to sub-Gaussian).

Theorem 4.1-4.3

It is not minimax when $q < 3$

	Worst case	Low complexity
A1	$O(n^{-2/3}) + O(n^{-1+1/q})$	$O(m/n) + O(n^{-1+1/q})$
A2	$O(n^{-2/3}) + O(n^{-1})$	$O(m/n) + O(n^{-1})$

$R^{\text{mono}}_{n,L} + R^{\text{CV}}_{n,L}$

Convergence Rates of the Estimator

It is not minimax when $q < 3$

	Worst case	Low complexity
A1	$O(n^{-2/3}) + O(n^{-1+1/q})$	$O(m/n) + O(n^{-1+1/q})$

Lipschitz est. $\Rightarrow$ Monotone est. $+$ Model selection

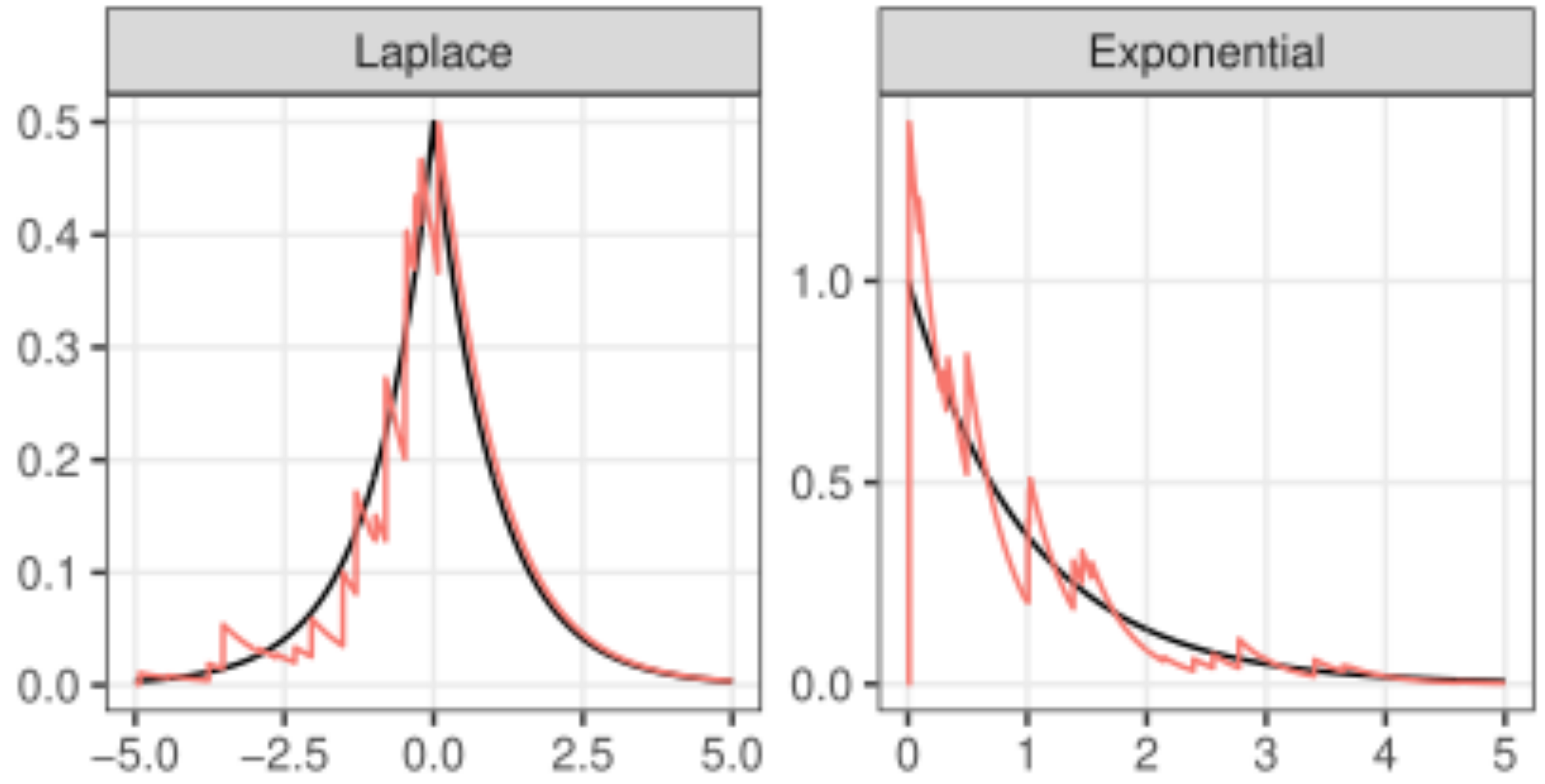
Under finite q moments, cross-validation becomes slower than estimation. This is because cross-validation based on squared error is *suboptimal* for heavy-tailed problems. We show that model selection based on median-of-means achieves the optimal rate under both **A1** and **A2**.

Open Problem: Irregular Density Estimation

Observe $X_1, \ldots, X_n$ from a density function f_0 such that $\log f_0(x) = g(x) - Lx$ where g is monotone.

This problem can be considered as *log Lipschitz estimation* in compaction to log concave estimation.

The example of the worst case is Laplace distribution and the adaptive case is exponential distribution.



Open Problem: Shape Detection

Suppose that the estimator is obtained by the proposed method:

$$\hat{f}(x) = \widehat{g}(x) - \widehat{L}x.$$

Intuitively, $|\widehat{L}| \approx 0$ should indicate that g_0 is monotone while $|\widehat{L}| \gg 0$ can be the evidence against monotonicity.

(Caveat) The population L_0 is not identifiable.

Summary

A nonparametric regression can be decomposed into two problems: (1) shape-restricted estimation and (2) cross-validation.

We propose a new nonparametric estimator that converges at adaptive rates under the some assumption on the error ε_i (e.g., sub-Weibull).

When the error ε_i has finite q moments, cross-validation can be “harder” than estimation. This cannot be improved without considering an alternative model selection procedure.

In the manuscript, we discuss some “higher-order” (e.g., convex + quadratic) and multivariate extensions.

Higher-order Methods

Lipschitz fn. $\Rightarrow$ Monotone fn. $+$ Linear fn.

Higher-order Methods

Lipschitz fn. $\Rightarrow$ Monotone fn. $+$ Linear fn.

Lipschitz deriv. $\Rightarrow$ Convex fn. $+$ Quadratic fn.

Thank you

From Isotonic to Lipschitz Regression:
A New Interpolative Perspective on Shape-restricted Estimation

Kenta Takatsu, Tianyu Zhang, and Arun Kumar Kuchibhotla

Department of Statistics and Data Science, Carnegie Mellon University

Abstract

This manuscript seeks to bridge two seemingly disjoint paradigms of nonparametric regression: estimation based on smoothness assumptions and shape constraints. The proposed approach is motivated by a conceptually simple observation: Every Lipschitz function is a sum of monotonic and linear functions. This principle is further generalized to the higher-order monotonicity and multivariate covariates. A family of estimators is proposed based on a sample-splitting procedure, which inherits desirable methodological, theoretical, and computational properties of shape-restricted estimators. The theoretical analysis provides convergence guarantees of the estimator under heteroscedastic and heavy-tailed errors, as well as adaptivity properties to the unknown complexity of the true regression function. The generality of the proposed decomposition framework is demonstrated through new approximation results, and extensive numerical studies validate the theoretical properties of the proposed estimation framework.

Keywords— Nonparametric regression, Model selection, Shape-restricted estimation, Constructive Approximation, Heavy-tailed data

arXiv:2307.05732

Appendix

Oracle Inequality for Model Selection

We develop a new general oracle inequality regarding an LSE:

Let $\hat{f}$ be an LSE over an arbitrary uniformly bounded and finite class $\mathscr{F}$, and ε_i has finite conditional q moments, then for any $\delta > 0$

$$\mathbb{E}\|\hat{f} - f_0\|^2 \leq (1+\delta) \inf_{f \in \mathscr{F}} \|f - f_0\|^2 + \left(n^{-1+1/q} + \frac{1}{\delta n}\right) \log(|\mathscr{F}|)$$

Theorem 4.4

The issue is the term $n^{-1+1/q}$, and this is *not generally improvable* for LSEs (Han and Wellner (2018); Kuchibhotla and Patra (2022)).

Han, Q and J. A. Wellner. (2018). Robustness of shape-restricted regression estimators: An envelop perspective, arXiv preprint

Kuchibhotla, A. K. and R. K. Patra. (2022). On least squares estimation under heteroscedastic and heavy-tailed errors. The Annals of Statistics, 50(1):277-302.

Chatterjee, S., A. Guntuboyina, and B. Sen (2015). On risk bounds in isotonic and other shape restricted regression problems. The Annals of Statistics, 43(4):1774–1800.

Smirnov, N. V. (1952). Limit distributions for the terms of a variational series. *American Mathematical Society Translations, 6.*

Kim, I. and Ramdas, A. (2024). Dimension-agnostic inference using cross u-statistics. *Bernoulli*, 30(1): 683-711

Vogel, S. (2008b). Universal confidence sets for solutions of optimization problems. *SIAM Journal on Optimization, 19(3):1467–1488.*

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics, 18(1):191–219.*

Mourtada, J. (2022). Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics, 50(4):2157–2178.*

Thank You

Takatsu, K. Zhang, T., and Kuchibhotla, A. K. (2024). From Isotonic to Lipschitz Regression: A New Interpolative Perspective on Shape-restricted Estimation, arXiv:2307.05732.

Takatsu, K. and Kuchibhotla, A. K. (2025). Bridging Root-n and Non-standard Asymptotics: Dimension-agnostic Adaptive Inference in M-Estimation, arXiv:2501.07772.

Takatsu, K. and Kuchibhotla, A. K. (2024). Generalized van Trees Inequality: Local Minimax Bounds for Non-differentiable Functionals and Irregular Statistical Models. arXiv:2405.06437.

Convergence Rates of the Estimator

Recall $Y_i = f_0(X_i) + \varepsilon_i$

(A1) ε_i has finite q moments conditioning on X_i i.e., $\mathbb{E}[\,|\,\varepsilon_i\,|^q\,|\,X_i] \leq C$

(A2) ε_i is β-sub-Weibull conditioning on X_i ($\beta = 1$ corresponds to sub-exponential and $\beta = 2$ corresponds to sub-Gaussian)

Convergence Rates of the Estimator

Recall $Y_i = f_0(X_i) + \varepsilon_i$

(A1) ε_i has finite q moments conditioning on X_i i.e., $\mathbb{E}[\,|\,\varepsilon_i\,|^q\,|\,X_i] \leq C$

(A2) ε_i is β-sub-Weibull conditioning on X_i ($\beta = 1$ corresponds to sub-exponential and $\beta = 2$ corresponds to sub-Gaussian)

Recall $\mathscr{F}_1(L) := \left\{f : [0,1] \mapsto \mathbb{R} \;:\; g(x) - Lx, \exists g \text{ is non-decr}\right\}$ and let $\mathscr{L}$ be a finite set of linear parameter L

Theorem 4.1-4.3 (informal)

The following “two-layer“ oracle inequality holds for any $\varepsilon \in (0,1)$

$$\|\hat{f} - f_0\|^2 \leq \inf_{L \in \mathscr{L}} \left\{ C \inf_{f \in \mathscr{F}_1(L)} \|f - f_0\|^2 + \left(R^{\text{mono}}_{n,L} + \log(\,|\,\mathscr{L}\,|\,) R^{\text{CV}}_{n,L}\right) \right\}$$

For any L-Lipschitz function f, there exists a non-decreasing function g such that:
$f(x) = g(x) - L'x$ where $L' \geq L$.

Proposition 3.1

Define L-Lipschitz class as:

$$\Sigma_1(L) := \left\{ f : [0,1] \mapsto \mathbb{R} \ : \ \frac{|f(x_1) - f(x_2)|}{|x_1 - x_2|} \leq L, \forall x_1, x_2 \in [0,1] \right\}.$$

We introduce a new decomposition class:

$$\mathscr{F}_1(L) := \{ f : [0,1] \mapsto \mathbb{R} \ : \ f(x) = g(x) - Lx, \exists g \text{ is non-decr} \}.$$

We can show that $\Sigma_1(L) \subsetneq \mathscr{F}_1(L)$ and $\Sigma_1(L)$ is not dense in $\mathscr{F}_1(L)$ for fixed L.

New Function Spaces

For an integer $r \geq 1$ and h satisfying $h \in [0,1-rh]$, we define a forward operator:

$$\Delta_h^r(f,x) := \sum_{m=0}^{r} \binom{r}{m} (-1)^{r-m} f(x+mh).$$

Let $\mathscr{C}(k)$ be the collection of k-monotone functions (Chatterjee et al., 2015):

$$\mathscr{C}(k) := \left\{ g : [0,1] \to \mathbb{R} \ : \ \Delta_h^k(g,x) \geq 0 \text{ for all } x \in [0,1] \right\}$$

Examples: $\mathscr{C}(1) :=$ {non-decr. fns.} and $\mathscr{C}(2) :=$ {convex fns.}

We introduce a new function space:

$$\mathscr{F}(k,L) := \left\{ g(x) - (L/k!)x^k \text{ such that } g \in \mathscr{C}(k) \right\}.$$

Properties of the Estimator

Lipschitz est. $\Rightarrow$ Monotone est. $+$ Cross Validation

(Intuition) When monotone estimation is *harder* than cross-validation, the convergence rate of the proposed estimator should match that of a monotone estimator.

We introduce a new function space:

$$\mathscr{F}(k, L) := \left\{ g(x) - (L/k!)x^k \text{ such that } g \in \mathscr{C}(k) \right\}.$$

Define kth bounded L-Lipschitz class for an integer k as:

$$\Sigma_k(L) := \left\{ f : [0,1] \mapsto \mathbb{R} \ : \ \sum_{0 \le m \le k-1} \|D^m f\|_\infty + \frac{|D^{k-1}f(x_1) - D^{k-1}f(x_2)|}{|x_1 - x_2|} \le L, \forall x_1, x_2 \in [0,1] \right\}$$

We can show that $\Sigma_k(L) \subsetneq \mathscr{F}(k, L)$ and $\Sigma_k(L)$ is not dense in $\mathscr{F}(k, L)$

* $D^k f$ denotes the kth weak derivative and $D^0 f = f$