Locally adaptive nonparametric regression through shape-constrained classes

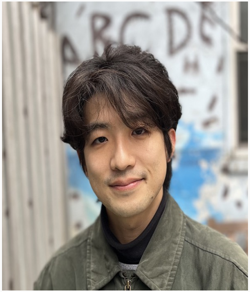
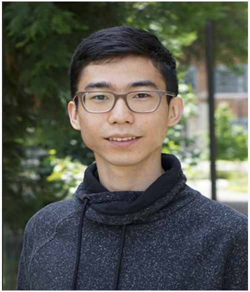
Smoothness versus shape

Arun Kumar Kuchibhotla

15 July, 2025

Carnegie Mellon University

Collaborators

Kenta Takatsu (CMU) Tianyu Zhang (CMU/UCSD)

This research was supported by NSF-DMS-2210662.
Available at ArXiv:2307.05732.

Table of contents

Introduction to Nonparametric Regression

Nonparametric Regression: smoothness classes

- ⋆ Suppose $(X, Y) \in \mathbb{R}^2$ is a random vector and we are interested in estimating

$$f_0(x) = \mathbb{E}[Y|X = x],$$

the conditional mean.

- ⋆ Traditional estimators include local averaging, series regression, least squares, and so on.

- ⋆ The convergence rates are crucially dependent on the smoothness of f_0.

- ⋆ If $f_0(\cdot)$ is known to be a Lipschitz function, then

$$\inf_{\widetilde{f}_n} \sup_{f_0 \in \mathrm{Lip}} \mathbb{E}_{f_0} \|\widetilde{f}_n - f_0\|^2 \asymp n^{-2/3}.$$

Nonparametric Regression: Shape-constrained classes

- $\star$ Instead of smoothness, suppose we know $f_0(\cdot)$ is monotonically non-decreasing.

- $\star$ A natural estimator is the least squares estimator (with no tuning parameters):

$$\widehat{f}_n := \mathop{\arg\min}_{f:\ \mathrm{non-dec}} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

- $\star$ Here as well, we have

$$\sup_{f_0\ \mathrm{non-dec}} \mathbb{E}\|\widehat{f}_n - f_0\|^2 \asymp \inf_{\widetilde{f}_n} \sup_{f_0} \mathbb{E}\|\widetilde{f}_n - f_0\|^2 \asymp n^{-2/3}.$$

- $\star$ Additionally, if f_0 is a constant, then

$$\|\widehat{f}_n - f_0\|^2 \asymp \frac{1}{n}, \quad \text{(ignoring logarithmic factors.)}$$

Comparison of smoothness and shape-constrained classes

- ⋆ Shape-constraints are easily interpretable and justifiable in many applications.
- ⋆ The minimax rate of convergence does not depend on the smoothness but more importantly on the metric entropy.
- ⋆ Metric entropy is $\log N(\varepsilon; \mathcal{F})$ where $N(\varepsilon; \mathcal{F})$ is the number of ε-radius balls needed to cover a function class.
- ⋆ We know

$$\log N(\varepsilon; \mathrm{Lip}) \asymp \log N(\varepsilon; \mathrm{non-dec}),$$

  which leads to the same minimax rate
- ⋆ Local geometries are significantly different. For example,

$$\begin{aligned} \text{Quadratic} + \text{Lipschitz} &= \text{Lipschitz}, \\ \text{Quadratic} + \text{non-dec} &\neq \text{non-dec}. \end{aligned}$$

Illustration of Local Neighborhood in Shape-constrained Class

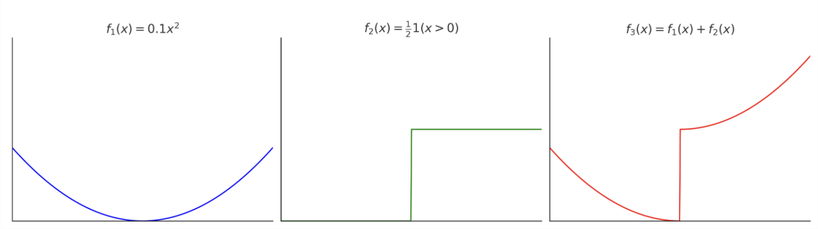


Figure 1: The neighborhood of f_2 cannot contain arbitrarily non-increasing functions. This implies a smaller metric entropy for the neighborhood of f_2 in the class of non-decreasing functions.

Question: Can we use shape-constrained classes to learn smoothness classes?

New Decomposition Spaces

A Result from Optimization Theory

- ⋆ Zlobec (2006, Optimization) proved that any twice differentiable function with a bounded second derivative is convexifiable.
- ⋆ The underlying principle is very simple and applies to all smoothness classes.
- ⋆ If f_0 is L-Lipschitz, i.e.,

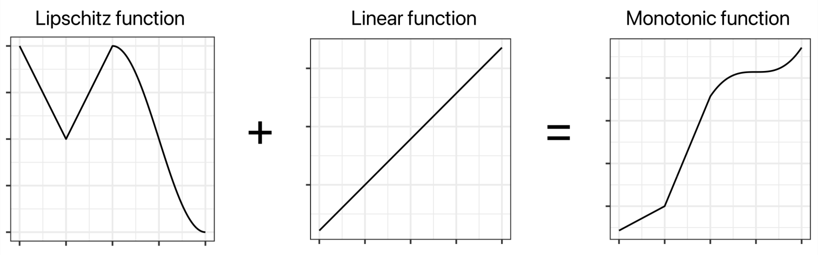
$$|f_0(x) - f_0(y)| \leq L|x - y|, \quad (\text{Think: } |f_0'(x)| \leq L \quad \text{for all} \quad x \in \mathbb{R})$$

  then there exists a non-decreasing function g_0 such that $f_0(x) = g_0(x) - Lx$ for all x.

- ⋆ **Proof:** Consider $g_0(x) = f_0(x) + Lx$.

$$g_0'(x) = f_0'(x) + L \geq 0 \quad \text{for all} \quad x \in \mathbb{R}.$$

Illustration



Extension

⋆ Define

$$\begin{aligned}
\mathcal{C}(1) &:= \{g : \mathbb{R} \to \mathbb{R} : \ g \text{ non-decreasing}\}, \\
\Sigma(1, L) &:= \{f : \mathbb{R} \to \mathbb{R} : \ |f(x) - f(y)| \le L|x - y| \quad \text{for all} \quad x, y \in \mathbb{R}\}, \\
\mathcal{F}(1, L) &:= \{f : \mathbb{R} \to \mathbb{R} |\, \exists g \in \mathcal{C}(1) \text{ such that } f(x) = g(x) - Lx \ \forall\, x \in \mathbb{R}\}.
\end{aligned}$$

⋆ We have shown $\Sigma(1, L) \subseteq \mathcal{F}(1, L)$. We can also show $\mathcal{C}(1) \subseteq \mathcal{F}(1, L)$.

⋆ We can extend these results to higher-order smoothness. Any function $f : \mathbb{R} \to \mathbb{R}$ with $\|f^{(k)}\|_\infty \le L$ can be decomposed as

$$f(x) = g(x) - L\frac{x^k}{k!},$$

for a function g that satisfies $g^{(k)}(x) \ge 0$ for all $x \in \mathbb{R}$. (k-monotone.)

⋆ We can similarly define interpolation spaces $\mathcal{F}(k, L)$ that contain $\mathcal{C}(k)$ and $\Sigma(k, L)$.

Estimation Procedure

Estimation

- ⋆ Recall

$$\mathcal{F}(1, L) := \{f : \mathbb{R} \to \mathbb{R} |\, \exists g \text{ non-dec } f(x) = g(x) - Lx \;\forall\, x \in \mathbb{R}\}.$$

- ⋆ This naturally suggests performing least squares on the class of non-decreasing functions and on $L \geq 0$.

- ⋆ Such a procedure will always interpolate the data.

- ⋆ As a remedy, we suggest sample splitting to avoid overfitting. Split $1, 2, \dots, n$ into two parts $\mathcal{I}_1$ and $\mathcal{I}_2$.

- ⋆ For each $L \geq 0$,

$$\widehat{g}_L := \operatorname*{arg\,min}_{g \text{ non-dec}} \sum_{i \in \mathcal{I}_1} (Y_i + LX_i - g(X_i))^2.$$

- ⋆ Compute

$$\widehat{L} := \operatorname*{arg\,min}_{L \geq 0} \sum_{i \in \mathcal{I}_2} (Y_i + LX_i - \widehat{g}_L(X_i))^2.$$

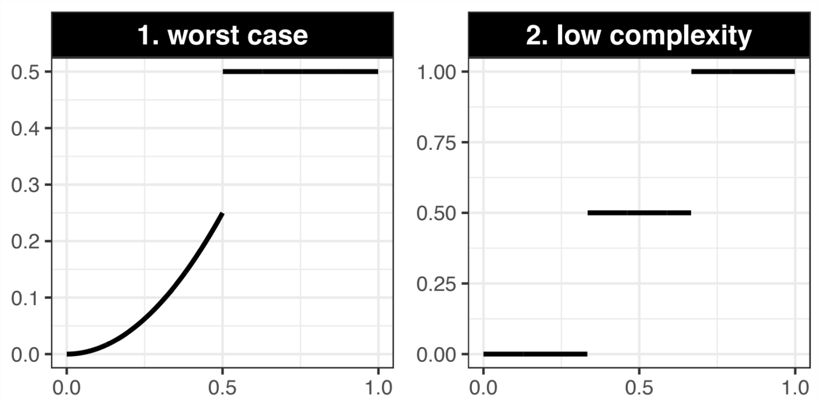


Figure 2: Monotone part of $f \in \mathcal{F}(1, L)$

Illustration

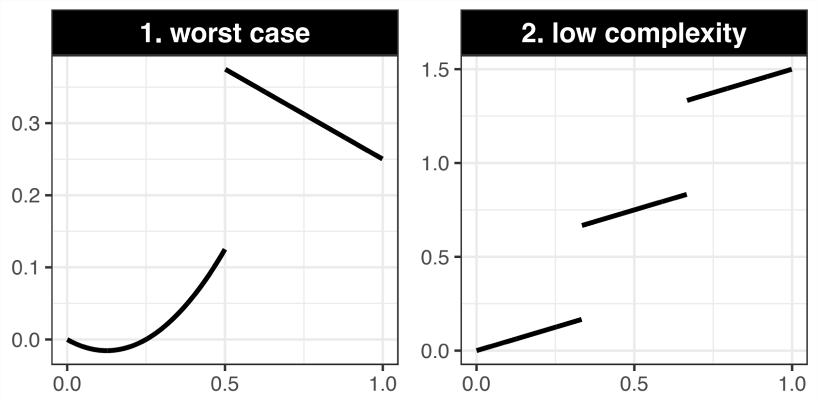


Figure 3: Functions in $\mathcal{F}(1, L)$

Illustration

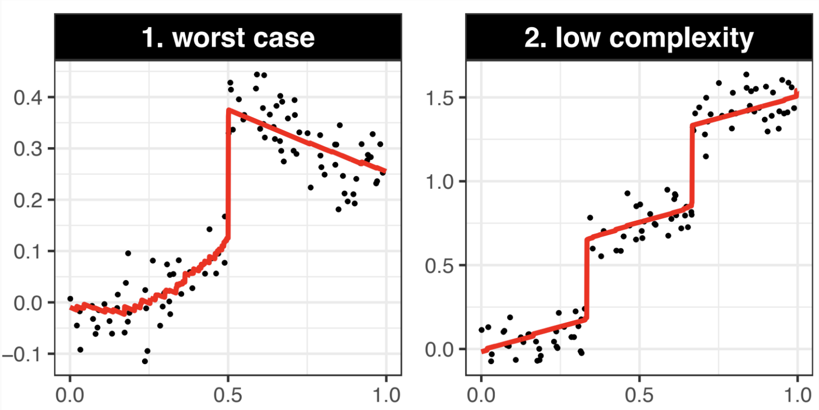


Figure 4: Estimation of functions in $\mathcal{F}(1, L)$. Visually, the estimator adapts to the flat pieces of the monotonic part of f.

Rates of Convergence and Adaptivity

What can be expected?

- ⋆ The least squares estimator on the class of non-decreasing functions admits adaptive rate of convergence:

$$\|\widehat{g}_n - g\|^2 \asymp \begin{cases} n^{-2/3}, & \text{for arbitrary non-decreasing g,} \\ \sqrt{m/n}, & \text{for } g \text{ m-piecewise constant.} \end{cases}$$

- ⋆ Hence, we can *expect* the following for $f \in \mathcal{F}(1, L)$:

$$\|\widehat{f}_n - f\|^2 \asymp \begin{cases} n^{-2/3}, & \text{for arbitrary } f \in \mathcal{F}(1, L), \\ \sqrt{m/n}, & \text{for } m\text{-piecewise non-dec } f, \\ \sqrt{m/n}, & \text{for } m\text{-piecewise non-dec + linear,} \end{cases}$$

- ⋆ Note that the third case include f being a linear function, if $f(x) + Lx$ is a constant (non-decreasing) function.

- ⋆ Hence, we can expect parametric rates if $f \in \mathcal{F}(1, L)$ is constant, or linear or m-piecewise non-decreasing.

Assumptions

- ⋆ For theoretical reasons, we restrict the second stage of the estimation procedure to

$$\widehat{L} := \arg\min_{L \in \mathcal{L}} \sum_{i \in \mathcal{I}_2} (Y_i + LX_i - \widehat{g}_L(X_i))^2,$$

  for some bounded set $\mathcal{L}$.

- ⋆ Let the largest element L_+ of $\mathcal{L}$ satisfy $L_+ = O(\log n)$.

- ⋆ Define $\xi = Y - \mathbb{E}[Y|X]$, and assume

$$\begin{aligned} \mathbb{E}[|\xi|^q|X] &\le K_q, \quad \text{for some} \quad q \ge 2 && (L_q) \\ \mathbb{E}[|\xi|^r|X] &\le Cr^{1/\alpha} \quad \text{for all} \quad r \ge 2, && (\mathrm{SW}_\alpha) \end{aligned}$$

- ⋆ Define

$$f_L^* = \arg\min_{f \in \mathcal{F}(1,L)} \|f_0 - f\|^2 : \quad \text{Projection of } f_0 \text{ onto } \mathcal{F}(1, L).$$

Well-specification: Oracle Inequality

- ⋆ If $f_0 \in \mathcal{F}(1, L_0)$ and $L_0 \leq L_+ = O(\log n)$, then

$$\|\widehat{f}_n - f_0\|^2 = O_p(1) \left\{ \frac{(\log n)^{4/3}}{n^{2/3}} + \frac{(\log n)^2}{n^{1-1/q}} \right\} \quad \text{under} \quad (L_q).$$

- ⋆ Under (SW_α), the second term becomes $(\log n)^{2+1/\alpha}/n$.

- ⋆ The second term arises from the selection of L.

- ⋆ Note the dependence on q. If $q < 3$, the second term dominates.

Well-specification: Oracle Inequality

- ⋆ If $f_0 \in \mathcal{F}_m(1, L_0)$ (m-piecewise constant non-doc + linear) and $L_0 \leq L_+ = O(\log n)$, then

$$\|\widehat{f}_n - f_0\|^2 = O_p(1)\left\{\frac{m(\log n)^2}{n} + \frac{(\log n)^2}{n^{1-1/q}}\right\} \quad \text{under} \quad (L_q).$$

- ⋆ Under (SW_α), the second term becomes $(\log n)^{2+1/\alpha}/n$.
- ⋆ The second term arises from the selection of L.
- ⋆ This result implies faster adaptive rates for low-complexity functions, e.g., constants, linear functions, and so on.

General Oracle Inequality

⋆ Recall

$$\widehat{L} := \arg\min_{L \in \mathcal{L}} \sum_{i \in \mathcal{I}_2} (Y_i + LX_i - \widehat{g}_L(X_i))^2,$$

for some bounded set $\mathcal{L}$.

$$\|\widehat{f}_n - f_0\|^2 \lesssim_P \inf_{L \in \mathcal{L}} \Bigg\{ \|f_L^* - f_0\|^2$$

Misspecification/Approximation Error

$$+ (\log n)^4 \inf_{1 \le m \le n} \left(\inf_{g \in \mathcal{C}_m(1)} \|g - g_L^*\|^2 + \frac{m \log^2(nL)}{n} \right) \Bigg\}$$

Oracle Inequality for Monotone Function Estimation

$$+ (\log n)^2 \begin{cases} n^{-1+1/q}, & \text{under } (L_q), \\ (\log n)^{2+1/\alpha}/n, & \text{under } (SW_q) \end{cases}$$

Error from Selection of L

Robust Estimation

- ⋆ The bounds obtained depend strongly on q and can prohibit (near) minimax optimality for small q.
- ⋆ This can be rectified using robust estimation of the mean in the selection of the L step.
- ⋆ Instead of

$$\widehat{L} := \arg\min_{L\in\mathcal{L}} \sum_{i\in\mathcal{I}_2} (Y_i + LX_i - \widehat{g}_L(X_i))^2,$$

we consider the median of means

$$\widehat{L} := \arg\min_{L\in\mathcal{L}} \max_{L'\in\mathcal{L}} \mathrm{MOM}_K \left\{ (Y_i + LX_i - \widehat{g}_L(X_i))^2 - (Y_i + LX_i - \widehat{g}_{L'}(X_i))^2 \right\},$$

with $K = 4\lceil \log(\mathrm{card}(\mathcal{L})) \rceil$.

- ⋆ All the previous results are valid as if the errors satisfy (SW_α) with $\alpha = 2$.

Conclusions

Conclusions

- ⋆ We have introduced new decomposition spaces that are more interpretable and adaptively estimable than classical smoothness classes.
- ⋆ Decomposition into monotone and linear pieces allows one to construct tests for monotonicity, for example.
- ⋆ Our two-stage estimation method yields near minimax rates simultaneously for the class of Lipschitz functions, the class of non-decreasing functions, and also for the class of linear functions.
- ⋆ Our decomposition methodology can be applied to other nonparametric settings, such as density estimation, NPIV, and classification problems.
- ⋆ Multivariate extensions are provided in the paper.