Adaptively Partitioning Max-Affine Estimators for Convex Regression

Gábor Balázs*
$\mathcal{G}\&\mathcal{G}$, Cartagena, Spain

Abstract

This paper considers convex shape-restricted nonparametric regression over subgaussian domain and noise with the squared loss. It introduces a tractable convex piecewise-linear estimator which precomputes a partition of the training data by an adaptive version of farthest-point clustering, approximately fits hyperplanes over the partition cells by minimizing the regularized empirical risk, and projects the result into the max-affine class. The analysis provides an upper bound on the generalization error of this estimator matching the rate of Lipschitz nonparametric regression and proves its adaptivity to the intrinsic dimension of the data mitigating the effect of the curse of dimensionality. The experiments conclude with competitive performance, improved overfitting robustness, and significant computational savings compared to existing convex regression methods.

1 INTRODUCTION

Convex (shape-restricted) nonparametric regression aims to estimate a convex regression function over some hypothesis class of convex functions. If the hypothesis class is chosen to be the set of *max-affine functions*, represented by the maximum of affine functions (also called hyperplanes), and the number of hyperplanes is chosen to be equal to the sample size, then the infinite dimensional minimization of the empirical risk can be written as a quadratic program and solved in polynomial time (Boyd and Vandenberghe, 2004, Section 6.5.5). Due to the simplicity of its construction and its tractable (polynomial-time) computability, this estimator has been dominating the convex regression literature. By today its convergence rate is well-understood in many settings (e.g., Seijo and Sen, 2011; Lim and Glynn, 2012; Lim, 2014; Balázs et al., 2015; Han and Wellner, 2016; Kur et al., 2020), and many algorithms have been proposed to reduce its polynomial but large computational cost (e.g., Lee et al., 2013; Aybat and Wang, 2014; Balázs et al., 2015; Mazumder et al., 2019; Chen and Mazumder, 2020).

It has been also observed that reducing the number of hyperplanes over the hypothesis class to significantly less than the number of samples can improve estimation properties, including reaching near-optimal minimax rate (e.g., Guntuboyina, 2012, Section 3.3), and adapting the convergence rate to some structure of the regression function (e.g., Han and Wellner, 2016, Section 4). However, there is no known tractable method to carry out the empirical risk minimization of these estimators, except when the hypothesis class is restricted to max-affine functions *inducing* a fixed partition of the data (Balázs et al., 2015, Section 4.4), that is when the max-affine function values for data points within the partition cells are provided by the same hyperplane. Although many clustering methods have been proposed to find an appropriate partition (e.g., Balázs et al., 2015; Siahkamari et al., 2020b), there have been no generalization bounds proved for these algorithms yet.

1.1 Contribution of the paper

This paper provides the first tractable partitioning max-affine estimator with an upper bound on the generalization error. The presented estimator (Algorithm 2) computes an approximate max-affine fit over a precomputed partition of the training data by minimizing the regularized empirical risk over a class of (not necessarily convex) piecewise-linear functions, and projects the result to the max-affine function class. The distance between the hypothesis class and the set of max-affine functions is controlled by regularization.

*email: gabalz@gandg.ai, web: gabalz.gandg.ai

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

The precomputed partition is delivered by an adaptive version of the farthest-point clustering method (Algorithm 1) which governs the complexity of the max-affine representation by setting the number of hyperplanes (or equivalently the partition size) appropriately for the regression task. The main result (Theorem 4.1) proves an adaptive (instance dependent) upper bound on the generalization error for this estimator as $O(n^{-2/(2+d_{\mathcal{X}})})$ for sample size n and intrinsic data dimension $d_{\mathcal{X}}$. Finally, an empirical evaluation is provided (Section 5) to demonstrate the competitive performance, the overfitting robustness, and the economical computational cost of the proposed method.

1.2 Related literature

There is a significant amount of work on max-affine estimators which minimize the empirical risk over the class of max-affine functions with a fixed number of hyperplanes being less than the number of samples including Guntuboyina (2012, Section 3.3), Han and Wellner (2016, Section 4), and Balázs (2016, Section 5.4). However, the tractability of such minimization has been only studied by Ghosh et al. (2019, 2021) for the special case when the regression function is max-affine with a known number of hyperplanes. They analyzed a carefully initialized version of the alternating minimization algorithm of Magnani and Boyd (2009) and proved a convergence rate for this case, while stated that the optimization of the empirical $\mathcal{L}_2$-risk over max-affine functions is intractable in general.

Balázs et al. (2015, Section 4.4) observed that if the partition is fixed, the empirical risk minimization over a max-affine function class inducing the fixed partition is tractable. Recently, Siahkamari et al. (2020b, Appendix A6) proposed using the Farthest-Point Clustering (FPC) algorithm which admits a 2-approximation guarantee (Gonzalez, 1985; Hochbaum and Shmoys, 1985), however they did not provide an analysis. Connecting the two results is challenging because the linearization of the convex regression function around the partition cell centers provides a max-affine function which might not induce the partition itself, and this makes it hard to prove an upper bound on the approximation error to the regression function by the max-affine class inducing the fixed partition. This work overcomes this problem in Section 4.1 by relaxing the hypothesis class to (not necessarily convex) piecewise-linear functions for the empirical risk minimization (3) which are close to the max-affine class inducing the fixed partition and contain a (max-affine) linearization of the regression function which admits bounded approximation error to the regression function itself.

The only tractable nonparametric convex estimator which provides a convergence rate guarantee for the subgaussian convex regression setting studied here is the max-affine estimator using as many hyperplanes as the number of samples, while the other tractable algorithms including LSPA (Magnani and Boyd, 2009), CAP (Hannah and Dunson, 2013) and AMAP (Balázs, 2016, Section 6.2.3) can provide consistency at best. However, for the theoretical guarantee, this max-affine estimator necessarily requires regularization to control the overfitting risk at the domain boundary (Balázs et al., 2015, Section 4.3), either by limiting the Lipschitz constant (Lim, 2014) or the magnitude of the estimator values (Han and Wellner, 2016, Section 3). The proposed algorithm (Algorithm 2) requires such regularization as well and it is designed to limit the Lipschitz constant, because imposing a Lipschitz constraint barely adds any cost to the training process while working with boundedness constraints requires the existence (and the knowledge) of the domain boundary which is not satisfied for infinite domains, and even for bounded settings with known boundaries, the extra optimization cost might be still significant.

This paper was also motivated by the literature on partitioning estimators for (non-convex) Lipschitz nonparametric regression (e.g., Györfi et al., 2002, Section 4), where adaptation to intrinsic dimension of the covariate data has been studied for kernel estimators and regression trees (Kpotufe, 2010; Kpotufe and Dasgupta, 2011). The model selection technique for choosing the number of hyperplanes in Algorithm 1 to adapt the rate to the intrinsic dimension in Theorem 4.1 is similar in spirit to the automatic stopping rule designed for random projection regression trees by Kpotufe (2009, Lemma 14). Adaptation to intrinsic dimension for convex regression has been studied by Han and Wellner (2016, Section 4) for max-affine estimators which learn the number of hyperplanes, but they do not provide a tractable training algorithm. They define the intrinsic dimension through the pseudodimension (Pollard, 1990, Section 4) of the set of max-affine mappings of the covariate data, which is different to the definition used here. Similar to Cutler (1993) and Clarkson (2006), in this paper the intrinsic dimension is defined on the data itself (Section 2.4), and it is propagated through the objective value of the clustering method (Lemma 4.2) and then the approximation error of the estimator to the regression function (4) to reduce the dimension dependence of the generalization bound (Theorem 4.1).

2 PROBLEM SETTING

Throughout the paper, let $[n] \doteq \{1, \dots, n\}$ for any positive integer $n \in \mathbb{N}$, denote the common asymptotic order of growth notations by $\Theta(\cdot)$, $O(\cdot)$, and their versions ignoring logarithmic terms by $\tilde{\Theta}(\cdot)$, $\tilde{O}(\cdot)$.

2.1 Convex regression

In general, a *random design regression* setting is defined by a probability measure μ over $\mathbb{R}^d \times \mathbb{R}$ for some *domain dimension* $d \in \mathbb{N}$. The learner is given a finite data set $\mathcal{D}_n \doteq \{(\boldsymbol{x}_i, y_i) : i \in [n]\} \subseteq (\mathbb{R}^d \times \mathbb{R})^n$ of size $n \in \mathbb{N}$ sampled independently and identically from the unknown distribution μ written as $(\boldsymbol{x}_i, y_i) \sim \mu$ for all $i \in [n]$. Its goal is to use the data $\mathcal{D}_n$ for constructing an estimate $f_n \in \mathcal{F}_n$ of some hypothesis class $\mathcal{F}_n \subseteq \{f \,|\, f : \mathbb{R}^d \to \mathbb{R}\}$ which minimizes the $\mathcal{L}_2$-error to the unknown *regression function* defined as $f_*(\boldsymbol{x}) \doteq \mathbb{E}_{(\boldsymbol{x},y)\sim\mu}[y|\boldsymbol{x}]$ almost surely (a.s.). Formally, the goal of the learner is to find an estimate f_n which minimizes the $\mathcal{L}_2$-error $\|f_n - f_*\|_\mu^2 \doteq \mathbb{E}_{(\boldsymbol{x},\cdot)\sim\mu}\big[|f_n(\boldsymbol{x}) - f_*(\boldsymbol{x})|^2\big]$ with high-probability with respect to (w.r.t.) the random sample $\mathcal{D}_n$ and the possible randomness of the estimate f_n. To address these regression problems, I use *least squares estimates* (LSEs) which minimize the empirical $\mathcal{L}_2$-risk $\mathcal{L}_n(f) \doteq \frac{1}{n}\sum_{i\in[n]} |f(\boldsymbol{x}_i) - y_i|^2$ for function $f \in \mathcal{F}_n$ over some hypothesis class $\mathcal{F}_n \subseteq \{h \,|\, h : \mathbb{R}^d \to \mathbb{R}\}$.

The paper considers *subgaussian convex regression* problems μ for which there exists a convex set $\mathcal{X}$ containing the *covariate data* $\mathcal{X}_n \doteq \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \subseteq \mathcal{X}$ with probability at least $1-\gamma$ for all $\gamma > 0$ (where $\mathcal{X}$ might depend on γ and n), and the regression function f_* belongs to the class of uniformly Lipschitz convex functions over $\mathcal{X}$ given by

$$\mathcal{F}_{\mathcal{X},L} \doteq \big\{f \,\big|\, f \text{ is convex and } L\text{-Lipschitz w.r.t. } \|\cdot\| \text{ on } \mathcal{X}\big\}$$

for some Lipschitz constant $L \geq 0$ and the Euclidean norm $\|\cdot\|$, that is $f_* \in \mathcal{F}_{\mathcal{X},L}$. Additionally, both the *covariate* $\boldsymbol{x}$ and the *observation noise* $(y - f_*(\boldsymbol{x}))|\boldsymbol{x}$ are subgaussian random variables $(\boldsymbol{x}, y) \sim \mu$ satisfying

$$\begin{aligned} \mathbb{E}\big[e^{\|\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}]\|^2/\rho^2}\big] &\leq 2\,, \\ \mathbb{E}\big[e^{|f_*(\boldsymbol{x})-y|^2/\sigma^2}\big|\boldsymbol{x}\big] &\leq 2 \text{ a.s.}\,, \\ \mathbb{E}[y - f_*(\boldsymbol{x}) \,|\, \boldsymbol{x}] &= 0 \text{ a.s.}\,, \end{aligned} \tag{1}$$

with some subgaussian parameters $\rho > 0$ and $\sigma > 0$.

The probabilistic statement of $\mathcal{X}_n \subseteq \mathcal{X}$ is needed for the problem setup to cover locally Lipschitz regression functions over covariates with unbounded support, for example the quadratic regression function over a covariate with standard normal distribution. Define the radius of $\mathcal{X}_n$ by $\rho_n \doteq \max_{i\in[n]} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|$, its center by $\bar{\boldsymbol{x}} \doteq \frac{1}{n}\sum_{i\in[n]} \boldsymbol{x}_i$, and the d-dimensional ball with radius $r > 0$ around center $\boldsymbol{x}_0 \in \mathbb{R}^d$ by $\mathcal{B}(\boldsymbol{x}_0, r) \doteq \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{x}_0\| \leq r\}$. Because the covariate data satisfies $\mathcal{X}_n \subseteq \mathcal{B}(\bar{\boldsymbol{x}}, \rho_n)$, it can be also bounded by the Chernoff and union bounds as

$$\mathbb{P}\{\rho_n > \rho_\gamma\} \leq 2n\, e^{-\rho_\gamma^2/(2\rho)^2} = \gamma/2 \tag{2}$$

with $\rho_\gamma \doteq 2\rho \ln(4n/\gamma)$ for any $\gamma > 0$. Then without loss of generality $\mathcal{X}$ can always be considered bounded as $\mathcal{X} \subseteq \mathcal{B}(\mathbb{E}[\boldsymbol{x}], 2\rho_\gamma)$ with probability at least $1-\gamma$ due to $\Pr\{\|\bar{\boldsymbol{x}} - \mathbb{E}[\boldsymbol{x}]\| > \rho_\gamma\} \leq \gamma/2$. For the example of the quadratic regression function $f_*^{\text{sq}}(\boldsymbol{x}) \doteq \|\boldsymbol{x}\|^2/2$ over the standard normal covariate, consider $f_*^{\text{sq}} \in \mathcal{F}_{\mathcal{X},L}$ with $\mathcal{X} = \mathcal{B}(\mathbb{E}[\boldsymbol{x}], L)$, $L = 2\rho_\gamma$, and $\rho = 1$.

2.2 Max-affine functions and partitions

Throughout the paper the hypothesis classes for the LSEs are chosen to be "close to" *max-affine function classes* with at most $K \in \mathbb{N}$ hyperplanes defined as

$$\mathcal{H}_K \doteq \big\{f \,\big|\, f(\boldsymbol{x}) = \max_{k\in[K]} \boldsymbol{a}_k^\top \boldsymbol{x} + b_k\big\}.$$

Max-affine functions induce data *partitions*. Define a partition $\mathcal{P}_K$ over the sample indices $[n]$ with $K \in \mathbb{N}$ cells by $\mathcal{P}_K \doteq \{\mathcal{C}_k : k \in [K]\}$ where the cells are nonempty and disjoint $k \neq l \iff \mathcal{C}_k \cap \mathcal{C}_l = \emptyset$, and cover all indices as $[n] = \cup_{k\in[K]}\mathcal{C}_k$.

2.3 Set covers and Voronoi partitions

Partitions are related to set covers. For metric space $(\mathcal{Z}, \ell)$ and $\epsilon > 0$, the set $\{\boldsymbol{z}_k \in \mathcal{Z} : k \in [K]\}$ is an *ϵ-cover* of $\mathcal{Z}$ w.r.t. the distance $\ell : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$ if the union of the ϵ-balls around $\boldsymbol{z}_k$ covers $\mathcal{Z}$, that is $\mathcal{Z} \subseteq \cup_{k\in[K]}\{\boldsymbol{z} \in \mathcal{Z} : \ell(\boldsymbol{z}_k, \boldsymbol{z}) \leq \epsilon\}$. The cardinality of the smallest ϵ-cover of $\mathcal{Z}$ is called the *ϵ-covering number* of $\mathcal{Z}$ w.r.t. ℓ and denoted by $N_\ell(\mathcal{Z}, \epsilon)$. Due to the *volume argument* (e.g., Pollard, 1990, Lemma 4.1), the ϵ-covering number of a d-dimensional ball $\mathcal{B}(\boldsymbol{x}_0, r)$ with radius $r > 0$ around any center $\boldsymbol{x}_0 \in \mathbb{R}^d$ satisfies $N_{\|\cdot\|}\big(\mathcal{B}(\boldsymbol{x}_0, r), \epsilon\big) \leq \max\{1, (3r/\epsilon)^d\}$ for any $\epsilon > 0$.

Define the *Voronoi partition* on $\mathcal{X}_n$ around K centers $\boldsymbol{x}_{i_1}, \dots, \boldsymbol{x}_{i_K}$ by $\mathcal{P}_V(\{\boldsymbol{x}_{i_1}, \dots, \boldsymbol{x}_{i_K}\}) \doteq \{\mathcal{C}_k : k \in [K]\}$ with *Voronoi cells* $\mathcal{C}_k \doteq \{i \in [n] : \|\boldsymbol{x}_i - \boldsymbol{x}_{i_k}\| = \min_{l\in[K]} \|\boldsymbol{x}_i - \boldsymbol{x}_{i_l}\|\}$ for all $k \in [K]$ where ties are broken arbitrarily. When the centers define an ϵ-cover of $\mathcal{X}_n$, the Voronoi cell radii are also bounded as $\max_{k\in[K]} \max_{i\in\mathcal{C}_k} \|\boldsymbol{x}_i - \boldsymbol{x}_{i_k}\| \leq \epsilon$.

2.4 Intrinsic dimension of the data

To discuss adaptive (instance dependent) rates, consider the *box dimension* (similar to Cutler, 1993; Clarkson, 2006) of $\mathcal{X}_n$ which is the smallest constant $d_{\mathcal{X}} \in (0, d]$ such that $N_{\|\cdot\|}(\mathcal{X}_n, \epsilon) \leq (3\rho_n/\epsilon)^{d_{\mathcal{X}}}$ a.s. holds for all $\epsilon \in (0, 2\rho_n]$ and $n \in \mathbb{N}$ if $\mathcal{X}_n \subseteq \mathcal{X}$. Notice that as $\mathcal{X}_n \subseteq \mathcal{B}(\bar{\boldsymbol{x}}, \rho_n)$, the general case can be always recovered by setting $d_{\mathcal{X}} = d$, so the box dimension is well-defined. However, the box dimension $d_{\mathcal{X}}$ (which can be non-integer for fractals) generalizes extensions of the VC-dimension like the pseudodimension (Pollard, 1990, Section 4), and Assouad's

doubling dimension capturing sparse data and Riemannian manifolds with bounded condition numbers (Kpotufe and Dasgupta, 2011, Section 2). Notice that such structured data might appear for convex regression in a nonlinear way as well because there is no convexity restriction on the support of the covariate distribution.

3 THE PROPOSED ALGORITHM

Let $\mathcal{P}_K = \{\mathcal{C}_k : k \in [K]\}$ be a fixed partition, $\beta \geq 0$, and consider the quadratic programming problem,

$$\begin{aligned} \min_{\substack{\boldsymbol{a}_1,\dots,\boldsymbol{a}_K \in \mathbb{R}^d,\\ b_1,\dots,b_K \in \mathbb{R},\\ V \in \mathbb{R}}} \; & \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} \left| \boldsymbol{a}_k^\top \boldsymbol{x}_i + b_k - y_i \right|^2 + \beta V^2 \\ \text{subject to } & \boldsymbol{a}_k^\top \boldsymbol{x}_i + b_k \geq \boldsymbol{a}_l^\top \boldsymbol{x}_i + b_l - V \\ \text{and } & \|\boldsymbol{a}_k\|_\infty \leq L \text{ for all } i \in \mathcal{C}_k \text{ and } k, l \in [K] \,, \end{aligned} \tag{3}$$

where $\|\cdot\|_\infty$ denotes the max-norm. Using the optimized values $\{(\boldsymbol{a}_k, b_k) : k \in [K]\}$, the max-affine estimate is $f_n(\boldsymbol{x}) \doteq \max_{k \in [K]} \boldsymbol{a}_k^\top \boldsymbol{x} + b_k$ for all $\boldsymbol{x} \in \mathbb{R}^d$.

Notice that for $\beta = \infty$ (using $0 \cdot \infty \doteq 0$), $K = n$, and trivial partition $\mathcal{P}_n = \big\{\{i\} : i \in [n]\big\}$, (3) computes the *Convex Nonparametric Least Squares* (CNLS) estimator (Boyd and Vandenberghe, 2004, Section 6.5.5) which is a LSE over the class of convex functions $\mathcal{F}_{\mathcal{X},L}$. However, when $\beta < \infty$ and $K < n$, the solution of (3) yields a LSE over some class of (not necessarily continuous) piecewise-linear functions and its result is projected into the max-affine class $\mathcal{H}_K$ by the definition of f_n. The role of variable V in (3) is to bound the distance of this projection.

To compute a partition $\mathcal{P}_K$, consider a slightly modified version of the Farthest-Point Clustering (FPC) algorithm, which is named here *Adaptive FPC* (AFPC) and it is given by Algorithm 1. Just as FPC, AFPC

Algorithm 1 $\mathcal{P}_K \leftarrow$ **AFPC**$(\mathcal{X}_n)$

input: $\mathcal{X}_n = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$
Set $i_1 \in [n]$ arbitrarily, $K \leftarrow 1$ and $\mathcal{P}_K \leftarrow \{[n]\}$
while $\rho_n^2 K < n\,\Delta^2(\mathcal{P}_K)$ and $K < n^{d/(2+d)}$ **do**
  $K \leftarrow K + 1$
  $i_K \in \operatorname{argmax}_{i \in [n]} \min_{j \in \{i_1,\dots,i_{K-1}\}} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$
  $\mathcal{P}_K \leftarrow \mathcal{P}_V(\{\boldsymbol{x}_{i_1}, \dots, \boldsymbol{x}_{i_K}\})$
end while
output: $\mathcal{P}_K$

also minimizes the objective function $\Delta(\cdot)$ defined by

$$\Delta(\mathcal{P}) \doteq \max_{k \in [K]} \max_{i,j \in \mathcal{C}_k} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \,, \quad \mathcal{P} = \{\mathcal{C}_k : k \in [K]\},$$

over the set of all Voronoi partitions $\mathcal{P}$ of size K on the covariate data $\mathcal{X}_n$. The only modification of AFPC is the early termination rule $\rho_n^2 K < n\,\Delta^2(\mathcal{P}_K)$ which can detect if the clustering error $\Delta^2(\mathcal{P}_K)$ achieves a sufficient level for the regression task given by the threshold $\rho_n^2 K/n$. Hence, AFPC inherits the 2-approximation property of FPC (Gonzalez, 1985; Hochbaum and Shmoys, 1985) that is for every K these algorithms find a partition $\mathcal{P}_K$ which has no worse objective value $\Delta(\mathcal{P}_K)$ than twice the optimal.

Finally, consider the *Adaptively Partitioning CNLS* (APCNLS) estimator as given by Algorithm 2, which is the topic of this paper. APCNLS uses AFPC

Algorithm 2 $f_n \leftarrow$ **APCNLS**$(\mathcal{D}_n, L)$

input: $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i) : i \in [n]\}$, $L > 0$
$\mathcal{P}_K \leftarrow \text{AFPC}(\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\})$
$\{(\boldsymbol{a}_k, b_k) : k \in [K]\} \leftarrow$ solution to (3) with partition $\mathcal{P}_K$, Lipschitz constant L, and $\beta = d \ln(n)$
output: $f_n(\boldsymbol{x}) = \max_{k \in [K]} \boldsymbol{a}_k^\top \boldsymbol{x} + b_k$

to compute a partition $\mathcal{P}_K$, and searches through a class of piecewise-linear functions which are approximately max-affine inducing $\mathcal{P}_K$. The distance of the piecewise-linear estimate from the max-affine class $\mathcal{H}_K$ is controlled by quadratic penalization in (3) which allows bounding the distance between the projected max-affine estimate f_n and the regression function f_*.

Notice that APCNLS requires the Lipschitz constant L which is an upper bound on the Lipschitz constant of the regression function $f_* \in \mathcal{F}_{\mathcal{X},L}$. The role of L is to bound the steepness of the estimate and so the amount of overfitting at the domain boundary thus avoiding infinite $\mathcal{L}_2$-error (Balázs et al., 2015, Section 4.3). In the absence of such guess, one can choose $L = \Theta\big(\ln(n)\big)$ as proposed by Blanchet et al. (2019) to ensure that the generalization bound holds for large enough sample size n. The empirical results of Section 5 suggest that APCNLS is much less sensitive to this Lipschitz parameter than CNLS.

4 THEORETICAL GUARANTEES

The main result of this paper is presented by Theorem 4.1 which proves an adaptive generalization bound for APCNLS estimates scaling exponentially by the intrinsic dimension of the covariate data $d_{\mathcal{X}}$ instead of the domain dimension d.

Theorem 4.1. *Let f_n be the estimate of Algorithm 2 (APCNLS). Then for any $\gamma > 0$, it holds with probability at least $1 - \gamma$ that*

$$\|f_n - f_*\|_\mu^2 = O\Big(d^2\, n^{-2/(2+d_{\mathcal{X}})} R_\mu\Big),$$

where $R_\mu \doteq \big(L^2 \rho_\gamma^2 + \sigma^2 \ln(B/\gamma)\big) \ln(n) \ln(n/\gamma)$, and $B \doteq n\big(\ln(1/\gamma) + dL^2\rho_\gamma^2/\sigma^2\big)^{1/2}$.

The bound $\tilde{O}(n^{-2/(2+d_{\mathcal{X}})})$ matches the rate of kernel regressors in the (non-convex) Lipschitz regression setting (Kpotufe, 2010, Theorem 21), and slightly improves the rate $\tilde{O}(n^{-2/(2+d_{\mathcal{X}}\ln d_{\mathcal{X}})})$ of tree-based estimators (Kpotufe and Dasgupta, 2011, Theorem 9).

In the worst case when $d_{\mathcal{X}} = d$, the APCNLS rate $\tilde{O}(n^{-2/(2+d)})$ is slightly worse than the CNLS rate $\tilde{O}(n^{-2/d})$ for $d > 4$ (e.g., Kur et al., 2020). However, the difference is $n^{-2/(d+2)}/n^{-2/d} = n^{4/(d^2+2d)}$ which decreases so rapidly in d that it is negligible for most practical settings (e.g., it is upper bounded by constant 50 for $d \geq 10$ and $n \leq 1.3 \cdot 10^{50}$ approximating the number of atoms in the Earth). The rate difference occurs as the generalization bound proofs of APCNLS and CNLS are quite different. For APCNLS the proof balances the estimation error of a parametric function class within $\mathcal{H}_K$ with its approximation error to the convex regression function f_* leading to the bound $\tilde{O}\big(\frac{K}{n} + K^{-2/d_{\mathcal{X}}}\big)$, which is vacuous for the setting of CNLS when $K = n$. On the other hand, CNLS uses the hypothesis class of a restricted set of convex functions which has no approximation error to f_*, and for $d > 4$ it balances the so-called entropy integral $\tilde{O}\big(n^{-1/2}\int_\delta^1 \epsilon^{-d/4}\,d\epsilon + \delta\big)$ with $\delta = \Theta(n^{-2/d})$. Hence proving adaptivity w.r.t. $d_{\mathcal{X}}$ for CNLS requires adaptive log-covering numbers of convex functions $\tilde{O}(\epsilon^{-d_{\mathcal{X}}/2})$ which is only straightforward for linear manifolds performing dimensionality reduction for max-affine functions by adapting Lemma 4.3 of Balázs et al. (2015). In contrast, the box dimension of Section 2.4 covers many nonlinear manifolds as well.

The rate $\tilde{O}(n^{-2/(2+d_{\mathcal{X}})})$ is near-optimal (i.e., optimal up to logarithmic factors) for Lipschitz regression (Stone, 1982), but it is unclear whether it could be improved to $\tilde{O}(n^{-4/(4+d_{\mathcal{X}})})$ for convex regression in general (considering nonsmooth regression functions). On one hand the complexity of hypothesis classes might be smaller for convex regression, but as there is no convexity restriction on the support of μ, it can be a subspace without containing a single line (e.g., the surface of a ball) which might fade the benefit of convexity. Section 4.2 briefly discusses the case of smooth convex regression when the rate improves to $\tilde{O}(n^{-4/(4+d_{\mathcal{X}})})$.

The result of Theorem 4.1 still provides an initial fast learning rate in a more realistic setting when the covariate data lies on a low-dimensional manifold only approximately as discussed in Section 4.3.

The bound of Theorem 4.1 can be improved by a factor of d if one uses $\|\cdot\|$ instead of $\|\cdot\|_\infty$ for the Lipschitz constraints in (3) which avoids using the bound $\|\cdot\| \leq \sqrt{d}\,\|\cdot\|_\infty$ in the proof. However, that turns (3) into a quadratically constrained quadratic program which is more challenging to solve numerically in practice.

The number of arithmetic operations of solving (3) is $\tilde{O}\big((nK)^{3/2}(dK)^2\big)$ using interior-point methods (Boyd and Vandenberghe, 2004, Section 11.5) which suggests a significant computational saving for APCNLS using $K \leq \lceil n^{d/(2+d)} \rceil$ in contrast to CNLS using $K = n$ even for the worst-case when $d_{\mathcal{X}} = d$. The experiments in Section 5 support this as well.

4.1 Proof sketch of Theorem 4.1

This section describes the ideas and the main steps for the proof of Theorem 4.1. The skipped technical details are presented in Appendix A. The proof is also conditioned on the high-probability events $\mathcal{X}_n \subseteq \mathcal{X}$ and $\rho_n \leq \rho_\gamma$ as discussed around (2).

For some Voronoi partition $\mathcal{P}_K = \{\mathcal{C}_1, \dots, \mathcal{C}_K\} = \mathcal{P}_V(\{\boldsymbol{x}_{i_1}, \dots, \boldsymbol{x}_{i_K}\})$, and $V_0 \geq 0$, define the class

$$\mathcal{G}_n(\mathcal{P}_K, V_0) \doteq \Big\{ g \,\Big|\, g(\boldsymbol{x}) = \sum_{k\in[K]} \mathbb{I}\{\boldsymbol{x} \in \mathcal{S}_k\}\,(\boldsymbol{a}_k^\top \boldsymbol{x} + b_k), \\ \boldsymbol{a}_k^\top \boldsymbol{x}_i + b_k \geq \boldsymbol{a}_l^\top \boldsymbol{x}_i + b_l - V_0, \\ i \in \mathcal{C}_k,\, k, l \in [K] \Big\},$$

with $\mathcal{S}_k \doteq \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{x}_{i_k}\| = \min_{l\in[K]} \|\boldsymbol{x} - \boldsymbol{x}_{i_l}\|\}$, where ties are broken arbitrarily and $\mathbb{I}\{\cdot\}$ denotes the $\{0,1\}$-valued indicator function. Then (3) computes a (regularized) LSE g_n over the class of piecewise-linear functions $\mathcal{G}_n(\mathcal{P}_K, V_0)$ for some $V_0 \geq 0$.

The key observation is that for a Voronoi partition $\mathcal{P}_K$ there exist hyperplane parameters $\{(\boldsymbol{a}_k^*, b_k^*) : k \in [K]\}$ defining a piecewise-linear function in $\mathcal{G}_n(\mathcal{P}_K, V_*)$ with max-affine approximation bound $V_* \doteq 2L\Delta(\mathcal{P}_K)$ which is also close to the regression function f_*. As for the construction of Balázs et al. (2015, Lemma 4.1), these parameters are given by $\boldsymbol{a}_k^* \doteq \nabla f_*(\boldsymbol{x}_{i_k})$ and $b_k^* \doteq f_*(\boldsymbol{x}_{i_k}) - \nabla f_*(\boldsymbol{x}_{i_k})^\top \boldsymbol{x}_{i_k}$ for all $k \in [K]$, where $\nabla f_*(\boldsymbol{x})$ denotes an arbitrarily fixed subgradient of the convex function f_* at $\boldsymbol{x} \in \mathcal{X}$. Define the max-affine function $h_*(\boldsymbol{x}) \doteq \max_{k\in[K]} (\boldsymbol{a}_k^*)^\top \boldsymbol{x} + b_k^*$ and $g_*(\boldsymbol{x}) \doteq \sum_{k\in[K]} \mathbb{I}\{\boldsymbol{x} \in \mathcal{S}_k\}\big((\boldsymbol{a}_k^*)^\top \boldsymbol{x} + b_k^*\big)$ which satisfy for any $i \in \mathcal{C}_k$ that

$$\begin{aligned} 0 &\leq f_*(\boldsymbol{x}_i) - h_*(\boldsymbol{x}_i) \\ &\leq f_*(\boldsymbol{x}_i) - g_*(\boldsymbol{x}_i) \\ &\leq \big(\nabla f_*(\boldsymbol{x}_i) - \nabla f_*(\boldsymbol{x}_{i_k})\big)^\top (\boldsymbol{x}_i - \boldsymbol{x}_{i_k}) \\ &\leq V_*, \end{aligned} \tag{4}$$

due to $g_* \leq h_*$ by definition, $f_* \in \mathcal{F}_{\mathcal{X},L}$, and $\mathcal{X}_n \subseteq \mathcal{X}$. Then (4) shows that $\max_{i\in[n]} |g_*(\boldsymbol{x}_i) - f_*(\boldsymbol{x}_i)| \leq V_*$, and also bounds the max-affine violation of g_* by V_* due to $g_*(\boldsymbol{x}_i) \geq h_*(\boldsymbol{x}_i) - V_*$, hence $g_* \in \mathcal{G}_n(\mathcal{P}_K, V_*)$.

Let $\{(\boldsymbol{a}_{n,k}, b_{n,k}) : k \in [K]\}, V_n$ be the solution to (3), and define the piecewise-linear function $g_n(\boldsymbol{x}) \doteq$

$\sum_{k\in[K]} \mathbb{I}\{\boldsymbol{x} \in \mathcal{S}_k\}(\boldsymbol{a}_{n,k}^\top \boldsymbol{x} + b_{n,k})$. As g_* is considered during the optimization of (3) due to $g_* \in \mathcal{G}_n(\mathcal{P}_K, V_*)$, it holds that $\mathcal{L}_n(g_n) + \beta V_n^2 \le \mathcal{L}_n(g_*) + \beta V_*^2$ which can be rewritten into

$$\|g_n - g_*\|_n^2 + \beta V_n^2 \le 2\langle g_* - g_n,\, g_* - y\rangle_n + \beta V_*^2 \quad (5)$$

using the empirical norm $\|h\|_n^2 \doteq \frac{1}{n}\sum_{i=1}^n |h(\boldsymbol{x}_i)|^2$ and inner product $\langle h_1, h_2\rangle_n \doteq \frac{1}{n}\sum_{i=1}^n h_1(\boldsymbol{x}_i)\, h_2(\boldsymbol{x}_i)$ for any $h, h_1, h_2 : \mathbb{R}^d \to \mathbb{R}$ with the slight abuse of notation of van de Geer (2000, Section 4.3) treating y as a function inside these empirical operators by defining $y(\boldsymbol{x}_i) \doteq y_i$ for all $i \in [n]$. Then using $\|g_* - f_*\|_n \le V_*$ by (4) with $ab \le a^2 + b^2/4$, (5) can be turned into

$$\begin{aligned}&\|g_n - g_*\|_n^2 + 2\beta V_n^2 \\ &\quad\le 4\langle g_* - g_n,\, f_* - y\rangle_n + 2(\beta+2)V_*^2.\end{aligned} \quad (6)$$

Later the proof will show that the right hand side of (6) is bounded by $\tilde{O}\big(\|g_n - g_*\|\sqrt{K/n} + K/n\big)$, so $\|g_n - g_*\|_n^2$, and βV_n^2 are both bounded by $\tilde{O}(K/n)$. This result, with the choice of β in Algorithm 2 and $\|f_n - g_n\|_n \le V_n$, implies $\|f_n - f_*\|_n^2 = \tilde{O}(K/n)$, which can be extended by the techniques of Balázs et al. (2016) to the bound $\|f_n - f_*\|_\mu^2 = \tilde{O}(K/n)$.

As $V_* = 2L\Delta(\mathcal{P}_K)$, the next step is to bound $\Delta(\mathcal{P}_K)$ for partition $\mathcal{P}_K$ computed by Algorithm 1 (AFPC). Denote the set of all Voronoi partitions of size K over a set $\mathcal{Z} \doteq \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_m\}$ with *cardinality* $|\mathcal{Z}| \doteq m \in \mathbb{N}$ by $\Gamma_K(\mathcal{Z}) \doteq \{\mathcal{P}_V(\{\boldsymbol{z}_{i_1}, \dots, \boldsymbol{z}_{i_K}\}) \,|\, i_1, \dots, i_K \in [m]\}$, and define the optimal clustering objective value by $\Delta_K^* \doteq \min_{\mathcal{P}\in\Gamma_K(\mathcal{X}_n)} \Delta_K(\mathcal{P})$. Although finding a partition achieving the optimal objective value Δ_K^* is NP-hard for $d \ge 2$, FPC (and so AFPC too) is proved to be a 2-approximation algorithm, which means that it finds a partition $\mathcal{P}_K$ in polynomial-time $O(ndK)$[2] which has no higher objective value than twice the optimal (Gonzalez, 1985; Hochbaum and Shmoys, 1985), that is $\Delta(\mathcal{P}_K) \le 2\Delta_K^*$. This result can be turned into an approximation bound as shown by Lemma 4.2.

Lemma 4.2. *Let $\mathcal{X}_n \subseteq \mathcal{X}$ and $\mathcal{P}_K$ be the Voronoi partition of size K returned by AFPC (Algorithm 1). Then $\Delta(\mathcal{P}_K) \le 12\rho_n \min\big\{K^{-1/d_\mathcal{X}}, \sqrt{K/n}\big\}$, and $K \le \min\big\{1 + (144\,n)^{d_\mathcal{X}/(2+d_\mathcal{X})}, \lceil n^{d/(2+d)}\rceil\big\}$.*

Proof. First, consider proving the claim $\Delta(\mathcal{P}_K) \le 12\rho_n K^{-1/d_\mathcal{X}}$. As discussed above, AFPC only differs from FPC in its stopping condition, so it also satisfies $\Delta(\mathcal{P}_k) \le 2\Delta_k^*$ for all $k \in [K]$. If $\Delta_K^* = 0$, then $\Delta(\mathcal{P}_K) \le 2\Delta_K^* = 0$ and the claim holds. As $\Delta_1^* \le 2\rho_n$, the claim also holds for $K = 1$. Then let $K > 1$, $\Delta_K^* > 0$, and suppose that $K > M$ where $M \doteq N_{\|\cdot\|}(\mathcal{X}_n, \Delta_K^*/\kappa)$ for some $\kappa > 2$. Denote a (Δ_K^*/κ)-cover of $\mathcal{X}_n$ by $\{\boldsymbol{x}_{i_1}, \dots, \boldsymbol{x}_{i_M}\}$ and define the Voronoi partition $\{\mathcal{C}_1, \dots, \mathcal{C}_M\} \doteq \mathcal{P}_V(\{\boldsymbol{x}_{i_1}, \dots, \boldsymbol{x}_{i_M}\})$. Then $\Delta_M^* \le \max_{k\in[M]} \max_{i,j\in\mathcal{C}_k} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \le (2/\kappa)\Delta_K^* < \Delta_K^*$ by the definition of Δ_M^* and the triangle inequality, which yields a contradiction as Δ_K^* is a non-increasing function of K. Hence, $K \le M$ holds and using the definition of the box dimension (Section 2.4) with $\mathcal{X}_n \subseteq \mathcal{X}$ yields $K \le N_{\|\cdot\|}(\mathcal{X}_n, \Delta_K^*/\kappa) \le (3\kappa\rho_n/\Delta_K^*)^{d_\mathcal{X}}$, which proves the claim by $\Delta(\mathcal{P}_K) \le 2\Delta_K^* \le 6\kappa\rho_n K^{-1/d_\mathcal{X}}$ and taking the limit $\kappa \to 2$.

If AFPC terminates by $\rho_n^2 K \ge n\Delta^2(\mathcal{P}_K)$, then $\Delta(\mathcal{P}_K) \le \rho_n\sqrt{K/n}$. Otherwise, if termination occurs with $K \ge n^{d/(2+d)}$, then $K^{-1/d_\mathcal{X}} \le K^{-1/d} \le n^{-1/(2+d)} = \sqrt{n^{d/(2+d)}/n} \le \sqrt{K/n}$ due to $d_\mathcal{X} \le d$, which proves the bound on $\Delta(\mathcal{P}_K)$.

Clearly, the AFPC termination condition directly imposes $K \le \lceil n^{d/(2+d)}\rceil$. Further, using the first part of the proof for the $(K-1)$-th (non-terminating) step, $\rho_n^2(K-1) < n\Delta^2(\mathcal{P}_{K-1}) \le 144\,n\,\rho_n^2(K-1)^{-2/d_\mathcal{X}}$, which proves the bound on K after rearrangement. □

Now consider the term $\langle g_* - g_n,\, f_* - y\rangle_n$ in (6). As $f_* - y$ is σ-subgausian noise by (1), and the estimate g_n lies in a class $\mathcal{G}_n(\mathcal{P}_K, V_0)$ having functions with $(d+1)K$ parameters, one expects to bound this term by $\tilde{O}\big(\|g_* - g_n\|_n\, \sigma\sqrt{dK/n}\big)$. Such bound requires the parameter space to be bounded, which holds by the Lipschitz constraints of (3) and a similar reasoning to Balázs (2016, Lemma 5.3) showing that the bias parameters are also bounded with high-probability as expressed by Lemma 4.3.

Lemma 4.3. *With probability at least $1-\gamma$ for $\gamma > 0$, $|b_{n,k} + \boldsymbol{a}_{n,k}^\top \mathbb{E}[\boldsymbol{x}] - \mathbb{E}[y]| \le \widehat{B}$ for all $k \in [K]$ with some $\widehat{B} > 0$ satisfying $\widehat{B}^2 = \Theta\big(dL^2\rho_\gamma^2 + \sigma^2\ln(1/\gamma)\big)$.*

Then the idea is to write

$$\langle g_* - g_n, f_* - y\rangle_n \le \|g_n - g_*\|_n \sup_{g\in\overline{\mathcal{G}}_n} \left\langle \frac{g_* - g}{\|g_* - g\|_n}, f_* - y \right\rangle_n$$

with the function class $\overline{\mathcal{G}}_n \doteq \{g \in \mathcal{G}_n(\mathcal{P}_K, V_0) : g(\boldsymbol{x}) = \sum_{k\in[K]} \mathbb{I}\{\boldsymbol{x} \in \mathcal{S}_k\}(\boldsymbol{a}_k^\top \boldsymbol{x} + b_k), \|\boldsymbol{a}_k\|_\infty \le L, |b_k + \boldsymbol{a}_k^\top \mathbb{E}[\boldsymbol{x}] - \mathbb{E}[y]| \le \widehat{B}\}$ for some large enough $V_0 > 0$, and use this to upper bound the term $\langle g_* - g_n,\, f_* - y\rangle_n$ as shown by Lemma 4.4.

Lemma 4.4. *With probability at least $1-\gamma$ for $\gamma > 0$,*

$$\begin{aligned}&\langle g_* - g_n,\, f_* - y\rangle_n \\ &\quad= O\Big(\|g_n - g_*\|_n\, \sigma\sqrt{\frac{dK}{n}\ln(B/\gamma)} + \sigma^2\frac{dK}{n}\ln(B/\gamma)\Big),\end{aligned}$$

where B is as defined for Theorem 4.1.

Combining $V_*^2 = O(L^2\rho_\gamma^2 K/n)$ due to Lemma 4.2 and $\rho_n \le \rho_\gamma$, Lemma 4.4 with (6) and $ab \le (a^2 + b^2)/2$

[2] There exist more efficient implementations (Feder and Greene, 1988; Har-Peled, 2004), but those are less incremental w.r.t. the partition size, so combining them with the AFPC stopping condition might be challenging.

directly yields bounds on $\|g_n - g_*\|_n$ and V_n as summarized by Lemma 4.5.

Lemma 4.5. *With probability at least* $1-\gamma$ *for* $\gamma > 0$,

$$\|g_n - g_*\|_n^2 = O\Big(\frac{dK}{n}\big(\sigma^2 \ln(B/\gamma) + \ln(n) L^2 \rho_\gamma^2\big)\Big),$$
$$V_n^2 = O\Big(\frac{K}{n}\big(\sigma^2 \frac{\ln(B/\gamma)}{\ln(n)} + L^2 \rho_\gamma^2\big)\Big).$$

Using $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, $V_*^2 = O(L^2 \rho_\gamma^2 K/n)$, and Lemma 4.5, the empirical distance between the APCNLS estimate f_n and the regression function f_* can be bounded with probability at least $1-\gamma$ as

$$\begin{aligned}\|f_n - f_*\|_n^2 &\le 3\|f_n - g_n\|_n^2 + 3\|g_n - g_*\|_n^2 + 3\|g_* - f_*\|_n^2 \\ &\le 3V_n^2 + 3\,\|g_n - g_*\|_n^2 + 3V_*^2 \\ &= O\Big(\frac{dK}{n}\big(\sigma^2 \ln(B/\gamma) + \ln(n) L^2 \rho_\gamma^2\big)\Big). \quad (7)\end{aligned}$$

Now by Lemma 4.2 there exists a (nonrandom) bound $\bar{K} > 0$ on the number of hyperplanes such that $K \le \bar{K} = O(n^{d_\mathcal{X}/(2+d_\mathcal{X})})$. Additionally, as the parameters $\{(\boldsymbol{a}_{n,k}, b_{n,k}) : k \in [K]\}$ are shared by f_n and g_n, Lemma 4.3 implies that f_n belongs to a max-affine function class with bounded parameter space, that is $f_n \in \overline{\mathcal{H}}_K \subseteq \overline{\mathcal{H}}_{\bar{K}}$, where

$$\overline{\mathcal{H}}_K \doteq \big\{f \in \mathcal{H}_K : f(\boldsymbol{x}) = \max_{k\in[K]} \boldsymbol{a}_k^\top \boldsymbol{x} + b_k,\ \|\boldsymbol{a}_k\|_\infty \le L,\ |b_k + \boldsymbol{a}_k^\top \mathbb{E}[\boldsymbol{x}] - \mathbb{E}[y]| \le \widehat{B}\big\},$$

and $\overline{\mathcal{H}}_{\bar{K}}$ is defined similarly.

Then, the approximation error of f_n to f_* is bounded by (7), the complexity of f_n is controlled through the (nonrandom) class $\overline{\mathcal{H}}_{\bar{K}}$, and the number of hyperplanes of f_n is bounded by $\bar{K} = O(n^{d_\mathcal{X}/(2+d_\mathcal{X})})$ due to Lemma 4.2, so empirical process theory of sieved estimators (e.g., van de Geer, 2000, Section 10.3), especially the results of Györfi et al. (2002, Theorem 11.5) as extended by Balázs et al. (2016, Theorem 1 with Lemma 3), can be used to combine the pieces together through the decomposition,

$$\begin{aligned}\|f_n - f_*\|_\mu^2 &\le \sup_{f \in \overline{\mathcal{H}}_{\bar{K}}} \|f - f_*\|_\mu^2 - 2\,\|f - f_*\|_n^2 \\ &\qquad + 2\,\|f_n - f_*\|_n^2 \\ &= \tilde{O}(d\bar{K}/n),\end{aligned}$$

and to prove the result of Theorem 4.1. The details of these steps are presented in Appendix A.2.

4.2 Smoothness of f_*

When the regression function f_* is (first-order) *smooth*, that is if $\|\nabla f_*(\boldsymbol{x}) - \nabla f_*(\widehat{\boldsymbol{x}})\| \le L_g \|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|$ for some $L_g > 0$ and all $\boldsymbol{x}, \widehat{\boldsymbol{x}} \in \mathcal{X}$, then (4) provides $V_* = O\big(L_g \Delta^2(\mathcal{P}_K)\big)$. By modifying the termination condition of Algorithm 1 (AFPC) to $\rho_n^4 K < n\,\Delta^4(\mathcal{P}_K)$ and $K < n^{d/(4+d)}$, the result of Lemma 4.2 changes to $\Delta(\mathcal{P}_K) = O\big(\rho_n \min\{K^{-1/d_\mathcal{X}}, (K/n)^{1/4}\}\big)$ and $K = O\big(K^{d_\mathcal{X}/(4+d_\mathcal{X})}\big)$. Then the same reasoning as above yields the generalization bound $\tilde{O}\big(\frac{K}{n} + V_*^2\big) = \tilde{O}\big(n^{-4/(4+d_\mathcal{X})}\big)$ for APCNLS using the modified AFPC termination condition. Unfortunately this adaptation of APCNLS to smoothness is not automatic, one might need cross-validation to make it so. It has been also stated as a challenging open problem to prove or disprove whether CNLS achieves this rate in the smooth convex regression setting (Kur et al., 2020, Section 6).

4.3 Measurement noise

It might happen that the covariate data comes from a low-dimensional manifold, but it is contaminated by "measurement noise". In this case the definition of intrinsic dimension could be $N_{\|\cdot\|}(\mathcal{X}_n, \epsilon) \le (\rho_n/\epsilon)^{d_\mathcal{X}} (\sigma_m/\epsilon)^d$ a.s. for all $\epsilon \in (0, 2\rho_n]$, where $\sigma_m \ge 0$ is the *level of measurement noise*. Because the box dimension was only used in the proof of Lemma 4.2 with $\epsilon = \Delta_K^*/2$, the fast APCNLS rate using $d_\mathcal{X}$ instead of d in Theorem 4.1 still holds until $\sigma_m \le \Delta_K^*/2$, and degrades towards the full-dimensional case afterwards.

5 EXPERIMENTS

This section provides a performance comparison between APCNLS (Algorithm 2) and CNLS on some selected synthetic problems. Because both algorithms require regularization of the Lipschitz constant L, two versions are considered for each. First, CNLS_* and APCNLS_* with L measured on the union of the training and test sets, which provides a tight setting. Second, $\text{CNLS}_{\ln}$ and $\text{APCNLS}_{\ln}$ with setting $L = \ln(n)$ as motivated by Blanchet et al. (2019), which is here often a looser setting. For reference, the results also include the ordinary least squares (OLS) estimate.

I used OSQP (Stellato et al., 2020) to solve (3), which is a first-order quadratic programming solver built on the idea of alternating direction methods of multipliers. The initial center of AFPC (Algorithm 1) was always taken to be the closest element within $\mathcal{X}_n$ to the center $\bar{\boldsymbol{x}}$ with respect to $\|\cdot\|$. For further details, Python implementation is provided in the supplementary material and at github.com/gabalz/cvxreg.

All plots show the average test errors $\|f_n - f_*\|_\mu^2$ measured on 10^6 independently drawn samples of 20 experiments with standard deviation error bars for each sample size $n \in \{100, 250, 500, 750, 1000\}$.

Denote the zero vector with appropriate size by $\mathbf{0}$, the d-dimensional identity matrix by I_d, the uniform distribution over a bounded set $\mathcal{S}$ by $\mathcal{U}(\mathcal{S})$, and the d-dimensional Gaussian distribution with mean $\boldsymbol{m} \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d \times d}$ by $\mathcal{N}(\boldsymbol{m}, \Sigma)$.

The regression function f_* was chosen to be either symmetric or truncated, and piecewise-linear or quadratic, that is $f_* \in \{f_*^{\mathrm{spl}}, f_*^{\mathrm{tpl}}, f_*^{\mathrm{sq}}, f_*^{\mathrm{tq}}\}$ with $f_*^{\mathrm{spl}}(\boldsymbol{x}) \doteq \|\boldsymbol{x}\|_1$, $f_*^{\mathrm{tpl}}(\boldsymbol{x}) \doteq \|\max\{\mathbf{0}, \boldsymbol{x}\}\|_1$, $f_*^{\mathrm{sq}}(\boldsymbol{x}) \doteq \|\boldsymbol{x}\|^2/2$, $f_*^{\mathrm{tq}}(\boldsymbol{x}) \doteq \|\max\{\mathbf{0}, \boldsymbol{x}\}\|^2/2$, where $\max\{\mathbf{0}, \boldsymbol{x}\}$ is meant to be coordinate-wise. These functions cover the nonsmooth and smooth settings and the truncated versions break the uniformity of the Lipschitz constant around the boundary. All experiments use Gaussian observation noise $(y - f_*(\boldsymbol{x}))|\boldsymbol{x} \sim \mathcal{N}(0, \sigma^2)$.

5.1 Without low-dimensional manifold

The experiments on nonsmooth and smooth problems (Fig. 1) over a full dimensional domain ($d_{\mathcal{X}} = d$) demonstrate a competitive performance of APCNLS to CNLS even in the absence of a manifold.

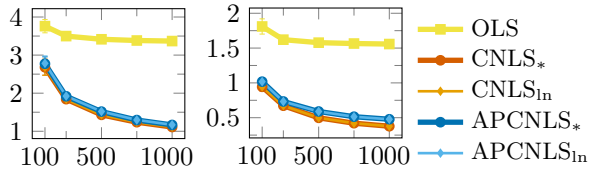

Figure 1: Estimating f_*^{spl} (left) over $\boldsymbol{x} \sim \mathcal{U}([-2,2]^{10})$, and f_*^{tq} (right) over $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, I_{10})$, both for $\sigma = 0.3$.

However, this is reached by reducing the number of hyperplanes (and the computational cost) of APCNLS quite significantly below the number of samples n (used by CNLS) as shown by Fig. 2.

$\boldsymbol{x} \setminus n$	100	250	500	750	1000
$\mathcal{U}([-2,2]^{10})$	34 ± 2	62 ± 4	100 ± 6	130 ± 7	158 ± 10
$\mathcal{N}(\mathbf{0}, I_{10})$	25 ± 3	48 ± 5	75 ± 10	100 ± 8	118 ± 10
$\boldsymbol{x}_{\mathrm{li},10}$	12 ± 1	17 ± 1	23 ± 1	26 ± 1	28 ± 2
$\boldsymbol{x}_{\mathrm{pe},10}$	5 ± 1	5 ± 1	7 ± 1	8 ± 1	9 ± 1

Figure 2: Number of hyperplanes K used by AFPC (and APCNLS) for some covariate distributions in $\mathbb{R}^{10}$.

However, the picture changes as the observation noise increases. Then, the results of Fig. 3 indicate that APCNLS is more robust than CNLS with respect to overfitting to a significantly higher observation noise (increased from $\sigma = 0.3$ of Fig. 1 to $\sigma = 3.0$).

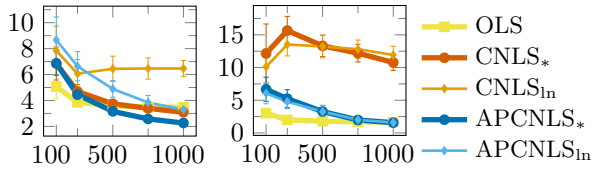

Figure 3: Estimating f_*^{spl} (left) over $\boldsymbol{x} \sim \mathcal{U}([-2,2]^{10})$, and f_*^{tq} (right) over $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, I_{10})$, both for $\sigma = 3.0$.

5.2 With linear low-dimensional manifold

Consider a linear 3-dimensional manifold ($d_{\mathcal{X}} = 3$), that is define the random vector $\boldsymbol{z}_3 \sim \mathcal{U}([-3,3]^3)$ and set $\boldsymbol{x}_{\mathrm{li},10} \sim \mathcal{N}([\boldsymbol{z}_3^\top \mathbf{0}^\top]^\top, \sigma_m^2 I_{10})$ with some measurement noise level $\sigma_m \geq 0$.

The results of Fig. 4 show that APCNLS delivers slightly better performance than CNLS, while it uses even less hyperplanes (and so requires even less computational effort) than for the full dimensional cases as presented above by Fig. 2. In particular for these experiments with sample size $n = 1000$, the CNLS estimators needed more than half an hour to train on average while the APCNLS estimators fitted in less than half a minute. These results also suggest that CNLS is adaptive to linear manifolds as well.

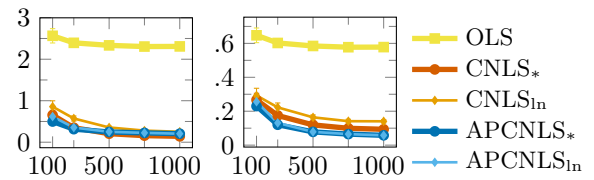

Figure 4: Estimating f_*^{spl} (left) and f_*^{tpl} (right) over $\boldsymbol{x}_{\mathrm{li},10}$ for $\sigma = 0.3$ and $\sigma_m = 0.1$.

5.3 With nonlinear low-dimensional manifold

Consider a nonlinear 1-dimensional manifold ($d_{\mathcal{X}} = 1$) when a scalar covariate is embedded into $\mathbb{R}^{10}$ by polynomial expansion with some measurement noise (Section 4.3). Formally, let $z_1 \sim \mathcal{U}([-1,1])$ and define $\boldsymbol{x}_{\mathrm{pe},10} \sim \mathcal{N}\big([z_1\, z_1^2\, \ldots\, z_1^{10}]^\top, \sigma_m^2 I_{10}\big)$ with some measurement noise level $\sigma_m \geq 0$.

The results of Fig. 5 show that APCNLS finds significantly better estimates than CNLS in these settings, and it does so by using even less resources as presented above by Fig. 2.

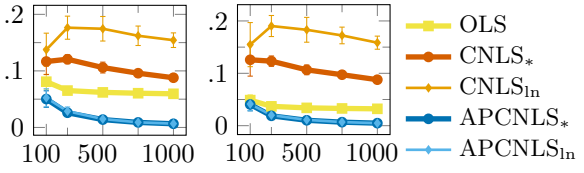

Figure 5: Estimating f_*^{sq} (left) and f_*^{tq} (right) over $\boldsymbol{x}_{\mathrm{pe},10}$ for $\sigma = 0.3$ and $\sigma_m = 0.1$.

The results of Fig. 6 demonstrate the cases of low (zero) and high (doubled) measurement noise. In the former no estimate exceeds OLS significantly (perhaps as the support of $\boldsymbol{x}$ does not contain any line), while in the latter APCNLS delivers far the lowest test error.

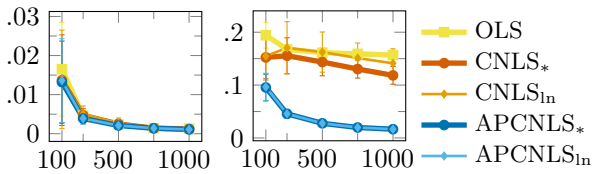

Figure 6: Estimating f_*^{sq} over $\boldsymbol{x}_{\mathrm{pe},10}$ for $\sigma = 0.3$ with $\sigma_m = 0.0$ (left) and $\sigma_m = 0.2$ (right).

Further experiments are provided in Appendix B.

6 CONCLUSIONS

This paper introduced a novel partitioning max-affine estimator (Algorithm 2) with an adaptive upper bound on the generalization error for the general subgaussian convex regression setting (Theorem 4.1) matching the adaptive rates of Lipschitz regression. This was obtained by a new analysis exploiting the properties of an appropriate Voronoi partition computed by an adaptive version of FPC, and the relaxed max-affine constraints in the empirical risk minimization process which considers the linearization of the regression function over the chosen partition. As CNLS has been recently extended to (non-convex) Lipschitz regression (Siahkamari et al., 2020a), perhaps one of the most interesting open question is whether the techniques of APCNLS carry over to those more general settings as well. For convex regression, the presented results indicate that APCNLS is a favorable alternative to CNLS in terms of theoretical guarantee, empirical performance, robustness, and computational cost.

Acknowledgements

This work was funded by Gema Sánchez Hernández and Gábor Balázs ($\mathcal{G}\&\mathcal{G}$). We thank the anonymous reviewers for their helpful suggestions.

References

Aybat, N. S. and Wang, Z. (2014). A parallel method for large scale convex regression problems. In *53rd IEEE Conference on Decision and Control*, pages 5710–5717.

Balázs, G. (2016). *Convex Regression: Theory, Practice, and Applications*. PhD thesis, University of Alberta.

Balázs, G., György, A., and Szepesvári, Cs. (2015). Near-optimal max-affine estimators for convex regression. In Lebanon, G. and Vishwanathan, S. V. N., editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 56–64, San Diego, California, USA. PMLR.

Balázs, G., Szepesvári, Cs., and György, A. (2016). Chaining bounds for empirical risk minimization. *arXiv preprint arXiv:1609.01872v1*.

Blanchet, J., Glynn, P. W., Yan, J., and Zhou, Z. (2019). Multivariate distributionally robust convex regression under absolute error loss. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Boucheron, S., Lugosi, G., and Massart, P. (2012). *Concentration Inequalities: A nonasymptotic theory of independence*. Clarendon Press.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Chen, W. and Mazumder, R. (2020). Multivariate convex regression at scale. *arXiv preprint arXiv:2005.11588v1*.

Clarkson, K. L. (2006). Nearest-neighbor searching and metric space dimensions. In Shakhnarovich, G., Darrell, T., and Indyk, P., editors, *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*, pages 15–59. MIT Press.

Cutler, C. D. (1993). A review of the theory and estimation of fractal dimension. In Tong, H., editor, *Dimension Estimation and Models: Nonlinear Time Series and Chaos Vol. I*, pages 1–107. World Scientific.

Feder, T. and Greene, D. H. (1988). Optimal algorithms for approximate clustering. In *Annual ACM symposium on Theory of computing*, volume 20, pages 434–444. ACM.

Ghosh, A., Pananjady, A., Guntuboyina, A., and Ramchandran, K. (2019). Max-affine regression: Provable, tractable, and near-optimal statistical estimation. *arXiv preprint arXiv:1906.09255v1*.

Ghosh, A., Pananjady, A., Guntuboyina, A., and Ramchandran, K. (2021). Max-affine regression: Parameter estimation for Gaussian designs. *IEEE Transactions on Information Theory*, Early Access.

Gonzalez, T. F. (1985). Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306.

Guntuboyina, A. (2012). Optimal rates of convergence for convex set estimation from support functions. *The Annals of Statistics*, 40(1):385–411.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer.

Han, Q. and Wellner, J. A. (2016). Multivariate convex regression: Global risk bounds and adaptation. *arXiv preprint arXiv:1601.06844v1*.

Hannah, L. A. and Dunson, D. B. (2013). Multivariate convex regression with adaptive partitioning. *Journal of Machine Learning Research*, 14(66):3261–3294.

Har-Peled, S. (2004). Clustering motion. *Discrete and Computational Geometry*, 31(4):545–565.

Hochbaum, D. S. and Shmoys, D. B. (1985). A best possible heuristic for the k-center problem. *Mathematics of Operations Research*, 10(2):180–184.

Kpotufe, S. (2009). Escaping the curse of dimensionality with a tree-based regressor. In *The 22nd Conference on Learning Theory*, Montreal, Quebec, Canada.

Kpotufe, S. (2010). *The curse of dimension in nonparametric regression*. PhD thesis, University of California.

Kpotufe, S. and Dasgupta, S. (2011). A tree-based regressor that adapts to intrinsic dimension. *Journal of Computer and System Sciences*, 78(5):1496–1515.

Kur, G., Gao, F., Guntuboyina, A., and Sen, B. (2020). Convex regression in multidimensions: Suboptimality of least squares estimators. *arXiv preprint arXiv:2006.02044v1*.

Lee, C.-Y., Johnson, A. L., Moreno-Centeno, E., and Kuosmanen, T. (2013). A more efficient algorithm for convex nonparametric least squares. *European Journal of Operational Research*, 227(2):391–400.

Lim, E. (2014). On convergence rates of convex regression in multiple dimensions. *INFORMS Journal of Computing*, 26(3):616–628.

Lim, E. and Glynn, P. W. (2012). Consistency of multidimensional convex regression. *Operations Research*, 60(1):196–208.

Magnani, A. and Boyd, S. P. (2009). Convex piecewise-linear fitting. *Optimization and Engineering*, 10(1):1–17.

Mazumder, R., Choudhury, A., Iyengar, G., and Sen, B. (2019). A computational framework for multivariate convex regression and its variants. *Journal of the American Statistical Association*, 114(525):318–331.

Pollard, D. (1990). Empirical processes: Theory and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 2:1–86.

Robbins, H. (1955). A remark on Stirling's formula. *The American Mathematical Monthly*, 62(1):26–29.

Seijo, E. and Sen, B. (2011). Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics*, 39(3):1633–1657.

Siahkamari, A., Gangrade, A., Kulis, B., and Saligrama, V. (2020a). Piecewise linear regression via a difference of convex functions. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8895–8904. PMLR.

Siahkamari, A., XIA, X., Saligrama, V., Castañón, D., and Kulis, B. (2020b). Learning to approximate a bregman divergence. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3603–3612. Curran Associates, Inc.

Stellato, B., Banjac, G., Goulart, P., Bemporad, A., and Boyd, S. (2020). OSQP: an operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12(4):637–672.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053.

van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

Supplementary Material: Adaptively Partitioning Max-Affine Estimators for Convex Regression

A Detailed proof of Theorem 4.1

This appendix provides the complete proofs for the lemmas of Section 4.1 and Theorem 4.1. It uses the notations of Section 4.1, and to improve readability, the steps ignore the scaling of the γ parameter by an appropriate absolute constant.

Recall that the result of APCNLS (Algorithm 2) is written as $f_n(\boldsymbol{x}) \doteq \max_{k\in[K]} \boldsymbol{a}_{n,k}^\top \boldsymbol{x} + b_{n,k}$ with parameters $\{(\boldsymbol{a}_{n,k}, b_{n,k}) \in \mathbb{R}^{d+1} : k \in [K]\}$, and $g_n(\boldsymbol{x}) \doteq \sum_{k\in[K]} \mathbb{I}\{\boldsymbol{x} \in \mathcal{S}_k\}(\boldsymbol{a}_{n,k}^\top \boldsymbol{x} + b_{n,k})$, where $g_n \in \mathcal{G}_n(\mathcal{P}_K, V_n)$ meaning that $0 \le f_n(\boldsymbol{x}_i) - g_n(\boldsymbol{x}_i) \le V_n$ for all $i \in [n]$. More generally, introduce the notation $\boldsymbol{a}_k(f), b_k(f)$, and $\boldsymbol{a}_k(g), b_k(g)$ to denote the parameters of the piecewise-linear functions $f \in \mathcal{H}_K$ and $g \in \mathcal{G}_n(\mathcal{P}_K, V_0)$, that is $f(\boldsymbol{x}) = \max_{k\in[K]} \boldsymbol{a}_k(f)^\top \boldsymbol{x} + b_k(f)$ and $g(\boldsymbol{x}) = \sum_{k\in[K]} \mathbb{I}\{\boldsymbol{x} \in \mathcal{S}_k\}\big(\boldsymbol{a}_k(g)^\top \boldsymbol{x} + b_k(g)\big)$. Clearly, $\boldsymbol{a}_{n,k} = \boldsymbol{a}_k(g_n)$ and $b_{n,k} = b_k(g_n)$ for all $k \in [K]$.

A.1 Proof of Lemmas 4.3, 4.4 and 4.5

First notice that the Lipschitzness of f_* and the subgaussian property of the observation noise (1) implies that $f_*(\boldsymbol{x})$ and y are also subgaussian random variables as given by Lemma A.1.

Lemma A.1. $\mathbb{E}\big[e^{|f_*(\boldsymbol{x})-\mathbb{E}[f_*(\boldsymbol{x})]|^2/(4L\rho)^2}\big]$ *and* $\mathbb{E}[e^{|y-\mathbb{E}[y]|^2/\xi^2}] \le 2$ *for* $\xi \doteq 2\max\{2L\rho, \sigma\}$.

Proof. Let $\widehat{\boldsymbol{x}}$ be an independent copy of $\boldsymbol{x}$, and use Jensen's inequality, the L-Lipschitzness of f_*, $(a+b)^2 \le 2(a^2+b^2)$, Jensen's inequality again, and (1) to get that

$$\mathbb{E}\big[e^{|f_*(\boldsymbol{x})-\mathbb{E}[f_*(\boldsymbol{x})]|^2/\xi_0^2}\big] \le \mathbb{E}\big[e^{|f_*(\boldsymbol{x})-f_*(\widehat{\boldsymbol{x}})|^2/\xi_0^2}\big] \le \mathbb{E}\big[e^{L^2\|\boldsymbol{x}-\widehat{\boldsymbol{x}}\|^2/\xi_0^2}\big] \le \mathbb{E}\big[e^{2L^2\|\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}]\|^2/\xi_0^2}\big]^2 \le 2,$$

with $\xi_0 \doteq 2L\rho$. Then, combine this with $(a+b)^2 \le 2(a^2+b^2)$, $\mathbb{E}[f_*(\boldsymbol{x})] = \mathbb{E}[y]$ by (1), the tower rule, Jensen's inequality, and (1) to show

$$\mathbb{E}\big[e^{|y-\mathbb{E}[y]|^2/\xi^2}\big] \le \mathbb{E}\Big[\mathbb{E}\big[e^{4|y-f_*(\boldsymbol{x})|^2/\xi^2}\big|\boldsymbol{x}\big]^{\frac{1}{2}} e^{2|f_*(\boldsymbol{x})-\mathbb{E}[f_*(\boldsymbol{x})]|^2/\xi^2}\Big] \le \sqrt{2}\,\mathbb{E}\big[e^{4|f_*(\boldsymbol{x})-\mathbb{E}[f_*(\boldsymbol{x})]|^2/\xi^2}\big]^{\frac{1}{2}} \le 2,$$

due to $\xi = 2\max\{\xi_0, \sigma\}$, which proves the claim. □

The next result (Lemma A.2) provides a data-dependent bound for the bias parameters of g_n.

Lemma A.2. $|b_{n,k} + \boldsymbol{a}_{n,k}^\top \bar{\boldsymbol{x}} - \bar{y}| \le 3\sqrt{d}L\rho_n$ *a.s. for all* $k \in [K]$.

Proof. Let $\widehat{\mathcal{G}}_n \doteq \{g \in \mathcal{G}_n(\mathcal{P}_K, V_n) : \|\boldsymbol{a}_k(g)\|_\infty \le L\}$, and notice that $g_n \in \widehat{\mathcal{G}}_n$. Additionally, as g_n is the solution to (3), it is a LSE over $\widehat{\mathcal{G}}_n$ satisfying $g_n \in \operatorname{argmin}_{g\in\widehat{\mathcal{G}}_n} \mathcal{L}_n(g)$. Because the set $\widehat{\mathcal{G}}_n$ is closed under constant shifting, it holds that $0 = \big[\partial_\lambda \mathcal{L}_n(g_n + \lambda)\big]_{\lambda=0}$ which implies $\bar{y} = \frac{1}{n}\sum_{i\in[n]} g_n(\boldsymbol{x}_i)$. Then, using $g_n(\boldsymbol{x}_i) = \boldsymbol{a}_{n,k}^\top \boldsymbol{x}_i + b_{n,k}$, the triangle and Jensen's inequalities, the L-Lipschitzness of g_n by $\|\boldsymbol{a}_{n,k}\|_\infty \le L$ with $\|\cdot\| \le \sqrt{d}\,\|\cdot\|_\infty$,

$$|b_{n,k} + \boldsymbol{a}_{n,k}^\top \bar{\boldsymbol{x}} - \bar{y}| = \Big|g_n(\boldsymbol{x}_i) - \boldsymbol{a}_{n,k}^\top(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) - \frac{1}{n}\sum_{j=1}^n g_n(\boldsymbol{x}_j)\Big| \le \|\boldsymbol{a}_{n,k}\|\,\rho_n + \frac{1}{n}\sum_{j=1}^n \big|g_n(\boldsymbol{x}_i) - g_n(\boldsymbol{x}_j)\big| \le 3\sqrt{d}L\rho_n,$$

which proves the claim. □

Notice that by the subgaussian property (1), (2), Lemma A.1, the Chernoff bound and Jensen's inequality, the following upper bounds hold with probability at least $1-\gamma$,

$$\rho_n \le \rho_\gamma, \quad \|f_* - y\|_n^2 = O\big(\sigma^2 \ln(1/\gamma)\big), \quad \|\bar{\boldsymbol{x}} - \mathbb{E}[\boldsymbol{x}]\|^2 = O\big(\rho^2 \ln(1/\gamma)\big), \quad |\bar{y} - \mathbb{E}[y]|^2 = O\big(\xi^2 \ln(1/\gamma)\big). \tag{8}$$

Then the combination of Lemma A.2, (8), and $\xi = O(L\rho + \sigma)$ yields that $|b_k(f_n) + \boldsymbol{a}_k(f_n)^\top \mathbb{E}[\boldsymbol{x}] - \mathbb{E}[y]|^2 = O\big(dL^2\rho_n^2 + (dL^2\rho^2 + \sigma^2)\ln(1/\gamma)\big) = O(\widehat{B}^2)$ with probability at least $1-\gamma$, which proves the claim of Lemma 4.3.

The log-covering number of a function class with bounded parameter space containing g_n is needed for bounding $\langle g_* - g_n, f_* - y\rangle_n$, so consider the bounds derived from (6) and Lemma A.2 as presented by Lemma A.3.

Lemma A.3. *Suppose that (8) holds for some $\gamma > 0$. Then for some $V_0 > 0$ satisfying $V_0^2 = O\big(\frac{\sigma^2 \ln(1/\gamma)}{d \ln(n)} + V_*^2\big)$, it holds with probability at least $1-\gamma$ that $g_n \in \overline{\mathcal{G}}_n$ for the function class $\overline{\mathcal{G}}_n \doteq \{g \in \mathcal{G}_n(\mathcal{P}_K, V_0) : \|\boldsymbol{a}_k(g)\|_\infty \le L,\ |b_k(g) - \boldsymbol{a}_k(g)^\top \mathbb{E}[\boldsymbol{x}] - \mathbb{E}[y]| \le \widehat{B},\ k \in [K]\}$.*

Proof. By (6), $ab \le (a^2+b^2)/2$, $\|f_* - y\|_n^2 = O\big(\sigma^2 \ln(1/\gamma)\big)$, $\beta = d\ln(n)$, and $\rho_n \le \rho_\gamma$, the bound $V_n^2 = O(V_0^2)$ follows, and so the claim follows as well by Lemma 4.3. □

Define the distance $\psi(g, \widehat{g}) \doteq \max_{k\in[K]} L^{-1}\|\boldsymbol{a}_k(g) - \boldsymbol{a}_k(\widehat{g})\| + \widehat{B}^{-1}\big|b_k(g) - b_k(\widehat{g}) + \big(\boldsymbol{a}_k(g) - \boldsymbol{a}_k(\widehat{g})\big)^\top \mathbb{E}[\boldsymbol{x}]\big|$ for any $g, \widehat{g} \in \overline{\mathcal{G}}_n$. By Lemma A.3, g_n lies in the class $\overline{\mathcal{G}}_n$ with probability at least $1-\gamma$, so we can condition on the event $g_n \in \overline{\mathcal{G}}_n$. Furthermore, by the volume argument (e.g., Pollard, 1990, Lemma 4.1) there exists an ϵ-cover of $\overline{\mathcal{G}}_n$ w.r.t. ψ of cardinality $N_\psi(\overline{\mathcal{G}}_n, \epsilon) \le \max\{1, (6\sqrt{d}/\epsilon)^{(d+1)K}\}$ for any $\epsilon > 0$. Additionally, for any $g, \widehat{g} \in \overline{\mathcal{G}}_n$ and $i \in [n]$ it holds that

$$\big|g(\boldsymbol{x}_i) - \widehat{g}(\boldsymbol{x}_i)\big| \le \sum_{k\in[K]} \mathbb{I}\{\boldsymbol{x}_i \in \mathcal{S}_k\}\Big|\big(\boldsymbol{a}_k(g) - \boldsymbol{a}_k(\widehat{g})\big)^\top \boldsymbol{x}_i + b_k(g) - b_k(\widehat{g})\Big| \le \psi(g,\widehat{g})\big(L\|\boldsymbol{x}_i - \mathbb{E}[\boldsymbol{x}]\| + \widehat{B}\big), \tag{9}$$

so $\max_{i\in[n]} |g(\boldsymbol{x}_i) - \widehat{g}(\boldsymbol{x}_i)| \le \psi(g,\widehat{g})(2L\rho_\gamma + \widehat{B})$ due to (2) and (8).

Let $\epsilon > 0$ to be chosen later, and $\overline{\mathcal{G}}_{n,\epsilon}$ be an ϵ-cover of $\overline{\mathcal{G}}_n$ w.r.t. ψ of minimal cardinality, and let $g_{n,\epsilon}$ be the closest element in $\overline{\mathcal{G}}_{n,\epsilon}$ to g_n, that is $g_{n,\epsilon} \in \operatorname{argmin}_{g\in\overline{\mathcal{G}}_{n,\epsilon}} \psi(g, g_n)$. Clearly, by the definition of ϵ-cover, $\psi(g_{n,\epsilon}, g_n) \le \epsilon$. Then by (9) and $\|f_* - y\|_n^2 = O\big(\sigma^2\ln(1/\gamma)\big)$ by (8), we have

$$\langle g_* - g_n, f_* - y\rangle_n \le \langle g_* - g_{n,\epsilon}, f_* - y\rangle_n + O\big(\epsilon\,(L\rho_\gamma + \widehat{B})\,\sigma\sqrt{\ln(1/\gamma)}\big). \tag{10}$$

Define the shorthand $\mathbb{P}_y\{\cdot\} \doteq \mathbb{P}\{\cdot|\mathcal{X}_n\}$, denote $\delta_i(g) \doteq \big(g_*(\boldsymbol{x}_i) - g(\boldsymbol{x}_i)\big)/\|g_* - g\|_n$, and let $t_0, t_1 > 0$ to be chosen later. Notice that the class $\overline{\mathcal{G}}_{n,\epsilon}$ only depends on the covariates $\mathcal{X}_n$ because AFPC (Algorithm 1) does not use the values $y_1, \ldots, y_n$ to compute the partition $\mathcal{P}_K$. Then by the union and Chernoff bounds, the independence of the samples $\{(\boldsymbol{x}_i, y_i) : i \in [n]\}$, the subgaussian property of the noise $f_*(\boldsymbol{x}_i) - y_i$ from (1) written as $\sup_{s\in\mathbb{R}} \mathbb{E}\big[e^{s(f_*(\boldsymbol{x}_i)-y_i) - 2s^2\sigma^2}\big|\boldsymbol{x}\big] \le 1$ a.s. (e.g., Boucheron et al., 2012, Section 2.3), $1 = \frac{1}{n}\sum_{i\in[n]} \delta_i^2(g)$ for any $g \in \overline{\mathcal{G}}_{n,\epsilon}$, it holds that

$$\begin{aligned}
\mathbb{P}_y\big\{\langle g_* - g_{n,\epsilon}, f_* - y\rangle_n > \|g_{n,\epsilon} - g_*\|_n\, t_0 t_1\big\} &\le \mathbb{P}_y\Big\{\sup_{g\in\overline{\mathcal{G}}_{n,\epsilon}} \Big\langle \frac{g_* - g}{\|g_* - g\|_n}, f_* - y\Big\rangle_n > t_0 t_1\Big\} \\
&\le \sum_{g\in\overline{\mathcal{G}}_{n,\epsilon}} \mathbb{P}_y\Big\{\frac{1}{n}\sum_{i=1}^n \delta_i(g)\big(f_*(\boldsymbol{x}_i) - y_i\big) > t_0 t_1\Big\} \\
&\le \sum_{g\in\overline{\mathcal{G}}_{n,\epsilon}} e^{-t_1} \prod_{i=1}^n \exp\Big(\frac{\delta_i(g)}{n t_0}\big(f_*(\boldsymbol{x}_i) - y_i\big)\Big) \\
&\le \sum_{g\in\overline{\mathcal{G}}_{n,\epsilon}} \exp\Big(\frac{2\sigma^2\sum_{i\in[n]}\delta_i^2(g)}{(n\,t_0)^2} - t_1\Big) \\
&= \big|\overline{\mathcal{G}}_{n,\epsilon}\big| \exp\Big(\frac{2\sigma^2}{n\,t_0^2} - t_1\Big) \\
&= \gamma,
\end{aligned} \tag{11}$$

by $t_1 \doteq 3\sigma^2/(nt_0^2)$ and $t_0 \doteq \sigma/\sqrt{n\ln(|\overline{\mathcal{G}}_{n,\epsilon}|/\gamma)}$. Then putting (10) and (11) together with $\ln|\overline{\mathcal{G}}_{n,\epsilon}| = \ln N_\psi(\overline{\mathcal{G}}_n,\epsilon) = O(dK\ln(d/\epsilon))$, and (9) with $\psi(g_n,g_{n,\epsilon}) \le \epsilon$, we get with probability at least $1-\gamma$ that

$$\begin{aligned}\langle g_* - g_n, f_* - y\rangle_n &= O\Big(\big(\|g_n - g_*\|_n + \epsilon(L\rho_\gamma + \widehat{B})\big)\sigma\sqrt{dK\ln\big(d/(\epsilon\gamma)\big)/n} + \epsilon\sigma(L\rho_\gamma+\widehat{B})\ln(1/\gamma)\Big)\\ &= O\Big(\|g_n - g_*\|_n\,\sigma\sqrt{\frac{dK}{n}\ln\big(nB_1/(\sigma\gamma)\big)} + \sigma^2\frac{dK}{n}\ln\big(nB_1/(\sigma\gamma)\big)\Big),\end{aligned}$$

with $\epsilon \doteq (\sigma/B_1)(dK/n)$ and $B_1 \doteq L\rho_\gamma + \widehat{B} = O(\widehat{B})$, which proves the claim of Lemma 4.4 by $n\widehat{B}/\sigma = O(B)$.

Next, combining (6) with Lemma 4.4 and $ab \le a^2 + b^2/4$ yields

$$\|g_n - g_*\|_n^2 + 4\beta V_n^2 = O\Big(\sigma^2\frac{dK}{n}\ln(B/\gamma) + \beta V_*^2\Big) = O\bigg(\frac{dK}{n}\Big(\sigma^2\ln(B/\gamma) + \ln(n)L^2\rho_\gamma^2\Big)\bigg)$$

due to $\beta = d\ln(n)$, $V_*^2 = O\big(L^2\Delta^2(\mathcal{P}_K)\big) = O(L^2\rho_\gamma^2 K/n)$ by Lemma 4.2, which proves both claims of Lemma 4.5.

A.2 Proof of Theorem 4.1

The proof goes similarly to the combination of Theorem 1 and Lemma 3 of Balázs et al. (2016), but here the $\mathcal{L}_2$-error is used instead of the excess risk.

As mentioned at the end of Section 4.1, the proof of Theorem 4.1 starts by conditioning on the events $\mathcal{X}_n \subseteq \mathcal{X}$, (8), and $f_n \in \overline{\mathcal{H}}_{\bar{K}}$ holding with probability at least $1-\gamma$ by the definition of $\mathcal{X}$, (1), Lemma A.1, and Lemma 4.3. Then it considers the decomposition of the $\mathcal{L}_2$-error as

$$\|f_n - f_*\|_\mu^2 \le r + \sup_{f\in\overline{\mathcal{H}}_{\bar{K}}(r)}\Big\{\|f-f_*\|_\mu^2 - 2\|f-f_*\|_n^2\Big\} + 2\|f_n - f_*\|_n^2, \tag{12}$$

where $\overline{\mathcal{H}}_{\bar{K}}(r) \doteq \{f \in \overline{\mathcal{H}}_{\bar{K}} : \|f-f_*\|_\mu^2 > r\}$ and $r \ge 0$ to be chosen later. As $\|f_n - f_*\|_n$ is bounded by (7), it remains to bound the first (supremum) term of (12) for an appropriate choice of r. For this, consider a slightly simplified version of the result of Balázs et al. (2016, Theorem 11) as stated by Lemma A.4.

Lemma A.4. *Let $(\mathcal{F},\psi)$ be a separable metric space, w be a random variable on some set $\mathcal{W}$, and $\Gamma : \mathcal{F}\times\mathcal{W} \to \mathbb{R}$ be a function. Define $\Lambda(f,w) \doteq \Gamma(f,w) - \mathbb{E}[\Gamma(f,w)]$ for all $f \in \mathcal{F}$, $w \in \mathcal{W}$, and suppose that the following conditions hold:*

(a) there exists scalars $\gamma > 0$, $T \ge 0$, and a function $\tau : \mathcal{W} \to [0,\infty)$ such that $\mathbb{P}\{\tau(w) > T\} \le \gamma/2$, and $\Lambda(f,w) - \Lambda(g,w) \le \psi(f,g)\,\tau(w)$ a.s. for all $f,g \in \mathcal{F}$ and $w \in \mathcal{W}$,

(b) there exists scalar $M > 0$ such that $\mathbb{E}\big[\exp\big(\Gamma(f,w)/M\big)\big] \le 1$ for all $f \in \mathcal{F}$.

Then for all $\epsilon \ge 0$, it holds with probability at least $1-\gamma$ that

$$\sup_{f\in\mathcal{F}}\Gamma(f,w) \le M\ln\big(3N_\psi(\mathcal{F},\epsilon)/\gamma\big) + 8T\epsilon.$$

Proof. See Theorem 11 of Balázs et al. (2016) with $S = \infty$ and $\delta = \epsilon$. □

In order to apply Lemma A.4, set $\mathcal{F} \doteq \overline{\mathcal{H}}_{\bar{K}}(r)$, $w \doteq \mathcal{X}_n \cup \{\boldsymbol{x}\}$, and $\Gamma(f,w) \doteq \|f-f_*\|_\mu^2 - 2\|f-f_*\|_n^2 = \frac{1}{n}\sum_{i\in[n]} w_i(f)$ with $w_i(f) \doteq \mathbb{E}\big[|f(\boldsymbol{x})-f_*(\boldsymbol{x})|^2\big] - 2|f(\boldsymbol{x}_i)-f_*(\boldsymbol{x}_i)|^2$ for all $i \in [n]$ and $f \in \mathcal{F}$. Then extend distance ψ to max-affine functions as $\psi(f,\widehat{f}) \doteq \max_{k\in[\bar{K}]} L^{-1}\|\boldsymbol{a}_k(f)-\boldsymbol{a}_k(\widehat{f})\| + \widehat{B}^{-1}\big|b_k(f)-b_k(\widehat{f}) + \big(\boldsymbol{a}_k(f)-\boldsymbol{a}_k(\widehat{f})\big)^\top\mathbb{E}[\boldsymbol{x}]\big|$ for any $f,\widehat{f} \in \mathcal{F}$, and notice that it still satisfies that

$$\big|f(\boldsymbol{x}) - \widehat{f}(\boldsymbol{x})\big| \le \max_{k\in[\bar{K}]}\big|\big(\boldsymbol{a}_k(f) - \boldsymbol{a}_k(\widehat{f})\big)^\top\boldsymbol{x} + b_k(f) - b_k(\widehat{f})\big| \le \psi(f,\widehat{f})\big(L\|\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]\| + \widehat{B}\big). \tag{13}$$

By writing any $f \in \overline{\mathcal{H}}_{\bar{K}}$ as $f(\boldsymbol{x}) = \max_{k\in[\bar{K}]}\boldsymbol{a}_k(f)^\top(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]) + \big(b_k(f) + \boldsymbol{a}_k(f)^\top\mathbb{E}[\boldsymbol{x}] - \mathbb{E}[y]\big) + \mathbb{E}[y]$ we have $|f(\boldsymbol{x})-f_*(\boldsymbol{x})|^2 - |\widehat{f}(\boldsymbol{x})-f_*(\boldsymbol{x})|^2 \le |f(\boldsymbol{x})-\widehat{f}(\boldsymbol{x})|\,\eta(\boldsymbol{x})$ with $\eta(\boldsymbol{x}) \doteq 2(\sqrt{d}L\|\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}]\| + \widehat{B}) + 2|f_*(\boldsymbol{x}) - \mathbb{E}[f_*(\boldsymbol{x})]|$ for any $f,\widehat{f} \in \overline{\mathcal{H}}_{\bar{K}}$ due to $\mathbb{E}[y] = \mathbb{E}[f_*(\boldsymbol{x})]$. Then by $\Lambda(f,w) = 2\big(\|f-f_*\|_\mu^2 - \|f-f_*\|_n^2\big)$ and (13), we also have

$$\Lambda(f,w) - \Lambda(\widehat{f},w) = 2\Big(\|f-f_*\|_\mu^2 - \|\widehat{f}-f_*\|_\mu^2\Big) + 2\Big(\|\widehat{f}-f_*\|_n^2 - \|f-f_*\|_n^2\Big) \le \psi(f,\widehat{f})\,\tau(w),$$

where $\tau(w) \doteq \frac{2}{n}\sum_{i\in[n]} \mathbb{E}[\eta^2(\boldsymbol{x})] + \eta^2(\boldsymbol{x}_i)$. As $\tau(w)$ satisfies $\mathbb{P}\{\tau(w) > T\} \leq \gamma/2$ for some $T > 0$ such that $T = O(\widehat{B}^2)$ due to (1), Lemma A.1, and $d(L\rho_\gamma)^2 = O(\widehat{B}^2)$, so the condition (a) of Lemma A.4 holds.

Now consider the following version of Bernstein's inequality for squared subgaussian random variables as presented by Lemma A.5.

Lemma A.5. *Let z be a subgaussian variable $\mathbb{E}[e^{z^2/\bar{\sigma}^2}] \leq 2$ for some $\bar{\sigma} > 0$. Then $\mathbb{E}[z^{2k}] \leq 2k!\,\mathbb{E}[z^2](c\,\bar{\sigma})^{2k-2}$ for any $k \in \mathbb{N}$ with $c \doteq 2\sqrt{\ln(3\mathbb{K}[z])}$ and kurtosis $\mathbb{K}[z] \doteq \mathbb{E}[z^4]/\mathbb{E}[z^2]^2$. Additionally, $\mathbb{E}\big[e^{\lambda(\mathbb{E}[z^2]-z^2)}\big] \leq \exp\big(4\lambda^2\mathbb{E}[z^2](c\,\bar{\sigma})^2/(1-(c\,\bar{\sigma})^2\lambda)\big)$ for all $\lambda \in (0, 1/(c\,\bar{\sigma})^2)$.*

Proof. The first claim is trivial for $k = 1$, so let $k \geq 2$. By the Cauchy-Schwartz inequality,

$$\begin{aligned}\mathbb{E}\big[z^{2k}\big] &= \mathbb{E}[z^{2k}\,\mathbb{I}\{|z| \leq c\,\bar{\sigma}\}] + \mathbb{E}[z^{2k}\,\mathbb{I}\{|z| > c\,\bar{\sigma}\}] \\ &\leq \mathbb{E}[z^2](c\,\bar{\sigma})^{2k-2} + \mathbb{E}[z^4]^{1/2}\,\mathbb{E}[z^{2(2k-2)}\,\mathbb{I}\{|z| > c\,\bar{\sigma}\}]^{1/2} \\ &\leq \mathbb{E}[z^2]\Big((c\,\bar{\sigma})^{2k-2} + \mathbb{K}[z]^{1/2}\,\mathbb{E}[z^{4(2k-2)}]^{1/4}\,\mathbb{P}\{|z| > c\,\bar{\sigma}\}^{1/4}\Big).\end{aligned} \tag{14}$$

By Lemma A.2 of Balázs (2016), $\mathbb{E}[z^{2s}] \leq 2(s/e)^s\,\bar{\sigma}^{2s}$ for $s = 2(2k-2)$, which can be combined with (14), $\mathbb{P}\{|z| > c\,\bar{\sigma}\} \leq \mathbb{E}\big[e^{z^2/\bar{\sigma}^2}\big]e^{-c^2} \leq 2e^{-c^2}$, $c > 2$ as $\mathbb{K}[z] \geq 1$, and $e(k/e)^k \leq k!$ for any $k \in \mathbb{N}$ (e.g., Robbins, 1955), we get

$$\mathbb{E}[z^{2k}] \leq \mathbb{E}[z^2](c\,\bar{\sigma})^{2k-2}\Big(1 + \mathbb{K}[z]^{1/2}\Big(4\big(\frac{4k-4}{e\,c^2}\big)^{4k-4}e^{-c^2}\Big)^{1/4}\Big) \leq \mathbb{E}[z^2](c\,\bar{\sigma})^{2k-2}\Big(1 + 2\,\mathbb{K}[z]^{1/2}(k-1)!\,e^{-c^2/4}\Big),$$

which proves the first claim on $\mathbb{E}[z^{2k}]$ by the definition of c and $1 + 2(k-1)! \leq 2k!$ for $k \geq 2$. Finally, the second claim on $\mathbb{E}[e^{\lambda(\mathbb{E}[z^2]-z^2)}]$ follows by the "standard version" of Bernstein's inequality (e.g., Boucheron et al., 2012, Theorem 2.10) using the first claim that $\mathbb{E}\big[|z^2|^k\big] \leq (k!/2)\big(4\,\mathbb{E}[z^2](c\,\bar{\sigma})^2\big)\big((c\,\bar{\sigma})^2\big)^{k-2}$. □

Let $z_f \doteq f(\boldsymbol{x}) - f_*(\boldsymbol{x})$ for a fixed $f \in \mathcal{F}$, and notice that if $f(\boldsymbol{x}) = \max_{k\in[\bar{K}]} \boldsymbol{a}_k^\top\boldsymbol{x} + b_k$, then $|f(\boldsymbol{x}) - f_*(\boldsymbol{x})| \leq \max_{k\in[\bar{K}]} \|\boldsymbol{a}_k\|\,\|\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]\| + |b_k + \boldsymbol{a}_k^\top\mathbb{E}[\boldsymbol{x}] - \mathbb{E}[y]| + |\mathbb{E}[y] - f_*(\boldsymbol{x})| \leq \sqrt{d}L\,\|\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]\| + \widehat{B} + |f_*(\boldsymbol{x}) - \mathbb{E}[f_*(\boldsymbol{x})]|$, so $\mathbb{E}\big[e^{z_f^2/t^2}\big] \leq e^{3\widehat{B}^2/t^2}\mathbb{E}\big[e^{12dL^2\|\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}]\|^2/t^2}\big]^{1/4}\mathbb{E}\big[e^{12|f_*(\boldsymbol{x})-\mathbb{E}[f_*(\boldsymbol{x})]|^2/t^2}\big]^{1/4} \leq 2$ for $t = 4\widehat{B}$ by $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, the Cauchy-Schwartz and Jensen's inequalities, and $\widehat{B}^2 = \Theta\big(dL^2\rho_\gamma^2 + \sigma^2\ln(1/\gamma)\big)$. Define $c_f \doteq 2\sqrt{\ln(3\mathbb{K}[z_f])}$, and use Bernstein's inequality (Lemma A.5) for z_f^2 with $\lambda = 2/M_f$ to get for all $i \in [n]$ that

$$\mathbb{E}\big[e^{w_i(f)/M_f}\big] = e^{-\mathbb{E}[z_f^2]/M_f}\,\mathbb{E}\big[e^{2(\mathbb{E}[z_f^2]-z_f^2)/M_f}\big] \leq \exp\left(\frac{\mathbb{E}[z_f^2]}{M_f}\Big(-1 + \frac{16(t\,c_f)^2}{M_f(1-2(t\,c_f)^2/M_f)}\Big)\right) = 1, \tag{15}$$

with $M_f \doteq 18(t\,c_f)^2 = O\big((\widehat{B}c_f)^2\big)$. Now choose $r \doteq t^2/n$, and notice that $\mathbb{E}[z_f^4] \leq 2(2/e)^2t^4$ due to Balázs (2016, Lemma A.2), so $\mathbb{K}[z_f] = O\big(t^4/r^2\big) = O(n^2)$ and we get $c_f^2 = O\big(\ln(n)\big)$ for all $f \in \overline{\mathcal{H}}_{\bar{K}}(r)$. Then by defining $M \doteq \max_{f\in\overline{\mathcal{H}}_{\bar{K}}(r)} M_f/n$ and using the independence of the random variables $w_1(f), \ldots, w_n(f)$, we get

$$\mathbb{E}\big[\exp\big(\Gamma(f,w)/M\big)\big] \leq \mathbb{E}\big[\exp\big(\Gamma(f,w)/(M_f/n)\big)\big] = \prod_{i\in[n]}\mathbb{E}\big[e^{w_i(f)/M_f}\big] \leq 1,$$

so condition (b) of Lemma A.4 holds with $M = O\big(\widehat{B}^2\ln(n)/n\big)$. Then by applying Lemma A.4, using $\ln N_\psi(\overline{\mathcal{H}}_{\bar{K}}, \epsilon) = O\big(d\bar{K}\ln(1/\epsilon)\big)$, and setting $\epsilon \doteq \bar{K}/n$, we get with probability at least $1-\gamma$ that

$$\begin{aligned}\sup_{f\in\overline{\mathcal{H}}_{\bar{K}}(r)} \|f - f_*\|_\mu^2 - 2\,\|f - f_*\|_n^2 &= O\Big(\frac{\widehat{B}^2\ln(n)\ln(N_\psi(\overline{\mathcal{H}}_{\bar{K}},\epsilon)/\gamma)}{n} + \widehat{B}^2\epsilon\Big) \\ &= O\Big(\frac{d\bar{K}\widehat{B}^2\ln(n)\ln(n/\gamma)}{n}\Big).\end{aligned} \tag{16}$$

Plugging in (7), (16), and $r = O(\widehat{B}^2/n)$ into (12), and using $K \leq \bar{K}$ with $\widehat{B}^2 = O\big(dL^2\rho_\gamma^2 + \sigma^2\ln(1/\gamma)\big)$, we get with probability at least $1-\gamma$ that

$$\|f_n - f_*\|_\mu^2 = O\Big(\frac{\widehat{B}^2}{n} + \frac{d\bar{K}\widehat{B}^2\ln(n)\ln(n/\gamma)}{n} + \frac{dK}{n}\big(\sigma^2\ln(B/\gamma) + \ln(n)L^2\rho_\gamma^2\big)\Big) = O\Big(\frac{d^2\bar{K}}{n}R_\mu\Big)$$

with $R_\mu = \big(L^2\rho_\gamma^2 + \sigma^2\ln(B/\gamma)\big)\ln(n)\ln(n/\gamma)$, which proves the claim of Theorem 4.1 by $\bar{K} = O(n^{d_\mathcal{X}/(2+d_\mathcal{X})})$.

B Further experiments

This appendix provides a few more experiments which could not fit the main text due to space limitations.

Figs. 7 to 10 complement the results of Section 5 by using the alternative piecewise-linear or quadratic choices for the regression function. Fig. 11 repeats the experiments for f_*^{tq} of Figs. 1 and 3 on a smaller domain (using $d = 5$ instead of $d = 10$). Finally, Fig. 12 provides results for the regression functions $f_*^{\mathrm{exp}}(\boldsymbol{x}) \doteq e^{\boldsymbol{x}^\top \mathbf{1}/d}$ with $\mathbf{1}$ being the full one vector of appropriate size, and $f_*^{\mathrm{plq}}(\boldsymbol{x}) \doteq \max\{f_*^{\mathrm{spl}}(\boldsymbol{x}), f_*^{\mathrm{tq}}(\boldsymbol{x})\}$.

The conclusion is again that while APCNLS uses smaller models than CNLS and so it slightly underfits complex functions like f_*^{plq} on Fig. 12, it is also more robust against overfitting for simpler targets like f_*^{exp} on Fig. 12, very noisy settings like on Fig. 8, and learning around a low-dimensional nonlinear manifold like on Fig. 10.

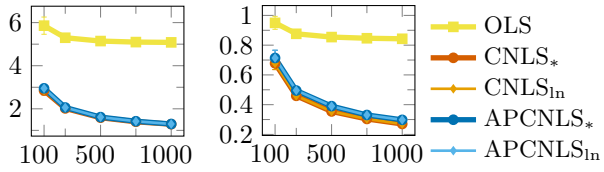

Figure 7: Estimating f_*^{sq} (left) over $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, I_{10})$, and f_*^{tpl} (right) over $\boldsymbol{x} \sim \mathcal{U}([-2,2]^{10})$, both for $\sigma = 0.3$.

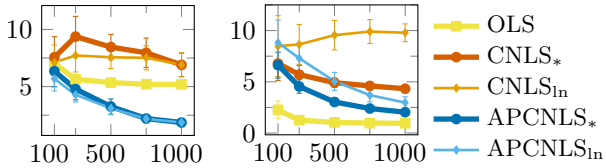

Figure 8: Estimating f_*^{sq} (left) over $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, I_{10})$, and f_*^{tpl} (right) over $\boldsymbol{x} \sim \mathcal{U}([-2,2]^{10})$, both for $\sigma = 3.0$.

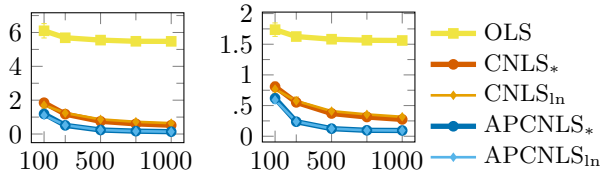

Figure 9: Estimating f_*^{sq} (left) and f_*^{tq} (right) over $\boldsymbol{x}_{\mathrm{li},10}$ for $\sigma = 0.3$ and $\sigma_m = 0.1$.

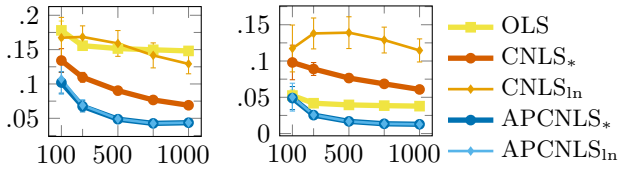

Figure 10: Estimating f_*^{spl} (left) and f_*^{tpl} (right) over $\boldsymbol{x}_{\mathrm{pe},10}$ for $\sigma = 0.3$ and $\sigma_m = 0.1$.

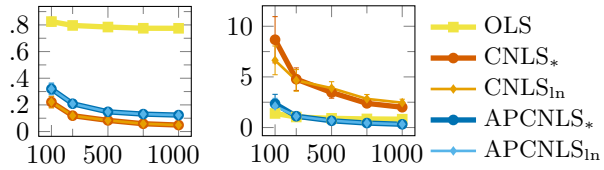

Figure 11: Estimating f_*^{tq} over $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, I_5)$ for $\sigma = 0.3$ (left) and $\sigma = 3.0$ (right).

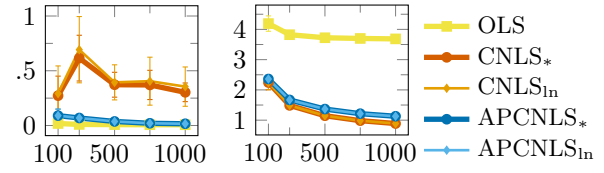

Figure 12: Estimating f_*^{exp} (left) and f_*^{plq} (right) over $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, I_{10})$ for $\sigma = 0.3$.