

# Near-optimal delta-convex estimation of Lipschitz functions

Gábor Balázs

G&G, CARTAGENA, SPAIN

GABALZ@GANDG.AI

## Abstract

This paper presents a tractable algorithm for estimating an unknown Lipschitz function from noisy observations and establishes an upper bound on its convergence rate. The approach extends max-affine methods from convex shape-restricted regression to the more general Lipschitz setting. A key component is a nonlinear feature expansion that maps max-affine functions into a subclass of delta-convex functions, which act as universal approximators of Lipschitz functions while preserving their Lipschitz constants. Leveraging this property, the estimator attains the minimax convergence rate (up to logarithmic factors) with respect to the intrinsic dimension of the data under squared loss and subgaussian distributions in the random design setting. The algorithm integrates adaptive partitioning to capture intrinsic dimension, a penalty-based regularization mechanism that removes the need to know the true Lipschitz constant, and a two-stage optimization procedure combining a convex initialization with local refinement. The framework is also straightforward to adapt to convex shape-restricted regression. Experiments demonstrate competitive performance relative to other theoretically justified methods, including nearest-neighbor and kernel-based regressors.

**Keywords:** nonparametric regression, Lipschitz function, squared loss, minimax rate, function approximation, delta-convex function, empirical risk minimization

## 1 Introduction

This paper considers the fundamental problem of estimating an unknown regression function from noisy observations in the random design setting. Suppose we observe  $n$  independent and identically distributed (i.i.d.) samples,  $\mathcal{D}_n \doteq \langle (\mathbf{X}_i, Y_i) : i \in [n] \rangle$ , for an unknown real-valued regression function  $f_* : \mathcal{X}_* \rightarrow \mathbb{R}$  over some unknown domain  $\mathcal{X}_* \subseteq \mathbb{R}^d$ , such that

$$\mathbf{X}_i \in \mathcal{X}_* \text{ almost surely (a.s.),} \quad Y_i \doteq f_*(\mathbf{X}_i) + \xi_i. \quad (1)$$

The noise  $\xi_i$  is centered, satisfying  $\mathbb{E}[\xi_i | \mathbf{X}_i] = 0$  a.s. for all  $i \in [n]$ , where  $[m] \doteq \{1, \dots, m\}$  for any positive integer  $m$ . We assume that the regression function  $f_*$  is  $\lambda_*$ -Lipschitz over  $\mathcal{X}_*$  with respect to (w.r.t.) the Euclidean norm  $\|\cdot\| \doteq \|\cdot\|_2$  for some unknown Lipschitz constant  $\lambda_* \in (0, \infty)$ . We evaluate the estimators using the excess risk under squared loss, for which the minimax (convergence) rate is known to be  $O(n^{-2/(2+d_*)})$  in terms of the sample size  $n$  and the intrinsic data dimension  $d_*$  (Stone, 1982). Throughout the paper, we use the standard asymptotic order of growth notations:  $\Omega(\cdot)$ ,  $\Theta(\cdot)$ , and  $O(\cdot)$ .

In convex (shape-restricted) regression (e.g., Lim, 2014; Han and Wellner, 2016; Balázs, 2016; Kur et al., 2024) the regression function  $f_*$  is known to be convex over a convex domain  $\mathcal{X}_*$ , and the goal is to estimate  $f_*$  by a convex function. In this setting, it is common to choose the estimator from the class of max-affine functions (functions defined as the maximum of finitely many affine functions) because they approximate any convex function at worst-case optimal rate (Balázs et al., 2015). Moreover, empirical risk minimization

over max-affine functions can be reformulated as a tractable convex optimization problem (solvable in polynomial-time w.r.t.  $d$  and  $n$ ; [Boyd and Vandenberghe, 2004](#), Section 6.5.5). Although several extensions of max-affine functions were proposed ([Bagirov et al., 2010](#); [Sun and Yu, 2019](#); [Siahkamari et al., 2020](#)), none have been shown to achieve the minimax rate up to logarithmic factors (i.e., near-minimax rate) in the general Lipschitz setting of (1). In this paper, we fill this gap by using the following extension of max-affine functions:

$$\mathcal{F}_{\triangleright}(\hat{\mathcal{X}}) \doteq \left\{ f : \mathbb{R}^d \rightarrow \mathbb{R} \mid f(\mathbf{x}) = \max_{k \in [k_0]} b_{f,k} + \mathbf{u}_{f,k}^\top (\mathbf{x} - \hat{\mathbf{x}}_k) + v_{f,k} \|\mathbf{x} - \hat{\mathbf{x}}_k\|_{\triangleright}, \right. \\ \left. \mathbf{x} \in \mathbb{R}^d, b_{f,k} \in \mathbb{R}, \mathbf{u}_{f,k} \in \mathbb{R}^d, v_{f,k} \in \mathbb{R}, k \in [k_0] \right\}, \quad (2)$$

where  $\hat{\mathcal{X}} \doteq \{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_{k_0}\} \subset \mathbb{R}^d$  is a nonempty, finite set of size  $k_0 \doteq |\hat{\mathcal{X}}|$ , and  $\|\cdot\|_{\triangleright}$  is a norm on  $\mathbb{R}^d$ .

The key observation, based on the function representation in the extension theorem of [McShane \(1934\)](#), is that when  $\hat{\mathcal{X}}$  is chosen to be an  $\epsilon$ -cover of the covariate data  $\mathcal{X}_n \doteq \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ , there exists a function within  $\mathcal{F}_{\triangleright}(\hat{\mathcal{X}})$  that is uniformly  $O(\epsilon)$ -close to the Lipschitz regression function  $f_*$  over  $\mathcal{X}_n$  (see Theorem 2). We use this fact to bound the approximation error of our estimators to  $f_*$  over  $\mathcal{X}_n$ , where the estimators “approximately” minimize the empirical risk over the training data  $\mathcal{D}_n$  within  $\mathcal{F}_{\triangleright}(\hat{\mathcal{X}})$ . A tradeoff arises in selecting the size of the cover  $\hat{\mathcal{X}}$ : increasing  $|\hat{\mathcal{X}}|$  improves approximation accuracy (i.e., smaller  $\epsilon$ ) but results in a more complex representation (i.e., more parameters), and vice versa. To balance this tradeoff, we construct  $\hat{\mathcal{X}}$  using the adaptive farthest-point clustering algorithm of [Balázs \(2022\)](#), which achieves  $|\hat{\mathcal{X}}| \approx n^{d_*/(2+d_*)}$  and  $\epsilon \approx n^{-1/(2+d_*)}$ . Our main result, stated in Theorem 1, shows that using this choice of  $\hat{\mathcal{X}}$  together with the class  $\mathcal{F}_{\triangleright}(\hat{\mathcal{X}})$  yields estimators that achieve a near-minimax rate under the statistical model (1) for subgaussian distributions.

Allowing the choice of norm via the parameter  $\triangleright$  provides flexibility in the design of the estimator. The most natural choice is the rotation invariant Euclidean norm, corresponding to  $\triangleright = 2$ . However, by choosing the max-norm with  $\triangleright = \infty$ , the elements of the set  $\{f \in \mathcal{F}_{\infty}(\hat{\mathcal{X}}) : v_{f,k} \leq 0, k \in [k_0]\}$  are max-min-affine functions (functions defined as the maximum of minima of finitely many affine functions), as discussed in Section 5.1. Since [Ovchinnikov \(2002\)](#) showed that max-min-affine functions can represent any continuous piecewise-linear function, there have been various developments of max-min-affine estimators ([Bagirov et al., 2010](#); [Toriello and Vielma, 2012](#); [Bagirov et al., 2022](#)). However, to the best of our knowledge, none of these come with theoretical guarantees. In this paper, we also address this gap by proposing a tractable max-min-affine estimator based on  $\mathcal{F}_{\infty}(\hat{\mathcal{X}})$ , which is proven to achieve a near-minimax rate under the nonparametric setting of (1) for subgaussian distributions, as shown in Theorem 1. In Section 2, we further discuss an extension to  $\mathcal{F}_1(\hat{\mathcal{X}})$ , based on the Manhattan norm (i.e.,  $\triangleright = 1$ ), which can be computed by maxout neural networks ([Goodfellow et al., 2013](#)).

Our algorithm regularizes the uniform Lipschitz constant of the estimator and guarantees a near-minimax rate in the general Lipschitz setting of (1), without requiring knowledge of the Lipschitz constant  $\lambda_*$  of the regression function  $f_*$ . In Section 6.2, we further show how our method can be easily adapted to the convex regression setting. Although [Blanchet et al. \(2019\)](#) proposed a similar uniform Lipschitz regularization for convex regression and

proved a convergence rate in probability, their result only holds for  $d > 4$  and for sufficiently large  $n$  satisfying  $\ln(n) \geq \lambda_*$ . [Lim \(2025\)](#) extended this result to  $d \leq 4$ , but it still only provides convergence rate in probability as  $n \rightarrow \infty$ . In contrast, we establish a probably approximately correct (PAC) bound that holds for all  $n$ . To the best of our knowledge, this adapted variant of our estimator is the first tractable method for convex regression to enjoy a PAC guarantee in the random design setting (albeit not necessarily near-minimax) without requiring knowledge of  $\lambda_*$  or a uniform bound on  $f_*$ .

The result of [Hiriart-Urruty \(1985, Section III.2\)](#) shows that the class  $\mathcal{F}_\triangleright(\hat{\mathcal{X}})$  lies within the class of delta-convex (DC) functions, whose elements can be expressed as the difference of two convex functions ([Hartman, 1959](#)). The classes of max-min-affine, weakly max-affine, and delta-max-affine functions are also contained within the class of DC functions, and all of them have been studied for estimator design (e.g., [Bagirov et al., 2010](#); [Sun and Yu, 2019](#); [Siahkamari et al., 2020](#)). However, none of these approaches has achieved a near-minimax rate guarantee in the setting of (1). In Sections 5.1 and 5.3, we provide approximation results for all of these classes, which may be of independent interest. To achieve the near-minimax rate in Theorem 1, we work with  $\mathcal{F}_\triangleright(\hat{\mathcal{X}})$  for two reasons. First, empirical risk minimization (ERM) over  $\mathcal{F}_\triangleright(\hat{\mathcal{X}})$  “leads” to a tractable convex optimization problem, as discussed in Section 4. In contrast, we are not aware of any tractable “approximation” of the ERM problem over the entire class of max-min-affine functions. Second, for the worst-case optimal approximation  $f \in \mathcal{F}_\triangleright(\hat{\mathcal{X}})$  to a  $\lambda_*$ -Lipschitz regression function  $f_*$ , the parameter magnitudes  $\max_{k \in [k_0]} \max\{\|\mathbf{u}_{f,k}\|, |v_{f,k}|\}$  are provably bounded above by  $O(\lambda_*)$ , as shown in Theorem 2. Importantly, this upper bound does not depend on the approximation accuracy. Since we cannot establish a similar bound for weakly max-affine and delta-max-affine functions, our proof technique does not apply to those cases in general (i.e., without further assuming smoothness of  $f_*$ ).

Finally, we note that several other methods have been proven to achieve a near-minimax rate in the regression setting of (1) with respect to the intrinsic data dimension. Specifically, the nearest-neighbor estimator ([Kulkarni and Posner, 1995, Corollary 3](#)), certain tree-based predictors ([Kpotufe and Dasgupta, 2011, Theorem 9](#)), and the Nadaraya-Watson regressor ([Kpotufe, 2010, Theorem 21](#)). We compare these methods to our algorithm through experiments in Section 3.

## 2 The proposed algorithm

Define the feature vector  $\phi_\triangleright(\mathbf{x}, \hat{\mathbf{x}}) \doteq [(\mathbf{x} - \hat{\mathbf{x}})^\top \|\mathbf{x} - \hat{\mathbf{x}}\|_\triangleright]^\top \in \mathbb{R}^{d+1}$  for all  $\triangleright \in \{1, 2, \infty\}$  and  $\mathbf{x}, \hat{\mathbf{x}} \in \mathbb{R}^d$ . Then, for any nonempty, finite set  $\hat{\mathcal{X}} \doteq \{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_{k_0}\} \subset \mathbb{R}^d$ , we can rewrite (2) in the compact form:

$$\begin{aligned} \mathcal{F}_\triangleright(\hat{\mathcal{X}}) = \{f : \mathbb{R}^d \rightarrow \mathbb{R} \mid f(\mathbf{x}) = \max_{k \in [k_0]} b_{f,k} + \mathbf{w}_{f,k}^\top \phi_\triangleright(\mathbf{x}, \hat{\mathbf{x}}_k), \\ \mathbf{x} \in \mathbb{R}^d, b_{f,k} \in \mathbb{R}, \mathbf{w}_{f,k} \in \mathbb{R}^{d_\triangleright}, k \in [k_0]\} \end{aligned}$$

where  $d_\triangleright \doteq d + 1$ . This formulation allows us to overload the notation and introduce the set  $\mathcal{F}_+(\hat{\mathcal{X}})$  by defining  $\phi_+(\mathbf{x}, \hat{\mathbf{x}}) \doteq [(\mathbf{x} - \hat{\mathbf{x}})_+^\top (\hat{\mathbf{x}} - \mathbf{x})_+^\top]^\top \in \mathbb{R}^{2d}$ , and  $d_+ \doteq 2d$ , where the ReLU operation  $(\mathbf{z})_+ \doteq \max\{\mathbf{0}_d, \mathbf{z}\}$  is applied elementwise to any vector  $\mathbf{z} \in \mathbb{R}^d$ , with  $\mathbf{0}_d$  denoting the zero vector of size  $d$ .

The functions of  $\mathcal{F}_+(\hat{\mathcal{X}})$  can be computed by the so-called maxout neural networks (Goodfellow et al., 2013). Further,  $\mathcal{F}_1(\hat{\mathcal{X}}) \subseteq \mathcal{F}_+(\hat{\mathcal{X}})$  holds because  $\mathbf{1}_{2d}^\top \phi_+(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\mathbf{x}}\|_1$  and  $[\mathbf{1}_d^\top - \mathbf{1}_d^\top] \phi_+(\mathbf{x}, \hat{\mathbf{x}}) = \mathbf{x} - \hat{\mathbf{x}}$  for all  $\mathbf{x}, \hat{\mathbf{x}} \in \mathbb{R}^d$ , where  $\mathbf{1}_s$  denotes the all-ones vector of size  $s$ . In the paper, we consider  $\mathcal{F}_\triangleright(\hat{\mathcal{X}})$  for all  $\triangleright \in \{1, 2, \infty, +\}$ , and discuss their approximation guarantees for Lipschitz functions in Section 4.1.

Let  $\{\mathcal{C}_1(\hat{\mathcal{X}}), \dots, \mathcal{C}_K(\hat{\mathcal{X}})\}$  denote the Voronoi partition of the entire space  $\mathbb{R}^d$  induced by the center points  $\hat{\mathcal{X}} \doteq \{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_{k_0}\} \subset \mathbb{R}^d$ . That is, for each  $k \in [k_0]$ , define  $\mathcal{C}_k(\hat{\mathcal{X}}) \doteq \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \hat{\mathbf{x}}_k\| = \min_{\hat{\mathbf{x}} \in \hat{\mathcal{X}}} \|\mathbf{x} - \hat{\mathbf{x}}\|\}$ , with ties broken arbitrarily but consistently (e.g., by selecting the center with the smaller index).

## 2.1 Delta-convex fitting (DCF)

Suppose we are given the training data  $\mathcal{D}_n$  from (1). First, we use the adaptive farthest-point clustering (AFPC; see Algorithm 1 below) method of Balázs (2022) to compute a finite set of distinct center points  $\hat{\mathcal{X}}_K \doteq \{\hat{\mathbf{X}}_1, \dots, \hat{\mathbf{X}}_K\} \subseteq \mathcal{X}_n$  for some  $K \in [n]$ . Then, the core of our algorithm is the following convex optimization problem:

$$\begin{aligned} \min_{\substack{z \in \mathbb{R}, \\ b_1, \dots, b_K \in \mathbb{R}, \\ \mathbf{w}_1, \dots, \mathbf{w}_K \in \mathbb{R}^{d_\triangleright}}} \quad & \theta_1 z^2 + \sum_{k \in [K]} \theta_2 \|\mathbf{w}_k\|^2 + \frac{1}{n} \sum_{i \in [n]} \mathbb{I}\{\mathbf{X}_i \in \mathcal{C}_k(\hat{\mathcal{X}}_K)\} \left( b_k + \mathbf{w}_k^\top \phi_\triangleright(\mathbf{X}_i, \hat{\mathbf{X}}_k) - Y_i \right)^2 \\ \text{such that } & b_k \geq b_l + \mathbf{w}_l^\top \phi_\triangleright(\hat{\mathbf{X}}_k, \hat{\mathbf{X}}_l), \quad \|\mathbf{w}_k\| \leq z + \theta_0, \quad k, l \in [K]. \end{aligned} \quad (3)$$

where  $\mathbb{I}\{\cdot\}$  is the  $\{0, 1\}$ -valued indicator function, and  $\theta_0, \theta_1, \theta_2 \geq 0$  are fixed regularization parameters. The role of  $\theta_0$  is to mitigate the effect of conservative regularization on the uniform Lipschitz constant of the estimator.

Let  $(z_n, \langle (b_{n,k}, \mathbf{w}_{n,k}) : k \in [K] \rangle)$  be a solution to (3), and define the (initial) estimator as  $f_n(\mathbf{x}) \doteq \max_{k \in [K]} b_{n,k} + \mathbf{w}_{n,k}^\top \phi_\triangleright(\mathbf{x}, \hat{\mathbf{X}}_k)$  for all  $\mathbf{x} \in \mathbb{R}^d$ . Clearly,  $f_n \in \mathcal{F}_\triangleright(\hat{\mathcal{X}}_K)$ .

The minimization problem in (3) is similar to the APCNLS algorithm of Balázs (2022), which is designed for training max-affine estimators. Like APCNLS, our approach trains a partitioning estimator and maps it into a target function class,  $\mathcal{F}_\triangleright(\hat{\mathcal{X}}_K)$  in our case, instead of the class of max-affine functions. A key distinction, however, is that we impose the constraints only at the center points, resulting in just  $K^2$  constraints, which is considerably fewer than the  $nK$  constraints used in APCNLS. As shown in Section 4.2, despite this reduced constraint set, our algorithm preserves theoretical properties analogous to those of APCNLS.

Denote the empirical risk of any function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  by  $\mathcal{L}_n(f) \doteq \frac{1}{n} \sum_{i \in [n]} (f(\mathbf{X}_i) - Y_i)^2$ . Additionally, for all  $f \in \mathcal{F}_\triangleright(\hat{\mathcal{X}}_K)$  and  $c_0, c_1, c_2 \geq 0$ , define the regularization term

$$\mathcal{R}_{c_0, c_1, c_2}(f) \doteq c_1 \max_{k \in [K]} (\|\mathbf{w}_{f,k}\| - c_0)_+^2 + c_2 \sum_{k \in [K]} \|\mathbf{w}_{f,k}\|^2,$$

and the largest slope parameter magnitude by  $\lambda_f \doteq \max_{k \in [K]} \|\mathbf{w}_{f,k}\|$ . We call any estimator  $\hat{f}_n^+ \in \mathcal{F}_\triangleright(\hat{\mathcal{X}}_K)$  to be a refinement of  $f_n$  that satisfies

$$\mathcal{L}_n(\hat{f}_n^+) + \mathcal{R}_n(\hat{f}_n^+) \leq \mathcal{L}_n(f_n) + \mathcal{R}_n(f_n), \quad \mathcal{R}_n(\cdot) \doteq \mathcal{R}_{\theta_3 \lambda_{f_n}, \theta_{f_n}, \theta_2}(\cdot), \quad (4)$$

where  $\theta_{f_n} \doteq \lambda_{f_n}^{-2}(\mathcal{L}_n(f_n) + \mathcal{R}_{0,0,\theta_2}(f_n))$  if  $\lambda_{f_n} > 0$ , and  $\theta_3 \geq 1$  is a fixed regularization parameter. Note that  $\lambda_{f_n} = \max_{k \in [K]} \|\mathbf{w}_{n,k}\|$  by definition. In the degenerate case, when  $\lambda_{f_n} = 0$ , we set  $\hat{f}_n^+ = f_n$  and  $\theta_{f_n} = 0$ .

Clearly, the choice  $\hat{f}_n^+ = f_n$  always satisfies (4). However, based on the experimental results in Section 3, we strongly recommend choosing  $\hat{f}_n^+$  as an approximate local solution to the non-convex optimization problem  $\min_{f \in \mathcal{F}_b(\hat{\mathcal{X}}_K)} \mathcal{L}_n(f) + \mathcal{R}_n(f)$ , initialized at  $f_n$ . Our near-minimax convergence rate guarantee in Theorem 1 holds in both cases.

To define the final estimator, we prune all parameters of  $\hat{f}_n^+$  which do not affect its empirical risk, and center its average response on  $\mathcal{X}_n$ . Formally, we define the final estimator for all  $\mathbf{x} \in \mathbb{R}^d$  as

$$\begin{aligned} f_n^+(\mathbf{x}) &\doteq C_n^+ + \max_{k \in \mathcal{I}_n^+} b_{\hat{f}_n^+,k} + \mathbf{w}_{\hat{f}_n^+,k}^\top \phi(\mathbf{x}, \hat{\mathbf{X}}_k), & \mathcal{I}_n^+ &\doteq \mathcal{I}_n(\hat{f}_n^+), \\ \mathcal{I}_n(f) &\doteq \{k \in [K] : f(\mathbf{X}_i) = b_{f,k} + \mathbf{w}_{f,k}^\top \phi(\mathbf{X}_i, \hat{\mathbf{X}}_k) \text{ for some } i \in [n]\}, \end{aligned} \quad (5)$$

where the index set  $\mathcal{I}_n(\cdot)$  is defined for all  $f \in \mathcal{F}_b(\hat{\mathcal{X}}_K)$ , and the centering constant  $C_n^+$  is given as  $C_n^+ \doteq \bar{Y} - \frac{1}{n} \sum_{i=1}^n \hat{f}_n^+(\mathbf{X}_i)$  with  $\bar{Y} \doteq \frac{1}{n} \sum_{i=1}^n Y_i$ . We also set  $b_{f_n^+,k} \doteq b_{\hat{f}_n^+,k} + C_n^+$  and  $\mathbf{w}_{f_n^+,k} \doteq \mathbf{w}_{\hat{f}_n^+,k}$  for all  $k \in \mathcal{I}_n^+$ . Besides potentially reducing inference-time computational costs, this step also facilitates bounding the magnitude of the unregularized bias parameters  $\{b_{f_n^+,k} : k \in \mathcal{I}_n^+\}$ , which is required for applying the concentration inequality used to prove the near-minimax rate of the estimator. We also extend  $\lambda_f$  and  $\mathcal{R}_n(f)$  to  $f = f_n^+$  by replacing  $[K]$  with  $\mathcal{I}_n^+$  in their definitions.

In Section 4, we show that the initial solution  $f_n$  already approximates  $\mathcal{L}_n(f_*) + \theta_1 \lambda_*^2$  with accuracy  $O(K/n)$ , which is upper bounded by the near-minimax rate. For the constant function,  $f_c^{\text{const}}(\mathbf{x}) \doteq c$  for all  $\mathbf{x} \in \mathbb{R}^d$  and  $c \in \mathbb{R}$ , the empirical risk  $\mathcal{L}_n(f_c^{\text{const}} + \hat{f}_n^+) = c^2 - 2cC_n^+ + \mathcal{L}_n(\hat{f}_n^+)$  is minimized at  $c = C_n^+$  with minimum value  $\mathcal{L}_n(f_n^+)$ , hence the centering in (5) ensures  $\mathcal{L}_n(f_n^+) \leq \mathcal{L}_n(\hat{f}_n^+)$ . Moreover, we have  $\mathcal{R}_n(f_n^+) \leq \mathcal{R}_n(\hat{f}_n^+)$  since the slope parameters of  $f_n^+$  are a subset of those of  $\hat{f}_n^+$ . Therefore, the estimator  $f_n^+$  is also a refinement of  $f_n$  that satisfies

$$\mathcal{L}_n(f_n^+) + \mathcal{R}_n(f_n^+) \leq \mathcal{L}_n(\hat{f}_n^+) + \mathcal{R}_n(\hat{f}_n^+) \leq \mathcal{L}_n(f_n) + \mathcal{R}_n(f_n). \quad (6)$$

We rely on (6) to show that  $f_n^+$  inherits the near-minimax rate guarantee of  $f_n$ . To this end, we also use the regularizer  $\mathcal{R}_n(\cdot)$  in (4), which limits the largest slope magnitude of  $\hat{f}_n^+$  (and hence of  $f_n^+$ ) from exceeding that of  $f_n$  by more than a constant factor of  $\theta_3$ . That is,  $\lambda_{f_n^+} \leq \lambda_{\hat{f}_n^+} = O(\theta_3 \lambda_{f_n})$ , as explained in Theorem 5.

We call our algorithm delta-convex fitting (DCF) and summarize it in Algorithm 2. For completeness, Algorithm 1 also presents the AFPC method of Balázis (2022). To describe it, let the covariate data radius be defined as  $R_{\mathcal{X}_n} \doteq \max_{i \in [n]} \|\mathbf{X}_i - \bar{\mathbf{X}}\|$  with  $\bar{\mathbf{X}} \doteq \frac{1}{n} \sum_{i \in [n]} \mathbf{X}_i$ . Define the clustering objective as  $\epsilon_n(\hat{\mathcal{X}}) \doteq \max_{\mathbf{X} \in \mathcal{X}_n} \min_{\hat{\mathbf{X}} \in \hat{\mathcal{X}}} \|\mathbf{X} - \hat{\mathbf{X}}\|$ , and the partition size limit as  $\hat{k}(\hat{\mathcal{X}}) \doteq \min \{n(\epsilon_n(\hat{\mathcal{X}})/R_{\mathcal{X}_n})^2, n^{d/(2+d)}\}$  for any set of center points  $\hat{\mathcal{X}} \subseteq \mathcal{X}_n$ .

**Algorithm 1**  $\hat{\mathcal{X}} \doteq \text{AFPC}(\mathcal{X}_n)$ 


---

```

1:  $\hat{\mathcal{X}} \leftarrow \{\hat{\mathbf{X}}\}$  with arbitrary  $\hat{\mathbf{X}} \in \mathcal{X}_n$ 
2: while  $|\hat{\mathcal{X}}| < \hat{k}(\hat{\mathcal{X}})$  do
3:    $\tilde{\mathbf{X}} \in \operatorname{argmax}_{\mathbf{X} \in \mathcal{X}_n} \min_{\hat{\mathbf{X}} \in \hat{\mathcal{X}}} \|\mathbf{X} - \hat{\mathbf{X}}\|$ 
4:    $\hat{\mathcal{X}} \leftarrow \hat{\mathcal{X}} \cup \{\tilde{\mathbf{X}}\}$ 
5: end while
6: return  $\hat{\mathcal{X}}$ 

```

---

**Algorithm 2**  $f_n^+ \doteq \text{DCF}(\mathcal{D}_n, \phi_{\triangleright})$ 


---

```

1:  $\hat{\mathcal{X}}_K \doteq \text{AFPC}(\mathcal{X}_n)$ ,  $K \doteq |\hat{\mathcal{X}}_K|$ 
2:  $(z_n, \{(b_{n,k}, \mathbf{w}_{n,k}) : k \in [K]\}) \leftarrow \text{solution to (3),}$   

   using  $\mathcal{D}_n$ ,  $\phi_{\triangleright}$ ,  $\hat{\mathcal{X}}_K$ , and  $\theta_0, \theta_1, \theta_2$ 
3:  $f_n(\cdot) \doteq \max_{k \in [K]} b_{n,k} + \mathbf{w}_{n,k}^\top \phi_{\triangleright}(\cdot, \hat{\mathbf{X}}_k)$ 
4:  $f_n^+ \leftarrow \text{refinement of } f_n \text{ via (4) and (5),}$   

   using  $\mathcal{D}_n$ ,  $\phi_{\triangleright}$ ,  $\hat{\mathcal{X}}_K$ ,  $f_n$ , and  $\theta_0, \theta_1, \theta_2, \theta_3$ 
5: return  $f_n^+$ 

```

---

The stopping condition of AFPC (Algorithm 1) ensures that  $K - 1 < \hat{k}(\hat{\mathcal{X}}_K) \leq K$ , which can be rearranged into the equation  $\epsilon_n^2(\hat{\mathcal{X}}_K) \approx R_{\mathcal{X}_n}^2 K/n$ . Using this, Balázs (2022, Lemma 4.2) showed that AFPC guarantees both a complexity bound of  $K = O(n^{d_*/(2+d_*)})$  and an accuracy bound of  $\epsilon_n^2(\hat{\mathcal{X}}_K) = O(K/n) = O(n^{-2/(2+d_*)})$ . These bounds balance the tradeoff between complexity and accuracy in our setting, as discussed under (2). The term  $n^{d/(2+d)}$  inside  $\hat{k}(\hat{\mathcal{X}})$  serves as a straightforward upper bound for the “worst-case” scenario when  $d_* \approx d$ . Furthermore, AFPC stops immediately if  $\epsilon_n(\hat{\mathcal{X}}) = 0$ , ensuring that it always produces distinct center points, justifying the representation of  $\hat{\mathcal{X}}$  as a set.

We consider feature maps  $\phi_{\triangleright}$  for all  $\triangleright \in \{1, 2, \infty, +\}$ , and analyze the DCF algorithm under the following choice of regularization parameters:

$$\begin{aligned} 0 \leq \theta_0 &= O((R_{\mathcal{Y}_n}/\max\{1, R_{\mathcal{X}_n}\}) \ln(n)), & \theta_1 &= \Theta(\max\{1, R_{\mathcal{X}_n}^2\} (dK/n)), \\ 0 \leq \theta_2 &\leq \theta_1/K, & 1 \leq \theta_3 &= O(\ln(n)), \end{aligned} \quad (7)$$

where the response data radius is defined as  $R_{\mathcal{Y}_n} \doteq \max_{i \in [n]} |Y_i - \bar{Y}|$  with  $\mathcal{Y}_n \doteq \{Y_1, \dots, Y_n\}$ , and  $K$  is the size of the AFPC-computed partition as defined in Algorithm 2.

## 2.2 Theoretical guarantees of DCF

Let  $\mathcal{B}(\mathbf{x}_0, r) \doteq \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{x}_0\| \leq r\}$  denote the closed ball in  $\mathbb{R}^d$  centered at  $\mathbf{x}_0 \in \mathbb{R}^d$  with radius  $r > 0$ . We write  $Z \sim P$  to indicate that the random variable  $Z$  is sampled from the distribution  $P$ . We consider the statistical model (1), where the regression function  $f_* : \mathcal{X}_* \rightarrow \mathbb{R}$  is  $\lambda_*$ -Lipschitz over its domain  $\mathcal{X}_* \subseteq \mathbb{R}^d$ .

Let  $d_{\circ}$  denote the *doubling dimension* of  $\mathcal{X}_*$  (e.g., Gupta et al., 2003). That is,  $d_{\circ}$  is the smallest number such that for any  $\mathbf{x} \in \mathbb{R}^d$  and  $r > 0$ , the set  $\mathcal{B}(\mathbf{x}, r) \cap \mathcal{X}$  can be covered by the union of at most  $2^{d_{\circ}}$  balls of radius  $r/2$ . Since  $d_{\circ}$  can exceed  $d$  by a constant factor,<sup>1</sup> we define the intrinsic dimension as  $d_* \doteq \min\{d_{\circ}, d\}$  to ensure that the convergence rate is bounded by  $n^{-2/(2+d)}$  in the worst-case when  $d \leq d_*$ . On the other hand, the doubling dimension  $d_{\circ}$  (and thus  $d_*$ ) can be significantly smaller than  $d$ , helping to mitigate the curse of dimensionality. Practical examples where this occurs include affine subspaces, Riemannian manifolds (Dasgupta and Freund, 2008, Theorem 22), sparse data, and unions of these (Kpotufe and Dasgupta, 2011, Lemmas 3 and 4).

Theorem 1 presents the main result of the paper, providing an adaptive near-minimax rate for DCF estimators with respect to (w.r.t.) the intrinsic dimension  $d_*$ .

---

<sup>1</sup>It is known that exactly 7 discs of radius 1/2 are needed to cover the unit disc (e.g., Zahn, 1962, Section I.2), so the doubling dimension for any set of positive area in  $\mathbb{R}^2$  is at least  $\log_2(7) > 2$ .



**Theorem 1** Consider the estimation problem (1), where the  $n$  i.i.d. samples  $\mathcal{D}_n$  are drawn from an unknown distribution  $P_*$  over  $\mathcal{X}_* \times \mathbb{R}$ , and the regression function  $f_*$  is  $\lambda_*$ -Lipschitz over  $\mathcal{X}_*$  w.r.t.  $\|\cdot\|$ . Suppose the covariate and noise distributions are subgaussian with parameters  $\rho, \sigma > 0$  such that

$$\mathbb{E}[e^{\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|^2/\rho^2}] \leq 2, \quad \mathbb{E}[e^{(f_*(\mathbf{X}) - Y)^2/\sigma^2} | \mathbf{X}] \leq 2 \text{ a.s.}, \quad (\mathbf{X}, Y) \sim P_*. \quad (8)$$

Let  $\triangleright \in \{1, 2, \infty, +\}$ , and  $f_n^+$  be the DCF estimator computed by Algorithm 2 using regularization parameters satisfying (7). Then, for all  $\gamma \in (0, 1)$ , it holds with probability at least  $1 - \gamma$  w.r.t. the randomness of the sample  $\mathcal{D}_n$  and the estimator (i.e., choosing the initial point of AFPC) that

$$\mathbb{E}_{(\mathbf{X}, \cdot) \sim P_*} [(f_n^+(\mathbf{X}) - f_*(\mathbf{X}))^2] = O\left(d(1 + d\mathbb{I}\{\triangleright \neq 2\}) n^{-2/(2+d_*)} \beta\right),$$

where  $\beta \doteq \theta_3^2(1 + \rho^2 \ln(n/\gamma))(\lambda_*^2 + \sigma^2 \ln(\beta_{\text{in}}/\gamma)) \ln^3(n) \ln(dn/\gamma)$ , and  $\beta_{\text{in}} \doteq n(1 + \lambda_* \rho/\sigma)$ .

**Proof** See Section 4.4. ■

Stone (1982, Theorem 1) showed that the minimax rate for the estimation problem (1) under the squared loss is  $\Omega(n^{-2/(2+d_*)})$  whenever  $[0, 1]^{d_*} \times \{0\}^{d-d_*} \subseteq \mathcal{X}_*$ . Therefore, the convergence rate established in Theorem 1 is near-minimax, since  $\beta = O(\ln^6(n))$ . Furthermore, our result provides a PAC bound, which can be converted into an expectation bound via integration (e.g., Balázs et al., 2016, Eq. 2).

The DCF algorithm can be adapted to use alternative function classes. As discussed in Section 6, it can be applied with the “complementary” set  $\mathcal{F}_{\triangleright}^-(\hat{\mathcal{X}}_K) \doteq \{f : -f \in \mathcal{F}_{\triangleright}(\hat{\mathcal{X}}_K)\}$  or the “symmetric” set  $\mathcal{F}_{\triangleright}^\Delta(\hat{\mathcal{X}}_K) \doteq \{f_1 - f_2 : f_1, f_2 \in \mathcal{F}_{\triangleright}(\hat{\mathcal{X}}_K)\}$ . The convergence rate established in Theorem 1 extends to both cases. In Section 5.1, we further describe how the DCF algorithm yields max-min-affine estimators and extend its theoretical guarantees to this setting. Finally, by restricting  $\mathcal{F}_{\triangleright}(\hat{\mathcal{X}}_K)$  to convex functions (including max-affine functions), the DCF algorithm specializes to convex regression, as discussed in Section 6.2. This generalization subsumes the APCNLS algorithm of Balázs (2022), matching its convergence-rate bound while eliminating the need to know the Lipschitz constant  $\lambda_*$  of the regression function.

Scaling all the elements of  $\mathcal{X}_n$  and  $\mathcal{Y}_n$  by the same positive constant can alter the slope variables  $\mathbf{w}_1, \dots, \mathbf{w}_K$  returned by the DCF algorithm. This is due to the  $\max\{1, R_{\mathcal{X}_n}\}$  terms in the definitions of the parameters  $\theta_0$  and  $\theta_1$  in (7). The positive lower bound on these parameters is necessary to keep the regularization active, thereby preserving the guarantee of Theorem 1 in degenerate cases where  $\mathbb{E}[\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|^2]$  converges to zero as  $n$  grows while the noise level  $\sigma > 0$  remains fixed. The choice of 1 as the lower bound can be relaxed, for example to  $1/\ln(n)$  at the cost of introducing additional  $\ln(n)$  factors in the bound of Theorem 1.

Finally, note that the DCF algorithm runs in polynomial-time with respect to both  $d$  and  $n$ . Since the AFPC algorithm differs from farthest-point clustering only in its stopping rule (using  $|\hat{\mathcal{X}}| < \hat{k}(\hat{\mathcal{X}})$  in Algorithm 1 instead of a fixed cardinality threshold), it can be computed in  $O(dKn)$  time (Gonzalez, 1985; Hochbaum and Shmoys, 1985). Constructing

the second-order cone program (SOCP) of (3) takes  $O(d^2n + K^2)$  time and yields a problem with  $O(dK)$  variables and  $O(K^2)$  constraints. Its solution can be approximated to accuracy  $\delta > 0$  using interior-point methods (e.g., Nesterov and Nemirovskii, 1994, Section 6.2), yielding a conservative worst-case runtime  $O(d^2n + d^2K^5 \ln(K/\delta))$ , ignoring sparsity. Lastly, the refinement step is optional with respect to theoretical guarantees and can be performed to a desired accuracy using smoothing techniques (Nesterov, 2005) in a tractable manner. Evaluating the DCF estimators  $f_n$  and  $f_n^+$  at a new point for inference takes  $O(dK)$  and  $O(d|\mathcal{I}_n^+|)$  time, respectively, where  $|\mathcal{I}_n^+| \leq K$ .

### 3 Experiments

We compared DCF (Algorithm 2) with other theoretically justified estimators that achieve near-minimax rates, namely the  $k$ -nearest neighbors estimator ( $k$ -NN; e.g., Györfi et al., 2002, Chapter 6) and the Nadaraya-Watson kernel regressor (NW; Nadaraya, 1964; Watson, 1964). As baselines, we also evaluated ordinary least squares regression (OLS) and state-of-the-art tree-based estimators, namely random forests (RF; Breiman, 2001) and the XGBoost gradient boosting algorithm (XGB; Chen and Guestrin, 2016).

We present experimental results on three public datasets from the Delve Project (University of Toronto).<sup>2</sup> The `cpusmall` (comp-active/cpuSmall) dataset consists of 8192 samples, where the task is to predict the portion of time that CPUs run in user mode based on 12 system activity measures. The `pumadyn` datasets also contain 8192 samples each and involve predicting the angular acceleration of one of the links of a simulated Puma 560 robot arm. We selected the versions designed for highly nonlinear estimation with 8 input dimensions and varying noise levels: `pumadyn-8nm` (medium noise) and `pumadyn-8nh` (high noise). Despite their small size, these problems effectively illustrate the strengths and weaknesses of DCF estimators.

For each experiment, we drew  $n \in \{1024, 2048, 4096\}$  training samples from the datasets and used the remaining data for evaluation, measuring the mean squared error (MSE) of the estimators on the test set. Each experiment was repeated 20 times, and we report the average results along with standard deviation error bars. All algorithms except the tree-based ones are sensitive to feature scaling, so we tested both min-max scaling (MM) and Z-score normalization (STD). The features of the `pumadyn` datasets are already reasonably scaled, so we also conducted experiments on these without applying any additional scaling (noFS). To ensure comparability across problems, we also standardized the response variables in each dataset by centering and scaling them to have unit variance.

Figure 1 shows the partition size and the average cell size distribution of the Voronoi partitions computed by AFPC. On the `cpusmall` dataset, AFPC terminates relatively early under both scaling methods, with roughly half as many cells in the STD case compared with MM. The `pumadyn` data includes noise in both the covariates and response variables, causing AFPC to return partition sizes close to the upper bound, even at the medium noise level. In all cases, AFPC produces many cells with fewer points than the domain dimension  $d$ , making regression underdetermined within those cells. Minimizing the slope of the estimator along unconstrained directions within such cells is an effective safeguard against overfitting. In DCF, this is enforced by the regularizer  $\theta_2 \sum_{k \in [K]} \|\mathbf{w}_k\|^2$  in (3), and

<sup>2</sup><https://www.cs.toronto.edu/~delve/data/datasets.html>



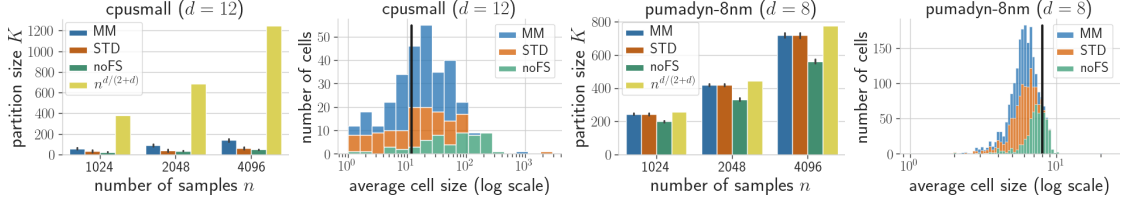


Figure 1: AFPC partition size ( $K$ ) for sample sizes  $n \in \{1024, 2048, 4096\}$ , and average cell size distribution for  $n = 4096$ . The upper bound of  $K$  is  $n^{d/(2+d)}$ . The black vertical lines on the average cell size axes mark the value of  $d$ . The plots for `pumadyn-8nh` are similar to those of `pumadyn-8nm` and are omitted for brevity.

analogously in (4), as has been done in convex regression practice (Aybat and Wang, 2016; Chen and Mazumder, 2024).

For the regression experiments, we used the default parameter settings for XGB and RF. The implementation of RF and  $k$ -NN were taken from scikit-learn (Pedregosa et al., 2011), while NW was implemented using scikit-fda (Ramos-Carreño et al., 2024). The number of neighbors for  $k$ -NN was selected via 5-fold cross-validation (CV) from the range 1 to  $\ln(n)n^{2/(2+d)}$  as motivated by Györfi et al. (2002, Theorem 6.2). For NW, we used the Gaussian and triweight kernels (referred to as NW-G and NW-T, respectively), selecting the bandwidth via 5-fold CV among 100 equidistant values up to  $(2R_{\mathcal{X}_n})^{d/(2+d)}(R_{\mathcal{Y}_n}^2/n)^{1/(2+d)}$  as motivated by Kpotufe (2010, Theorem 21).

DCF was implemented in Python, using local optimization for its refinement step as recommended in Section 2.1. We employed a quadratic penalty method (e.g., Nocedal and Wright, 2006, Section 17.1) with penalty parameter of  $10^6$  to transform the SOCP initialization problems (3) and (35), as well as the refinement steps (4) and (36), into unconstrained minimization problems, which were then solved using L-BFGS (e.g., Nocedal and Wright, 2006, Section 7.2).<sup>34</sup> The objective functions in the refinement steps were smoothed using the soft maximum approximation  $\max_{k \in [K]} \alpha_k \approx \mu \ln(\sum_{k \in [K]} e^{\alpha_k/\mu})$  for all  $\alpha_1, \dots, \alpha_K \in \mathbb{R}$ , with smoothing parameter  $\mu \doteq 10^{-6}$ .<sup>5</sup> Since the gradients of these objective functions are either not Lipschitz or have very large Lipschitz constants (on the order of  $1/\mu$ ), we stabilized the L-BFGS algorithm by reverting to a gradient step and resetting the L-BFGS memory whenever the backtracking line search failed. To further improve numerical stability, the training data were centered and scaled to unit variance.

We evaluated DCF using the sets  $\mathcal{F}_{\triangleright}(\cdot)$ ,  $\mathcal{F}_{\triangleright}^-(\cdot)$ , and  $\mathcal{F}_{\triangleright}^{\Delta}(\cdot)$  for norms  $\triangleright \in \{1, 2, \infty, +\}$ . We denote the corresponding estimators by  $\text{DCF}_{\triangleright}$ ,  $\text{DCF}_{\triangleright}^-$ , and  $\text{DCF}_{\triangleright}^{\Delta}$ , respectively. As the results for  $\triangleright \in \{1, 2\}$  were similar to those for  $\triangleright = \infty$ , we only report the latter in the plots. For reference, we denote by  $\text{i-DCF}_{+}^{\Delta}$  the initial estimator based on  $\mathcal{F}_{+}^{\Delta}(\cdot)$ , i.e., before applying the refinement step (36). We also compare against the max-min-affine

<sup>34</sup>Our implementation is available at <https://github.com/gabalz/cvxreg>.

<sup>4</sup>We also implemented the SOCP problems (3) and (35) using the Clarabel interior-point solver (Goulart and Chen, 2024), which yielded similar results but with significantly higher computational time.

<sup>5</sup>More precisely, smoothing was applied only to the gradient computations, which perturbs the objective by at most  $\mu \ln(K)$ , a quantity we ignored.

estimator based on  $\mathcal{F}_\infty(\cdot)$ , denoted MMA, and its symmetrized variant, denoted  $\text{MMA}^\Delta$ , which are described in Section 5.1. We use regularization parameters  $\theta_0 = (R_{Y_n}/R_{X_n}) \ln(n)$ ,  $\theta_1 = \max\{1, R_{X_n}^2\}(dK/n)$ ,  $\theta_2 \in \{(R_{X_n}/n)^2, R_{X_n}^2/n\}$ , and  $\theta_3 = \ln(n)$ , which satisfy (7). Unless otherwise indicated, we use the stronger regularizer  $\theta_2 = R_{X_n}^2/n$ .

The results are summarized in Figure 2. Across all problems, DCF (and MMA) using the symmetric sets ( $\text{DCF}_\Delta^\Delta$ ) achieved lower and more stable MSEs than their other variants with less expressive function representations. These symmetrized DCF variants performed

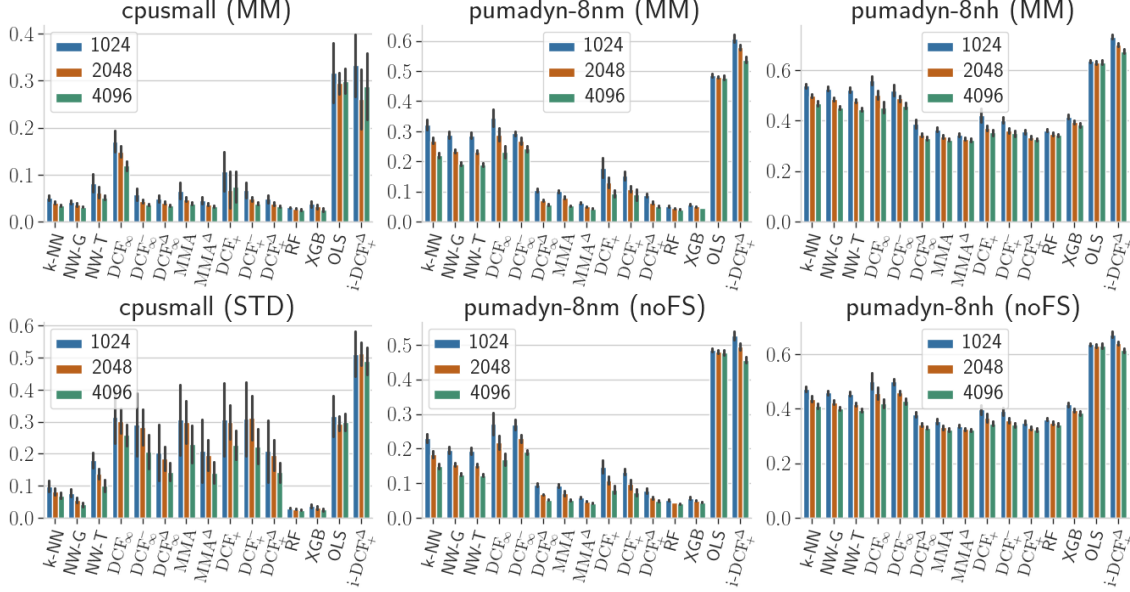


Figure 2: Test MSEs of the estimators trained on samples sizes  $n \in \{1024, 2048, 4096\}$ . The performance is very similar across all estimators for both MM and STD scalings of the **pumadyn** datasets; therefore, the plots for the latter are omitted for brevity.

at least as well as the other theoretically justified methods ( $k$ -NN and NW) on all problems, except for the **cpusmall** dataset with STD scaling. In that case, Figure 1 shows that AFPC stopped earlier, producing only half as many cells as in the MM case, which led DCF to underfit under the stronger regularizer  $\theta_2 = R_{X_n}^2/n$ . The tree-based methods (RF and XGB) performed quite well on these datasets, and DCF nearly matched their performance in many cases, unlike  $k$ -NN and NW. Since the initial estimator  $\text{i-DCF}_+^\Delta$  is forced to fit a continuous function only over the center points in (3) and (35), its performance was relatively poor, making the refinement step strongly recommended.

The training times of the estimators are shown in the left panel of Figure 3 for the **pumadyn** dataset,<sup>6</sup> which represents the least favorable case for DCF due to the large AFPC-computed partition size  $K$  (see Figure 1). On this dataset, the training time of the most expensive DCF variants ( $\text{DCF}_\infty^\Delta$  and  $\text{DCF}_+^\Delta$ ) is close to the training time of the NW algorithms (around 5 minutes for  $n = 4096$ ). The refinement step in training MMA estimators takes significantly longer, as they use  $d$  times more parameters than the corresponding DCF

<sup>6</sup>We used an Intel Core i5-2400S CPU with 4 cores at 2.50 GHz and ran two experiments in parallel.

variants. On the `cpusmall` dataset with MM scaling, where AFPC yields relatively few cells compared to the upper bound, DCF trains faster than NW (under 2 minutes for DCF, versus about 6 minutes for NW). The center panel of Figure 3 shows that the inference time of

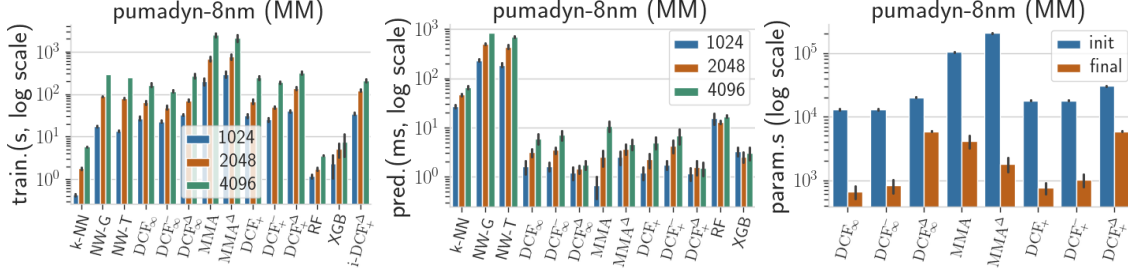


Figure 3: Training and prediction times (in seconds and milliseconds, respectively) are shown for the `pumadyn-8nm` dataset with MM scaling in the left and center panels. Prediction times are measured on the entire test set (whose size varies with  $n$ ) and normalized to 1000 samples. The right panel shows the number of parameters used by the initial and final DCF estimators,  $f_n$  and  $f_n^+$ , respectively.

DCF on the `pumadyn` dataset is quite fast. The right panel explains why: a large portion of DCF’s original parameters remain unused and are pruned in the final step of constructing the estimator, as described in (5) and (37), respectively.

Although Theorem 1 holds for all  $\theta_2 \in [0, \max\{1, R_{\mathcal{X}_n}^2\}(d/n)]$ , as defined in (7), Figure 4 shows that the practical performance of DCF is sensitive to this parameter. Using the

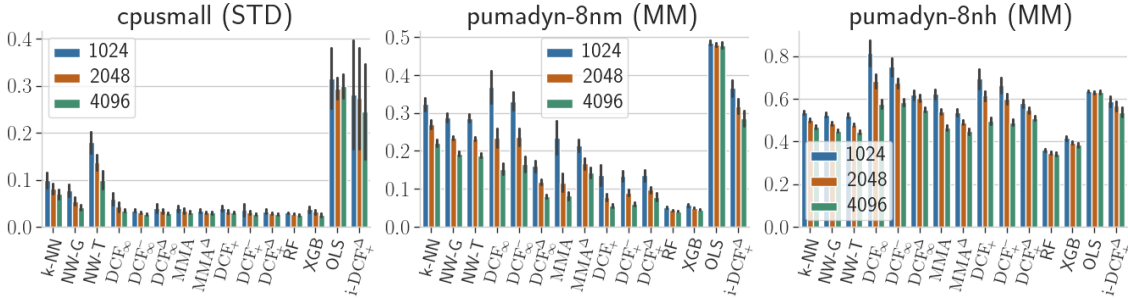


Figure 4: Test MSEs, using the same notations as above, where all DCF models are trained with the weaker regularizer  $\theta_2 = (R_{\mathcal{X}_n}/n)^2$ .

weaker regularizer  $\theta_2 = (R_{\mathcal{X}_n}/n)^2$ , the left panel shows that all DCF methods perform well on the `cpusmall` dataset with STD scaling (unlike using the stronger regularizer in Figure 2). However, as seen in the center and right panels, this choice allows the DCF algorithms to overfit on the `pumadyn` datasets, most notably on the noisier `pumadyn-8nh` variant shown in the right panel.

## 4 Analysis of DCF

This section is dedicated to the proof of Theorem 1. First, Section 4.1 presents our main approximation result in Theorem 2, which underpins the analysis of DCF (Algorithm 2). Then, in Section 4.2, we examine the properties of the DCF algorithm by relating it to empirical risk minimization. Finally, after briefly reviewing the guarantees of AFPC (Algorithm 1) in Section 4.3, we apply techniques from empirical process theory in Section 4.4 to establish the near-minimax rate.

### 4.1 Approximation of Lipschitz functions

For a metric space  $(\mathcal{Z}, \psi)$ , some  $\epsilon > 0$ , and some  $\mathcal{Z}_0 \subseteq \mathcal{Z}$ , the finite set  $\{\mathbf{z}_k \in \mathcal{Z}_0 : k \in [K]\}$  is called an (internal)  $\epsilon$ -cover of  $\mathcal{Z}_0$  w.r.t. the metric  $\psi$  if the union of the  $\epsilon$ -balls centered at  $\mathbf{z}_k$  covers  $\mathcal{Z}_0$ , that is  $\mathcal{Z}_0 \subseteq \cup_{k \in [K]} \{\mathbf{z} \in \mathcal{Z} : \psi(\mathbf{z}_k, \mathbf{z}) \leq \epsilon\}$ . The cardinality of the smallest such cover is called the  $\epsilon$ -covering number of  $\mathcal{Z}_0$  w.r.t.  $\psi$  and denoted by  $N_\psi(\mathcal{Z}_0, \epsilon)$ .

Let  $\mathcal{F}_{\lambda, \mathcal{X}}$  denote the class of  $\lambda$ -Lipschitz functions on a set  $\mathcal{X} \subseteq \mathbb{R}^d$  w.r.t.  $\|\cdot\|$ , defined as

$$\mathcal{F}_{\lambda, \mathcal{X}} \doteq \left\{ f : \mathcal{X} \rightarrow \mathbb{R} \mid \sup_{\mathbf{x}, \hat{\mathbf{x}} \in \mathcal{X}, \mathbf{x} \neq \hat{\mathbf{x}}} \|\mathbf{x} - \hat{\mathbf{x}}\|^{-1} (f(\mathbf{x}) - f(\hat{\mathbf{x}})) \leq \lambda \right\}.$$

McShane (1934, Theorem 1) showed that every  $\lambda$ -Lipschitz function  $f : \mathcal{X} \rightarrow \mathbb{R}$  on some  $\mathcal{X} \subset \mathbb{R}^d$  can be extended to  $\mathbb{R}^d$  via the function  $\tilde{f}(\cdot) \doteq \sup_{\hat{\mathbf{x}} \in \mathcal{X}} f(\hat{\mathbf{x}}) - \lambda \|\cdot - \hat{\mathbf{x}}\|$ , satisfying  $\tilde{f}(\mathbf{x}) = f(\mathbf{x})$  for all  $\mathbf{x} \in \mathcal{X}$ . The key observation in this paper is the uniform  $O(\lambda\epsilon)$  approximation bound in Theorem 2, which replaces the supremum in  $\tilde{f}$  with a maximum over a finite  $\epsilon$ -cover of  $\mathcal{X}$ .

**Theorem 2** *Let  $\mathcal{X}_\epsilon \subseteq \mathcal{X}$  be an  $\epsilon$ -cover of  $\mathcal{X} \subset \mathbb{R}^d$  w.r.t.  $\|\cdot\|$ , and  $\|\cdot\|_\triangleright$  be a norm on  $\mathbb{R}^d$  such that  $t_0 \|\cdot\|_\triangleright \leq \|\cdot\| \leq t_1 \|\cdot\|_\triangleright$  with some constants  $t_0, t_1 > 0$ . Suppose  $f \in \mathcal{F}_{\lambda, \mathcal{X}}$  for some Lipschitz constant  $\lambda > 0$ , and define  $\hat{f}(\mathbf{x}) \doteq \max_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} f(\hat{\mathbf{x}}) - t_1 \lambda \|\mathbf{x} - \hat{\mathbf{x}}\|_\triangleright$  for all  $\mathbf{x} \in \mathbb{R}^d$ . Then, for all  $\mathbf{x} \in \mathcal{X}$  and all  $\tilde{\mathbf{x}} \in \mathcal{X}_\epsilon$ , the following hold:*

$$0 \leq f(\mathbf{x}) - \hat{f}(\mathbf{x}) \leq (1 + t_0^{-1} t_1) \lambda \epsilon, \quad \hat{f}(\tilde{\mathbf{x}}) = f(\tilde{\mathbf{x}}), \quad \hat{f} \in \mathcal{F}_{(t_1/t_0)\lambda, \mathbb{R}^d}.$$

**Proof** Take any  $\mathbf{x} \in \mathcal{X}$  and  $\tilde{\mathbf{x}} \in \mathcal{X}_\epsilon$  arbitrarily. Since  $f \in \mathcal{F}_{\lambda, \mathcal{X}}$  and  $\|\cdot\| \leq t_1 \|\cdot\|_\triangleright$ , we get  $\hat{f}(\mathbf{x}) - f(\mathbf{x}) = \max_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} f(\hat{\mathbf{x}}) - f(\mathbf{x}) - t_1 \lambda \|\mathbf{x} - \hat{\mathbf{x}}\|_\triangleright \leq 0$ . Similarly,  $\hat{f}(\tilde{\mathbf{x}}) - f(\tilde{\mathbf{x}}) = \max_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} f(\hat{\mathbf{x}}) - f(\tilde{\mathbf{x}}) - t_1 \lambda \|\tilde{\mathbf{x}} - \hat{\mathbf{x}}\|_\triangleright \geq 0$ , hence  $\mathcal{X}_\epsilon \subseteq \mathcal{X}$  implies  $\hat{f}(\tilde{\mathbf{x}}) = f(\tilde{\mathbf{x}})$ . By the  $\epsilon$ -covering condition, we have  $\min_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} \|\mathbf{x} - \hat{\mathbf{x}}\| \leq \epsilon$ . Combining this with  $\|\cdot\|_\triangleright \leq t_0^{-1} \|\cdot\|$ , we obtain  $f(\mathbf{x}) - \hat{f}(\mathbf{x}) = \min_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} f(\mathbf{x}) - f(\hat{\mathbf{x}}) + t_1 \lambda \|\mathbf{x} - \hat{\mathbf{x}}\|_\triangleright \leq (1 + t_0^{-1} t_1) \lambda \epsilon$ . Moreover,  $\hat{f}$  is  $((t_1/t_0)\lambda)$ -Lipschitz on  $\mathbb{R}^d$  w.r.t.  $\|\cdot\|$  because the max function is 1-Lipschitz and  $\|\cdot\|_\triangleright$  is  $(t_0^{-1})$ -Lipschitz on  $\mathbb{R}^d$  w.r.t.  $\|\cdot\|$ . That is,  $\hat{f} \in \mathcal{F}_{(t_1/t_0)\lambda, \mathbb{R}^d}$ .  $\blacksquare$

Theorem 2 provides a lower approximation of  $f \in \mathcal{F}_{\lambda, \mathcal{X}}$  by a “max-concave” function  $\hat{f}$  that satisfies  $\hat{f} \leq f$  on  $\mathcal{X}$ . Similarly, one can construct an upper approximation of  $f$  using a “min-convex” function (Hiriart-Urruty, 1980), defined by  $\check{f}(\mathbf{x}) \doteq \min_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} f(\hat{\mathbf{x}}) + t_1 \lambda \|\mathbf{x} - \hat{\mathbf{x}}\|_\triangleright$  for all  $\mathbf{x} \in \mathbb{R}^d$ . We also present a variant of Theorem 2 with an improved approximation rate for smooth functions in Section 5.2.

The approximation error of  $\hat{f}$  and  $\check{f}$  w.r.t.  $f$  is zero at the points in  $\mathcal{X}_\epsilon$ , and increases proportionally with the distance from those points. Figure 5 illustrates several examples of these approximations. The functions  $\hat{f}$  and  $\check{f}$  provide an  $O(\lambda\epsilon)$  approximation rate of

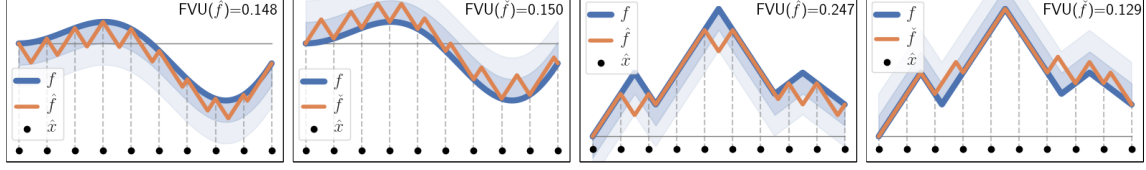


Figure 5: Approximation of a function  $f \in \mathcal{F}_{\lambda, \mathcal{X}}$  by the max-concave  $\hat{f}$  and the min-convex  $\check{f}$  as defined above. The left two plots use  $f(x) = x \sin(x)$ , while the right two plots use  $f(x) = \max\{1 - |x - 1|, 2 - |x - 3|, 1 - |x - 5|/2\}$ , both over  $\mathcal{X} = [0, 6]$ . The shaded regions represent  $\lambda\epsilon$  and  $2\lambda\epsilon$  bounds around  $f$ . Black circles mark the locations of the 10 equidistant centers  $\mathcal{X}_\epsilon$ , forming an  $\epsilon$ -cover of  $\mathcal{X}$  with  $\epsilon = 1/3$ . The horizontal line is shown at the height of zero. FVU (fraction of variance unexplained) is calculated over  $n = 1000$  equidistant points  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathcal{X}$  with  $y_i = f(\mathbf{x}_i)$  as:  $\text{FVU}(\hat{f}) \doteq \sum_{i \in [n]} (y_i - \hat{f}(\mathbf{x}_i))^2 / \sum_{i \in [n]} (y_i - \bar{y})^2$ , where  $\bar{y} \doteq (1/n) \sum_{i \in [n]} y_i$ .

the  $\lambda$ -Lipschitz function  $f$  while maintaining an  $O(\lambda)$  Lipschitz constant, as guaranteed by Theorem 2. Their appropriate variants (using the norm  $\|\cdot\|_\triangleright$ ) are included in the function sets  $\mathcal{F}_\triangleright(\hat{\mathcal{X}})$  and  $\mathcal{F}_\triangleright^-(\hat{\mathcal{X}})$  for  $\triangleright \in \{1, 2, \infty\}$ . Hence, these classes, as well as their superclasses such as  $\mathcal{F}_+(\hat{\mathcal{X}})$  and  $\mathcal{F}_\triangleright^\Delta(\hat{\mathcal{X}})$ , inherit the same guarantees.

## 4.2 Properties of DCF estimators

Let  $\triangleright \in \{1, 2, \infty, +\}$ ,  $k_0 \in \mathbb{N}$ , and for all nonempty, finite set  $\hat{\mathcal{X}} \doteq \{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_{k_0}\}$ , define the function class

$$\mathcal{G}_\triangleright(\hat{\mathcal{X}}) \doteq \left\{ g : \mathbb{R}^d \rightarrow \mathbb{R} \mid g(\mathbf{x}) \doteq \sum_{k \in [k_0]} \mathbb{I}\{\mathbf{x} \in \mathcal{C}_k(\hat{\mathcal{X}})\} h_{g,k}(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d, \right. \\ \left. h_{g,k}(\mathbf{x}) \doteq b_{g,k} + \mathbf{w}_{g,k}^\top \phi_\triangleright(\mathbf{x}, \hat{\mathbf{x}}_k), b_{g,k} \in \mathbb{R}, \mathbf{w}_{g,k} \in \mathbb{R}^{d_\triangleright}, k \in [k_0] \right\}.$$

Note that  $\mathcal{G}_\triangleright(\hat{\mathcal{X}})$  is a vector space over  $\mathbb{R}$ , since for any  $g_1, g_2 \in \mathcal{G}_\triangleright(\hat{\mathcal{X}})$  and  $c_1, c_2 \in \mathbb{R}$ , the function  $c_1 g_1 + c_2 g_2$  also belongs to  $\mathcal{G}_\triangleright(\hat{\mathcal{X}})$ , with  $b_{c_1 g_1 + c_2 g_2, k} = c_1 b_{g_1, k} + c_2 b_{g_2, k}$  and  $\mathbf{w}_{c_1 g_1 + c_2 g_2, k} = c_1 \mathbf{w}_{g_1, k} + c_2 \mathbf{w}_{g_2, k}$  for all  $k \in [k_0]$ .

Since the parameter spaces of  $\mathcal{F}_\triangleright(\hat{\mathcal{X}})$  and  $\mathcal{G}_\triangleright(\hat{\mathcal{X}})$  coincide, we can extend  $\mathcal{R}_{\theta_0, \theta_1, \theta_2}(f)$  and  $\lambda_f$  from Section 2.1, originally defined over  $\mathcal{F}_\triangleright(\hat{\mathcal{X}})$ , to any function  $f \in \mathcal{F}_\triangleright(\hat{\mathcal{X}}) \cup \mathcal{G}_\triangleright(\hat{\mathcal{X}})$ . We also define the map  $\pi : \mathcal{G}_\triangleright(\hat{\mathcal{X}}) \rightarrow \mathcal{F}_\triangleright(\hat{\mathcal{X}})$  by  $\pi(g)(\mathbf{x}) \doteq \max_{k \in [K]} h_{g,k}(\mathbf{x})$  for all  $\mathbf{x} \in \mathbb{R}^d$  and  $g \in \mathcal{G}_\triangleright(\hat{\mathcal{X}})$ . Clearly, for all  $g \in \mathcal{G}_\triangleright(\hat{\mathcal{X}})$ , we have  $b_{\pi(g), k} = b_{g, k}$  and  $\mathbf{w}_{\pi(g), k} = \mathbf{w}_{g, k}$  for all  $k \in [k_0]$ , and  $\pi(g) \geq g$  over  $\mathbb{R}^d$  holds since  $\mathcal{C}_1(\hat{\mathcal{X}}), \dots, \mathcal{C}_K(\hat{\mathcal{X}})$  are pairwise disjoint.

Consider the setting of Theorem 1, and let  $\hat{\mathcal{X}}_K = \{\hat{\mathbf{X}}_1, \dots, \hat{\mathbf{X}}_K\} \subseteq \mathcal{X}_n$  be the set of center points computed by AFPC (Algorithm 1). Since  $\phi_\triangleright(\hat{\mathbf{x}}, \hat{\mathbf{x}}) = \mathbf{0}_{d_\triangleright}$  for any  $\hat{\mathbf{x}} \in \mathbb{R}^d$  and  $\hat{\mathbf{X}}_k \in \mathcal{C}_k(\hat{\mathcal{X}}_K)$  for all  $k \in [K]$ , we obtain  $g(\hat{\mathbf{X}}_k) = h_{g,k}(\hat{\mathbf{X}}_k) = b_{g,k}$  for all  $g \in \mathcal{G}_\triangleright(\hat{\mathcal{X}}_K)$ .

and  $k \in [K]$ . Thereby, if  $b_{g,k} \geq h_{g,l}(\hat{\mathbf{X}}_k)$  for all  $k, l \in [K]$ , then  $g(\hat{\mathbf{X}}_k) \geq \pi(g)(\hat{\mathbf{X}}_k)$  for all  $k \in [K]$ , implying that the functions  $g$  and  $\pi(g)$  coincide over the set  $\hat{\mathcal{X}}_K$ . Therefore, the convex optimization problem (3) in DCF (Algorithm 2) is equivalent to the following regularized empirical risk minimization task:

$$\min_{g \in \mathcal{G}_{\triangleright}(\hat{\mathcal{X}}_K)} \mathcal{L}_n(g) + \mathcal{R}_{\theta}(g) \text{ such that } g = \pi(g) \text{ over } \hat{\mathcal{X}}_K, \quad (9)$$

where  $\mathcal{R}_{\theta}(\cdot) \doteq \mathcal{R}_{\theta_0, \theta_1, \theta_2}(\cdot)$ . Let  $g_n$  be a solution to (9) that matches the solution of (3) in parameters, i.e.,  $b_{g_n,k} = b_{n,k}$  and  $\mathbf{w}_{g_n,k} = \mathbf{w}_{n,k}$  for all  $k \in [K]$ . Then, by the definition of the initial DCF estimator  $f_n$  in Section 2.1, we have  $f_n = \pi(g_n)$ .

Recall that  $\hat{\mathcal{X}}_K$  is an  $\epsilon_n(\hat{\mathcal{X}}_K)$ -cover of the covariate data  $\mathcal{X}_n$  by definition. Then, the following result provides a uniform bound on the distance between  $f_n$  and  $g_n$  over  $\mathcal{X}_n$ :

**Lemma 3** *For all  $i \in [n]$ , it holds that  $0 \leq f_n(\mathbf{X}_i) - g_n(\mathbf{X}_i) \leq 2\lambda_{f_n}\epsilon_n(\hat{\mathcal{X}}_K)$ .*

**Proof** The lower bound follows directly from  $f_n = \pi(g_n) \geq g_n$ . Fix  $i \in [n]$  arbitrarily, and let  $k \in [K]$  be such that  $\mathbf{X}_i \in \mathcal{C}_k(\hat{\mathcal{X}}_K)$ . From (9), we have  $g_n = \pi(g_n)$  over  $\hat{\mathcal{X}}_K$ , which implies  $0 \leq h_{g_n,k}(\hat{\mathbf{X}}_k) - h_{g_n,l}(\hat{\mathbf{X}}_k)$  for all  $l \in [K]$ . By the definition of  $\pi$ , we also have  $h_{f_n,l}(\mathbf{x}) = h_{g_n,l}(\mathbf{x})$  for all  $l \in [K]$  and  $\mathbf{x} \in \mathbb{R}^d$ . Therefore:

$$\begin{aligned} f_n(\mathbf{X}_i) - g_n(\mathbf{X}_i) &= \max_{l \in [K]} h_{f_n,l}(\mathbf{X}_i) - h_{g_n,k}(\mathbf{X}_i) \\ &\leq \max_{l \in [K]} h_{g_n,l}(\mathbf{X}_i) - h_{g_n,l}(\hat{\mathbf{X}}_k) + h_{g_n,k}(\hat{\mathbf{X}}_k) - h_{g_n,k}(\mathbf{X}_i) \\ &\leq \max_{l \in [K]} (\|\mathbf{w}_{n,l}\| + \|\mathbf{w}_{n,k}\|) \|\mathbf{X}_i - \hat{\mathbf{X}}_k\|, \end{aligned}$$

where we used the Cauchy-Schwartz inequality in the last step. The claim then follows from the definitions of  $\lambda_{f_n}$  and  $\epsilon_n(\hat{\mathcal{X}}_K)$  using  $\mathbf{X}_i \in \mathcal{C}_k(\hat{\mathcal{X}}_K)$ .  $\blacksquare$

Theorem 2 bounds the uniform approximation error of  $\mathcal{F}_{\triangleright}(\hat{\mathcal{X}}_K)$  to the  $\lambda_*$ -Lipschitz regression function  $f_*$  by  $O(\lambda_*\epsilon_n(\hat{\mathcal{X}}_K))$ . The next result transfers this bound to  $\mathcal{G}_{\triangleright}(\hat{\mathcal{X}}_K)$ , thereby justifying its use in the ERM task (9), and allowing us to reformulate the problem as the tractable convex optimization task in (3).

**Lemma 4** *For  $f_* \in \mathcal{F}_{\lambda_*, \mathcal{X}_*}$ , there exists a function  $g_* \in \mathcal{G}_{\triangleright}(\hat{\mathcal{X}}_K)$  such that  $f_* \geq \pi(g_*) \geq g_*$  over  $\mathcal{X}_n$ , and*

$$\max_{i \in [n]} f_*(\mathbf{X}_i) - g_*(\mathbf{X}_i) \leq \tilde{\tau}_{\triangleright} \lambda_* \epsilon_n(\hat{\mathcal{X}}_K), \quad g_* = \pi(g_*) \text{ over } \hat{\mathcal{X}}_K, \quad \lambda_{g_*} \leq \tau_{\triangleright} \lambda_*,$$

where  $\tilde{\tau}_{\triangleright} \doteq 1 + \mathbb{I}\{\triangleright = 2\} + \mathbb{I}\{\triangleright \neq 2\}\sqrt{d}$ , and  $\tau_{\triangleright} \doteq \mathbb{I}\{\triangleright \in \{1, 2\}\} + \mathbb{I}\{\triangleright = \infty\}\sqrt{d} + \mathbb{I}\{\triangleright = +\}\sqrt{2d}$ .

**Proof** Fix the norm  $\triangleright \in \{1, 2, \infty\}$ , and define  $t_0 \doteq \mathbb{I}\{\triangleright \in \{2, \infty\}\} + \mathbb{I}\{\triangleright = 1\}/\sqrt{d}$  and  $t_1 \doteq \mathbb{I}\{\triangleright \neq \infty\} + \mathbb{I}\{\triangleright = \infty\}\sqrt{d}$ . Then,  $t_0\|\cdot\|_{\triangleright} \leq \|\cdot\| \leq t_1\|\cdot\|_{\triangleright}$ .

Let  $b_{*,k} \doteq f_*(\hat{\mathbf{X}}_k)$  and  $\mathbf{w}_{*,k} \doteq [\mathbf{0}_d^{\top} - t_1\lambda_*]^{\top}$  for all  $k \in [K]$ . Define  $\hat{f}_* \in \mathcal{F}_{\triangleright}(\hat{\mathcal{X}}_K)$  and  $g_* \in \mathcal{G}_{\triangleright}(\hat{\mathcal{X}}_K)$  such that  $b_{\hat{f}_*,k} = b_{g_*,k} = b_{*,k}$  and  $\mathbf{w}_{\hat{f}_*,k} = \mathbf{w}_{g_*,k} = \mathbf{w}_{*,k}$  for all  $k \in [K]$ . Then, we have  $\hat{f}_* = \pi(g_*) \geq g_*$ .



Let  $\epsilon \doteq \epsilon_n(\hat{\mathcal{X}}_K)$ . Since  $\hat{\mathcal{X}}_K$  is an  $\epsilon$ -cover of  $\mathcal{X}_n$ , we can apply Theorem 2 with  $f = f_*$ ,  $\hat{f} = \hat{f}_*$ ,  $\mathcal{X} = \mathcal{X}_n$ , and  $\mathcal{X}_\epsilon = \hat{\mathcal{X}}_K$ . This yields  $\hat{f}_*(\mathbf{X}_i) \leq f_*(\mathbf{X}_i)$  for all  $i \in [n]$ , proving  $f_* \geq \pi(g_*) \geq g_*$  over  $\mathcal{X}_n$ . Now fix  $i \in [n]$ , and let  $k \in [K]$  be such that  $\mathbf{X}_i \in \mathcal{C}_k(\hat{\mathcal{X}}_K)$ . By the definitions of  $g_*$  and  $\mathcal{C}_k(\hat{\mathcal{X}}_K)$ , we have  $g_*(\mathbf{X}_i) = f_*(\hat{\mathbf{X}}_k) - t_1 \lambda_* \|\mathbf{X}_i - \hat{\mathbf{X}}_k\|_\triangleright$ . Then, using  $f_* \in \mathcal{F}_{\lambda_*, \mathcal{X}_*}$ ,  $\mathcal{X}_n \subseteq \mathcal{X}_*$ , and  $\|\cdot\|_\triangleright \leq t_0^{-1} \|\cdot\|$ , we obtain

$$f_*(\mathbf{X}_i) - g_*(\mathbf{X}_i) = f_*(\mathbf{X}_i) - f_*(\hat{\mathbf{X}}_k) + t_1 \lambda_* \|\mathbf{X}_i - \hat{\mathbf{X}}_k\|_\triangleright \leq (1 + t_0^{-1} t_1) \lambda_* \|\mathbf{X}_i - \hat{\mathbf{X}}_k\|, \quad (10)$$

which implies  $f_*(\mathbf{X}_i) - g_*(\mathbf{X}_i) \leq \tilde{\tau}_\triangleright \lambda_* \epsilon$  for all  $i \in [n]$  as  $(1 + t_0^{-1} t_1) = \tilde{\tau}_\triangleright$  and  $\|\mathbf{X}_i - \hat{\mathbf{X}}_k\| \leq \epsilon$ . Additionally, setting  $\mathbf{X}_i = \hat{\mathbf{X}}_k$  in (10) and using  $f_* \geq g_*$  over  $\mathcal{X}_n$  yields  $g_*(\hat{\mathbf{X}}_k) = f_*(\hat{\mathbf{X}}_k)$  for all  $k \in [K]$ . Then, combining this with  $f_*(\hat{\mathbf{X}}_k) = \hat{f}_*(\hat{\mathbf{X}}_k)$  from Theorem 2, we obtain  $g_*(\hat{\mathbf{X}}_k) = f_*(\hat{\mathbf{X}}_k) = \hat{f}_*(\hat{\mathbf{X}}_k) = \pi(g_*)(\hat{\mathbf{X}}_k)$  for all  $k \in [K]$ , and thus  $g_* = \pi(g_*)$  over  $\hat{\mathcal{X}}_K$ . Finally,  $\lambda_{g_*} = \max_{k \in [K]} \|\mathbf{w}_{*,k}\| = t_1 \lambda_* = \tau_\triangleright \lambda_*$ , which proves the claim for all  $\triangleright \in \{1, 2, \infty\}$ .

The case  $\triangleright = +$  follows from the case  $\triangleright = 1$  by using  $\mathbf{w}_{*,k}^\top \phi_+(\mathbf{x}, \hat{\mathbf{x}}) = -\lambda_* \|\mathbf{x} - \hat{\mathbf{x}}\|_1$  for all  $\mathbf{x}, \hat{\mathbf{x}} \in \mathbb{R}^d$  with  $\mathbf{w}_{*,k} \doteq -\lambda_* \mathbf{1}_{2d}$ , which satisfies  $\|\mathbf{w}_{*,k}\| = \tau_\triangleright \lambda_*$  for all  $k \in [K]$ .  $\blacksquare$

Note that  $f_n$ ,  $\hat{f}_n^+$ , and  $f_n^+$  are Lipschitz continuous with Lipschitz constants bounded by  $\lambda_{f_n}$ ,  $\lambda_{\hat{f}_n^+}$ , and  $\lambda_{f_n^+}$ , respectively. In the refinement step (4), it is important to ensure that the Lipschitz constant of  $\hat{f}_n^+$  does not scale polynomially in the sample size  $n$ , as this would deteriorate the convergence rate of the DCF estimator. In Section 4.4.2, we prove that the Lipschitz constant of  $f_n$  grows only logarithmically with  $n$ . Therefore, we use  $\lambda_{f_n}$  as a reference for the Lipschitz constant of  $\hat{f}_n^+$ , and the following result shows that the Lipschitz constant of the refined estimators  $\hat{f}_n^+$  and  $f_n^+$  can exceed  $\lambda_{f_n}$  by at most an  $O(\theta_3)$  factor.

**Lemma 5** *Let  $\theta_3 \geq 1$ . Then,  $\lambda_{f_n^+} \leq \lambda_{\hat{f}_n^+} \leq (1 + \theta_3) \lambda_{f_n}$ .*

**Proof** By definition  $\lambda_{f_n^+} \leq \lambda_{\hat{f}_n^+}$ . Suppose, to the contrary, that  $\lambda_{\hat{f}_n^+} > (1 + \theta_3) \lambda_{f_n}$ . As  $\theta_3 \geq 1$  implies  $\mathcal{R}_n(f_n) = \mathcal{R}_{0,0,\theta_2}(f_n)$ , we get  $\mathcal{R}_n(\hat{f}_n^+) \geq \lambda_{f_n}^{-2} (\mathcal{L}_n(f_n) + \mathcal{R}_{0,0,\theta_2}(f_n)) (\lambda_{\hat{f}_n^+} - \theta_3 \lambda_{f_n})_+^2 > \mathcal{L}_n(f_n) + \mathcal{R}_n(f_n)$ , which contradicts (4), proving the claim.  $\blacksquare$

We will also need the following properties of the feature vector  $\phi_\triangleright$ : its output norm is bounded by a constant multiple of the Euclidean distance between its arguments, and  $\phi_\triangleright$  is Lipschitz in its second argument, as shown in the next result:

**Lemma 6** *Let  $\triangleright \in \{1, 2, \infty, +\}$ . Then for all  $\mathbf{x}, \hat{\mathbf{x}}, \tilde{\mathbf{x}} \in \mathbb{R}^d$ , the following inequalities hold:*

$$\|\phi_\triangleright(\mathbf{x}, \hat{\mathbf{x}})\| \leq \tau_{\phi_\triangleright} \|\mathbf{x} - \hat{\mathbf{x}}\|, \quad \|\phi_\triangleright(\mathbf{x}, \hat{\mathbf{x}}) - \phi_\triangleright(\mathbf{x}, \tilde{\mathbf{x}})\| \leq \lambda_{\phi_\triangleright} \|\hat{\mathbf{x}} - \tilde{\mathbf{x}}\|,$$

where  $\tau_{\phi_\triangleright} \doteq \sqrt{1 + \mathbb{I}\{\triangleright \in \{2, \infty\}\} + d \mathbb{I}\{\triangleright = 1\}}$ , and  $\lambda_{\phi_\triangleright} \doteq 1 + \mathbb{I}\{\triangleright \neq 1\} + \sqrt{d} \mathbb{I}\{\triangleright = 1\}$ .

**Proof** For  $\triangleright \in \{1, 2, \infty\}$ , the first claim follows from  $\|\phi_\triangleright(\mathbf{x}, \hat{\mathbf{x}})\|^2 = \|\mathbf{x} - \hat{\mathbf{x}}\|^2 + \|\mathbf{x} - \hat{\mathbf{x}}\|_\triangleright^2$ , and the inequalities  $\|\cdot\|_\infty \leq \|\cdot\|$  and  $\|\cdot\|_1 \leq \sqrt{d} \|\cdot\|$ . For  $\triangleright = +$ , we have  $\|\phi_+(\mathbf{x}, \hat{\mathbf{x}})\|^2 = \|(\mathbf{x} - \hat{\mathbf{x}})_+\|^2 + \|(\hat{\mathbf{x}} - \mathbf{x})_+\|^2 = \|\mathbf{x} - \hat{\mathbf{x}}\|^2$ .

For  $\triangleright \in \{1, 2, \infty\}$ , we have  $\phi_\triangleright(\mathbf{x}, \hat{\mathbf{x}}) - \phi_\triangleright(\mathbf{x}, \tilde{\mathbf{x}}) = [(\tilde{\mathbf{x}} - \hat{\mathbf{x}})^\top (\|\mathbf{x} - \hat{\mathbf{x}}\|_\triangleright - \|\mathbf{x} - \tilde{\mathbf{x}}\|_\triangleright)]^\top$ . From this, we prove the second claim using  $\|[\mathbf{u}^\top s]^\top\| \leq \|\mathbf{u}\| + |s|$  and the reverse triangle

inequality  $|\|\mathbf{u}\|_{\triangleright} - \|\mathbf{v}\|_{\triangleright}| \leq \|\mathbf{u} - \mathbf{v}\|_{\triangleright} \leq (\lambda_{\phi_{\triangleright}} - 1)\|\mathbf{u} - \mathbf{v}\|$ , which hold for all  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ ,  $s \in \mathbb{R}$ . For  $\triangleright = +$ , we have  $\phi_+(\mathbf{x}, \hat{\mathbf{x}}) - \phi_+(\mathbf{x}, \tilde{\mathbf{x}}) = \left[ \left( (\mathbf{x} - \hat{\mathbf{x}})_+ - (\mathbf{x} - \tilde{\mathbf{x}})_+ \right)^\top \left( (\hat{\mathbf{x}} - \mathbf{x})_+ - (\tilde{\mathbf{x}} - \mathbf{x})_+ \right)^\top \right]^\top$ , and the second claim follows by using  $\|[\mathbf{u}^\top \mathbf{v}^\top]^\top\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$  and  $\|(\mathbf{u})_+ - (\mathbf{v})_+\| \leq \|\mathbf{u} - \mathbf{v}\|$ , which hold for all  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ .  $\blacksquare$

The functions in  $\mathcal{F}_{\triangleright}(\hat{\mathcal{X}}_K)$  depend nonlinearly on the random center points  $\hat{\mathcal{X}}_K$ . In order to apply concentration inequalities, we exploit the Lipschitz property of  $\phi_{\triangleright}$  from Theorem 6, which enables us to “approximate” these random centers by fixed (non-random) ones in Section 4.4.3.

Since Theorems 3 and 4 both depend on  $\epsilon_n(\hat{\mathcal{X}}_K)$ , we first review the guarantees provided by AFPC for this quantity in the next section, before proceeding to the proof of Theorem 1.

### 4.3 AFPC guarantees

Recall the definition of the covering number from Section 4.1. We will often rely on the following well-known result on the covering number of bounded sets:

**Lemma 7 (e.g., Wainwright 2019, Lemma 5.7)** *Let  $\mathcal{Z}_0 \subseteq \mathcal{B}(\mathbf{z}_0, r) \subseteq \mathbb{R}^d$  for some  $d \in \mathbb{N}$ ,  $\mathbf{z}_0 \in \mathbb{R}^d$ , and  $r > 0$ . Then  $N_{\|\cdot\|}(\mathcal{Z}_0, \epsilon) \leq \max\{1, (3r/\epsilon)^d\}$  for all  $\epsilon > 0$ .*

Recall the definition of the covariate data radius  $R_{\mathcal{X}_n}$  from AFPC (Algorithm 1), and let  $d_o$  denote the doubling dimension of the domain  $\mathcal{X}_*$ , as introduced in Section 2.2. Because  $N_{\|\cdot\|}(\mathcal{X}_n, \epsilon) \leq \max\{1, (4R_{\mathcal{X}_n}/\epsilon)^{d_o}\}$  holds for all  $\epsilon > 0$  (e.g., Kpotufe and Dasgupta, 2011, Lemmas 6 and 7),<sup>7</sup> we can combine this with Theorem 7 to obtain  $N_{\|\cdot\|}(\mathcal{X}_n, \epsilon) \leq \max\{1, (4R_{\mathcal{X}_n}/\epsilon)^{d_*}\}$  for all  $\epsilon > 0$ . This allows us to apply the next result, which bounds the covering accuracy and the number of the center points returned by AFPC.

**Lemma 8 (Balázs 2022, Lemma 4.2)** *Suppose  $\hat{\mathcal{X}}$  is computed by AFPC (Algorithm 1) using the covariate data  $\mathcal{X}_n$ . Then there exists a (non-random) positive integer  $k_*$  such that  $K \doteq |\hat{\mathcal{X}}| \leq k_* = O(n^{d_*/(2+d_*)})$  a.s., and  $\hat{\mathcal{X}}$  is an  $\epsilon$ -cover of  $\mathcal{X}_n$  with  $\epsilon \doteq \epsilon_n(\hat{\mathcal{X}}) = O(R_{\mathcal{X}_n} \sqrt{K/n})$ .*

Combining the bounds on  $k_*$  and  $\epsilon$  from Theorem 8 yields  $\epsilon^2 = O(n^{-2/(2+d_*)})$ , which matches the minimax rate in the setting of Theorem 1.

### 4.4 Proof of the near-minimax rate of DCF

Consider the setting of Theorem 1, i.e., a regression problem as in (1) with an i.i.d. sample  $\mathcal{D}_n$  drawn from a distribution  $P_*$  corresponding to a regression function  $f_* \in \mathcal{F}_{\lambda_*, \mathcal{X}_*}$ .

To simplify notation, define  $\|f\|_*^2 \doteq \mathbb{E}_{(\mathbf{X}, \cdot) \sim P_*}[f^2(\mathbf{X})]$ ,  $\langle f, \hat{f} \rangle_n \doteq \frac{1}{n} \sum_{i \in [n]} f(\mathbf{X}_i) \hat{f}(\mathbf{X}_i)$ , and  $\|f\|_n^2 \doteq \langle f, f \rangle_n$  for all functions  $f, \hat{f} : \mathbb{R}^d \rightarrow \mathbb{R}$ . Further, in these contexts, we slightly abuse notation by treating  $y$  as a function, defining  $y(\mathbf{X}) \doteq Y$  for  $(\mathbf{X}, Y) \sim P_*$ , and  $y(\mathbf{X}_i) \doteq Y_i$  for all  $i \in [n]$ . E.g., we can write  $\mathcal{L}_n(f) = \|f - y\|_n^2$ ,  $\mathbb{E}_{(\mathbf{X}, Y) \sim P_*}[(f(\mathbf{X}) - Y)^2] = \|f - y\|_*^2$ , and  $\mathbb{E}_{(\mathbf{X}, \cdot) \sim P_*}[(f(\mathbf{X}) - f_*(\mathbf{X}))^2] = \|f - f_*\|_*^2$ .

<sup>7</sup>The definition of  $d_o$  in Section 2.2 does not require the covering balls to have centers within the covered set. Since we use internal covering numbers, our constant 4 in front of  $R_{\mathcal{X}_n}$  is looser by a factor of 2.

Let  $f_n^+$  be the estimator computed by DCF (Algorithm 2), and  $\tilde{c}_0 > 1$  be a constant. We decompose its expected squared error by following the approach of Györfi et al. (2002, Section 11.3):

$$\mathbb{E}_{(\mathbf{X}, \cdot) \sim P_*} [(f_n^+(\mathbf{X}) - f_*(\mathbf{X}))^2] = (\|f_n^+ - f_*\|_*^2 - \tilde{c}_0 E_{\text{approx}}) + \tilde{c}_0 E_{\text{approx}}, \quad (11)$$

where the approximation error is defined as  $E_{\text{approx}} \doteq \mathcal{L}_n(f_n^+) - \mathcal{L}_n(f_*)$ . The technique of offsetting with a factor greater than 1 allows the derivation of faster convergence rates in the nonparametric setting, where  $E_{\text{approx}}$  remains within the minimax rate. Using (6), we can upper bound  $E_{\text{approx}}$  as

$$\begin{aligned} E_{\text{approx}} &\leq \mathcal{L}_n(f_n) - \mathcal{L}_n(f_*) + \mathcal{R}_n(f_n) \\ &= \|f_n - y\|_n^2 - \|f_* - y\|_n^2 + \mathcal{R}_n(f_n) \\ &= \|f_n - g_n\|_n^2 + \|g_n - y\|_n^2 - \|f_* - y\|_n^2 + \mathcal{R}_n(f_n) \\ &\quad + 2\langle f_n - g_n, f_* - y \rangle_n + 2\langle f_n - g_n, g_n - f_* \rangle_n. \end{aligned} \quad (12)$$

Recall that  $g_*$  of Theorem 4 is considered in the optimization of (9) since  $g_* = \pi(g_*)$  holds over  $\hat{\mathcal{X}}_K$ . Because  $g_n$  is a solution to (9), it follows that

$$\mathcal{L}_n(g_n) + \mathcal{R}_\theta(g_n) \leq \mathcal{L}_n(g_*) + \mathcal{R}_\theta(g_*). \quad (13)$$

Furthermore, using  $2ab \leq a^2 + b^2$  for any  $a, b \in \mathbb{R}$ , we obtain

$$\begin{aligned} 2\langle f_n - g_n, g_n - f_* \rangle_n &= 2\langle f_n - g_n, g_* - f_* \rangle_n + 2\langle f_n - g_n, g_n - g_* \rangle_n \\ &\leq 2\|f_n - g_n\|_n^2 + \|g_* - f_*\|_n^2 + \|g_n - g_*\|_n^2. \end{aligned} \quad (14)$$

Notice that  $\theta_3 \geq 1$  implies  $\mathcal{R}_n(f_n) = \mathcal{R}_{0,0,\theta_2}(f_n)$ , and since  $f_n = \pi(g_n)$ , we also have  $\mathcal{R}_\theta(f_n) = \mathcal{R}_\theta(g_n)$ . Then, plugging (13) and (14) into (12), and using  $\mathcal{R}_n(f_n) = \mathcal{R}_{0,0,\theta_2}(f_n) \leq \mathcal{R}_\theta(f_n) = \mathcal{R}_\theta(g_n)$ , we get

$$\begin{aligned} E_{\text{approx}} &\leq \|f_n - g_n\|_n^2 + \|g_* - y\|_n^2 - \|f_* - y\|_n^2 + \mathcal{R}_\theta(g_*) \\ &\quad + 2\langle f_n - g_n, f_* - y \rangle_n + 2\langle f_n - g_n, g_n - f_* \rangle_n \\ &= \|f_n - g_n\|_n^2 + \|g_* - f_*\|_n^2 + 2\langle g_* - f_*, f_* - y \rangle_n + \mathcal{R}_\theta(g_*) \\ &\quad + 2\langle f_n - g_n, f_* - y \rangle_n + 2\langle f_n - g_n, g_n - f_* \rangle_n \\ &\leq 3\|f_n - g_n\|_n^2 + 2\|g_* - f_*\|_n^2 + \|g_n - g_*\|_n^2 + \mathcal{R}_\theta(g_*) \\ &\quad + 2\langle f_n - g_n, f_* - y \rangle_n + 2\langle g_* - f_*, f_* - y \rangle_n. \end{aligned} \quad (15)$$

Similarly to Balázs (2022, Section 4.1), using  $2ab = a(2b) \leq (a^2/2) + 2b^2$  for any  $a, b \in \mathbb{R}$ , we obtain

$$\begin{aligned} \mathcal{L}_n(g_*) - \mathcal{L}_n(g_n) &= \|g_* - y\|_n^2 - \|g_n - y\|_n^2 \\ &= -\|g_n - g_*\|_n^2 + 2\langle g_* - g_n, g_* - y \rangle_n \\ &= -\|g_n - g_*\|_n^2 + 2\langle g_* - g_n, g_* - f_* \rangle_n + 2\langle g_* - g_n, f_* - y \rangle_n \\ &\leq -(1/2)\|g_n - g_*\|_n^2 + 2\|g_* - f_*\|_n^2 + 2\langle g_* - g_n, f_* - y \rangle_n, \end{aligned}$$

which can be used to rearrange (13) as

$$(1/2)\|g_n - g_*\|_n^2 + \mathcal{R}_\theta(g_n) \leq 2\|g_* - f_*\|_n^2 + 2\langle g_* - g_n, f_* - y \rangle_n + \mathcal{R}_\theta(g_*). \quad (16)$$

Inequality (16) is an adaptation of the “basic inequality” of van de Geer (2000, Lemma 10.1), and it plays a key role in our analysis to bound the approximation error  $E_{\text{approx}}$  via (15).

Recall from Theorem 3 that  $\|f_n - g_n\|_n = O(\lambda_{f_n} \epsilon_n(\hat{\mathcal{X}}_K))$ . Additionally, we also have from Theorem 4 that  $\|g_* - f_*\|_n = O(\tilde{\tau}_\triangleright \lambda_* \epsilon_n(\hat{\mathcal{X}}_K))$  and  $\mathcal{R}_\theta(g_*) = O(\theta_1 \lambda_{g_*}^2) = O(\theta_1 \tau_\triangleright^2 \lambda_*^2)$  by  $\theta_1 \geq K\theta_2$  from (7). Combining these bounds with (15) and (16), using  $\mathcal{R}_\theta(g_n) \geq 0$ , we finally get

$$\begin{aligned} E_{\text{approx}} &= O\left((\lambda_{f_n}^2 + \tilde{\tau}_\triangleright^2 \lambda_*^2) \epsilon_n^2(\hat{\mathcal{X}}_K) + \theta_1 \tau_\triangleright^2 \lambda_*^2\right) \\ &\quad + 2\langle f_n - g_n, f_* - y \rangle_n + 2\langle g_* - f_*, f_* - y \rangle_n + 4\langle g_* - g_n, f_* - y \rangle_n. \end{aligned} \quad (17)$$

According to Theorem 1 of Stone (1982) and Theorem 8, the minimax rate is captured by  $K/n$ . This rate is also reflected in the asymptotic expression of (17), since both  $\epsilon_n^2(\hat{\mathcal{X}}_K)$  and  $\theta_1$  scale with  $K/n$ , as established by Theorem 8 and defined in (7), respectively.

It remains to show that the inner product terms of (17) and  $\|f_n^+ - f_*\|_*^2 - \tilde{c}_0 E_{\text{approx}}$  in (11) preserve the  $O(K/n)$  rate. To this end, we rely on concentration inequalities from empirical process theory.

#### 4.4.1 TECHNICAL PREPARATIONS

Since  $\mathbb{E}[\mathbf{X}]$  might not lie within  $\mathcal{X}_*$  for  $(\mathbf{X}, \cdot) \sim P_*$ , we need a reference point to leverage the Lipschitz continuity of  $f_*$ . To that end, fix  $\mathbf{x}_0 \in \arg\min_{\hat{\mathbf{x}} \in \mathcal{X}_*} \|\hat{\mathbf{x}} - \mathbb{E}[\mathbf{X}]\|$  independently of the data  $\mathcal{D}_n$ ,<sup>8</sup> and set  $y_0 \doteq f_*(\mathbf{x}_0)$ . For any fixed  $\gamma \in (0, 1)$ , we condition the entire proof on the event  $\mathcal{E}_\gamma$  defined in Theorem 9.

**Lemma 9** *Let  $\mathcal{D}_n$  be an i.i.d. subgaussian sample as in (8) for the regression function  $f_* \in \mathcal{F}_{\lambda_*, \mathcal{X}_*}$ . Fix  $r_\rho \doteq \rho \sqrt{\ln(2n/\gamma)}$  and  $r_\sigma \doteq \sigma \sqrt{\ln(2n/\gamma)}$ , and define the event*

$$\mathcal{E}_\gamma \doteq \left\{ \max_{i \in [n]} \|\mathbf{X}_i - \mathbb{E}[\mathbf{X}]\| \leq r_\rho, \max_{i \in [n]} |f_*(\mathbf{X}_i) - Y_i| \leq r_\sigma \right\}.$$

*Then  $\mathbb{P}\{\mathcal{E}_\gamma\} \geq 1 - 2\gamma$ . Furthermore,  $\mathcal{E}_\gamma$  implies  $R_{\mathcal{X}_n} \leq 2r_\rho$ ,  $R_{\mathcal{Y}_n} \leq 2(r_\sigma + \lambda_* r_\rho)$ , and*

$$\max_{i \in [n]} \|\mathbf{X}_i - \mathbf{x}_0\| \leq 2r_\rho, \quad \max_{i \in [n]} |Y_i - y_0| \leq r_\sigma + 2\lambda_* r_\rho, \quad \frac{R_{\mathcal{Y}_n}}{\max\{1, R_{\mathcal{X}_n}\}} \leq 2(r_\sigma + \lambda_*).$$

**Proof** The result  $\mathbb{P}\{\mathcal{E}_\gamma\} \geq 1 - 2\gamma$  follows from the subgaussian property (8) of  $P_*$ , using the union and Chernoff bounds. The implications of  $\mathcal{E}_\gamma$  follow from  $f_* \in \mathcal{F}_{\lambda_*, \mathcal{X}_*}$ ,  $\mathcal{X}_n \cup \{\mathbf{x}_0\} \subseteq \mathcal{X}_*$ , and  $y_0 = f_*(\mathbf{x}_0)$ , using the triangle and Jensen’s inequalities.  $\blacksquare$

Theorem 9 shows that, with high-probability, the data  $\mathcal{D}_n$  lies inside a ball of bounded radius centered at  $(\mathbf{x}_0, y_0)$ . This is needed for deriving upper bounds on the magnitudes of the estimator parameters.

<sup>8</sup>If the minimum is not attained, one may choose  $\mathbf{x}_0$  arbitrarily close to the infimum and shrink the gap to zero at the end of the analysis.

In the following sections, we work with parametric function sets and construct covers via their bounded parameter sets, as summarized in the next lemma:

**Lemma 10** *Let  $(\mathcal{F}, \psi)$  be a metric space, where  $\mathcal{F} \doteq \{f_{p_1, \dots, p_m} : p_j \in \mathcal{P}_j, j \in [m]\}$  with  $\mathcal{P}_j \subseteq \mathcal{B}(\mathbf{z}_j, r_j)$  for some  $t_j \in \mathbb{N}$ ,  $\mathbf{z}_j \in \mathbb{R}^{t_j}$ , and  $r_j > 0$  for all  $j \in [m]$ . Suppose there exist constants  $s_1, \dots, s_m > 0$  such that  $\psi(f_{p_1, \dots, p_m}, f_{\hat{p}_1, \dots, \hat{p}_m}) \leq \max_{j \in [m]} s_j \|p_j - \hat{p}_j\|$  for all  $p_j, \hat{p}_j \in \mathcal{P}_j$ ,  $j \in [m]$ . Then, for all  $\epsilon > 0$ , we have  $N_\psi(\mathcal{F}, \epsilon) \leq \max\{1, (3\beta/\epsilon)^t\}$ , where  $\beta \geq \max_{j \in [m]} s_j r_j$  and  $t \doteq \sum_{j \in [m]} t_j$ .*

**Proof** By using the conditions on  $\mathcal{F}$  and  $\psi$ , the result follows directly from Theorem 7 as  $N_\psi(\mathcal{F}, \epsilon) \leq \prod_{j=1}^m N_{\|\cdot\|}(\mathcal{P}_j, \epsilon/s_j) \leq \prod_{j=1}^m \max\{1, (3s_j r_j/\epsilon)^{t_j}\} \leq \max\{1, (3\beta/\epsilon)^t\}$ .  $\blacksquare$

We will make extensive use of the following concentration inequality, which generalizes Lemma 9 of Balázs (2022).

**Lemma 11** *Let  $\mathcal{D}_n$  be an i.i.d. subgaussian sample as in (8) for the regression function  $f_* : \mathbb{R}^d \rightarrow \mathbb{R}$ . Let  $\mathcal{H}_n \subseteq \{f : \mathbb{R}^d \rightarrow \mathbb{R}\}$ , and  $\psi_n$  be a metric on  $\mathcal{H}_n$  such that  $\|h - \hat{h}\|_n \leq \psi_n(h, \hat{h})$  holds for all  $h, \hat{h} \in \mathcal{H}_n$ . Assume that  $(\mathcal{H}_n, \psi_n)$  and  $\mathcal{Y}_n$  are conditionally independent given  $\mathcal{X}_n$ . Let  $h_n \in \mathcal{H}_n$  be a function, which may depend on the entire sample  $\mathcal{D}_n$ . Then, for any  $\gamma, \delta > 0$ , with probability at least  $1 - \gamma$  over the randomness of  $\mathcal{Y}_n | \mathcal{X}_n$ , it holds that*

$$\langle h_n, f_* - y \rangle_n \leq 3\sigma (\|h_n\|_n + \delta) \sqrt{\ln(N_{\psi_n}(\mathcal{H}_n, \delta)/\gamma)/n} + r_\sigma \delta.$$

**Proof** The claim is trivial when  $N_{\psi_n}(\mathcal{H}_n, \delta)$  is infinite, so suppose  $N_{\psi_n}(\mathcal{H}_n, \delta) < \infty$ . The claimed inequality always holds if  $\|h_n\|_n = 0$ . Therefore, without loss of generality, assume that  $\|h\|_n > 0$  for all  $h \in \mathcal{H}_n$ .

Let  $\mathcal{H}_\delta$  be a  $\delta$ -cover of  $\mathcal{H}_n$  w.r.t.  $\psi_n$  of minimal cardinality, so that  $|\mathcal{H}_\delta| = N_{\psi_n}(\mathcal{H}_n, \delta)$ . Note that  $\mathcal{H}_\delta$  and  $\mathcal{Y}_n$  are conditionally independent given  $\mathcal{X}_n$ . Further, leveraging the  $\delta$ -covering property, choose  $\hat{h}_n \in \mathcal{H}_\delta$  to be such that  $\|h_n - \hat{h}_n\|_n \leq \psi_n(h_n, \hat{h}_n) \leq \delta$ . By Theorem 9,  $\mathcal{E}_\gamma$  also implies  $\|f_* - y\|_n \leq r_\sigma$ . Then, by using the Cauchy-Schwartz inequality, we get

$$\langle h_n, f_* - y \rangle_n = \langle \hat{h}_n, f_* - y \rangle_n + \langle h_n - \hat{h}_n, f_* - y \rangle_n \leq \langle \hat{h}_n, f_* - y \rangle_n + r_\sigma \delta. \quad (18)$$

Define  $t_i(h) \doteq h(\mathbf{X}_i)/\|h\|_n$  for all  $i \in [n]$  and  $h \in \mathcal{H}_n$ , and let  $c_1, c_2 > 0$  be constants to be chosen later. Note that  $t_i(\cdot)$  only depends on  $\mathcal{X}_n$  for all  $i \in [n]$ . Then, using the union and Chernoff bounds, the independence of the samples  $\mathcal{D}_n$ , the subgaussian property (8) expressed as  $\sup_{s \in \mathbb{R}} \mathbb{E}[e^{s(f_*(\mathbf{X}_i) - Y_i) - 2s^2\sigma^2} | \mathbf{X}_i] \leq 1$  a.s. (Boucheron et al., 2013, Section 2.3),

and  $\sum_{i \in [n]} t_i^2(h) = n$  for any  $h \in \mathcal{H}_n$ , we obtain

$$\begin{aligned}
\mathbb{P}\left\{\langle \hat{h}_n, f_* - y \rangle_n > c_1 c_2 \|\hat{h}_n\|_n \mid \mathcal{X}_n\right\} &\leq \mathbb{P}\left\{\max_{\hat{h} \in \mathcal{H}_\delta} \left\langle \frac{\hat{h}}{\|\hat{h}\|_n}, f_* - y \right\rangle_n > c_1 c_2 \mid \mathcal{X}_n\right\} \\
&\leq \sum_{\hat{h} \in \mathcal{H}_\delta} \mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n t_i(\hat{h})(f_*(\mathbf{X}_i) - Y_i) > c_1 c_2 \mid \mathcal{X}_n\right\} \\
&\leq \sum_{\hat{h} \in \mathcal{H}_\delta} e^{-c_1} \prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{t_i(\hat{h})}{n c_2}(f_*(\mathbf{X}_i) - Y_i)\right) \mid \mathcal{X}_n\right] \quad (19) \\
&\leq \sum_{\hat{h} \in \mathcal{H}_\delta} \exp\left(\frac{2\sigma^2 \sum_{i \in [n]} t_i^2(\hat{h})}{(n c_2)^2} - c_1\right) \\
&= |\mathcal{H}_\delta| \exp\left(\frac{2\sigma^2}{n c_2^2} - c_1\right) \\
&= \gamma,
\end{aligned}$$

where we set  $c_1 \doteq 3\sigma^2/(n c_2^2)$  and  $c_2 \doteq \sigma/\sqrt{n \ln(|\mathcal{H}_\delta|/\gamma)}$ . Note that we also obtain  $c_1 c_2 = 3\sigma^2/(n c_2) = 3\sigma\sqrt{\ln(|\mathcal{H}_\delta|/\gamma)/n}$ . At last, the triangle inequality implies  $\|\hat{h}_n\|_n \leq \|h_n\|_n + \|\hat{h}_n - h_n\|_n \leq \|h_n\|_n + \delta$ , and we get the result by combining (18) and (19). ■

Finally, consider the following bound on the pointwise distance between functions in  $\mathcal{F}_\triangleright(\hat{\mathcal{X}})$  or  $\mathcal{G}_\triangleright(\hat{\mathcal{X}})$ , respectively.

**Lemma 12** *Let  $k_0 \in \mathbb{N}$  and  $\hat{\mathcal{X}} \doteq \{\mathbf{x}_1, \dots, \mathbf{x}_{k_0}\} \subset \mathbb{R}^d$ . Furthermore, let  $h, \hat{h} \in \mathcal{H}$  for  $\mathcal{H} \in \{\mathcal{F}_\triangleright(\hat{\mathcal{X}}), \mathcal{G}_\triangleright(\hat{\mathcal{X}})\}$ . Then it holds for all  $\mathbf{x} \in \mathbb{R}^d$  that*

$$|h(\mathbf{x}) - \hat{h}(\mathbf{x})| \leq \max_{k \in [k_0]} |b_{h,k} - b_{\hat{h},k}| + \tau_{\phi_\triangleright} \|\mathbf{x} - \hat{\mathbf{x}}_k\| \|\mathbf{w}_{h,k} - \mathbf{w}_{\hat{h},k}\|.$$

**Proof** Let  $f, \hat{f} \in \mathcal{F}_\triangleright(\hat{\mathcal{X}})$ , and  $g, \hat{g} \in \mathcal{G}_\triangleright(\hat{\mathcal{X}})$ . Then, we have

$$\begin{aligned}
|f(\mathbf{x}) - \hat{f}(\mathbf{x})| &\leq \max_{k \in [k_0]} \left| b_{f,k} - b_{\hat{f},k} + \phi_\triangleright(\mathbf{x}, \hat{\mathbf{x}}_k)^\top (\mathbf{w}_{f,k} - \mathbf{w}_{\hat{f},k}) \right|, \\
|g(\mathbf{x}) - \hat{g}(\mathbf{x})| &= \left| \sum_{k \in [k_0]} \mathbb{I}\{\mathbf{x} \in \mathcal{C}_k(\hat{\mathcal{X}})\} \left( b_{g,k} - b_{\hat{g},k} + \phi_\triangleright(\mathbf{x}, \hat{\mathbf{x}}_k)^\top (\mathbf{w}_{g,k} - \mathbf{w}_{\hat{g},k}) \right) \right|.
\end{aligned}$$

From Theorem 6, we also get  $\phi_\triangleright(\mathbf{x}, \hat{\mathbf{x}}_k) \leq \tau_{\phi_\triangleright} \|\mathbf{x} - \hat{\mathbf{x}}_k\|$  for all  $k \in [k_0]$ . Then, the claim for  $\mathcal{F}_\triangleright(\hat{\mathcal{X}})$  follows from the triangle and the Cauchy-Schwartz inequalities. The claim for  $\mathcal{G}_\triangleright(\hat{\mathcal{X}})$  follows similarly, where we also upper bound the sum by a max using that  $\mathcal{C}_1(\hat{\mathcal{X}}), \dots, \mathcal{C}_{k_0}(\hat{\mathcal{X}})$  are disjoint sets. ■



#### 4.4.2 BOUNDING THE APPROXIMATION ERROR

The goal of this section is to bound the approximation error  $E_{\text{approx}}$ . To achieve this, we use (17) and upper bound its inner product terms.

Let  $g_n$  be as in Section 4.2, and recall that the initial DCF estimator satisfies  $f_n = \pi(g_n)$ . Hence,  $f_n$  and  $g_n$  share the same parameters, and we have  $b_{n,k} = b_{f_n,k} = b_{g_n,k}$  and  $\mathbf{w}_{n,k} = \mathbf{w}_{f_n,k} = \mathbf{w}_{g_n,k}$  for all  $k \in [K]$ . First, consider the following bound on the regularized empirical risk of  $g_n$ :

**Lemma 13** *Suppose that event  $\mathcal{E}_\gamma$  holds. Then,  $\mathcal{L}_n(g_n) + \mathcal{R}_\theta(g_n) \leq (r_\sigma + 2\lambda_* r_\rho)^2$ .*

**Proof** Define the constant function  $g_0$  by  $g_0(\mathbf{x}) \doteq y_0$  for all  $\mathbf{x} \in \mathbb{R}^d$ . Observe that  $g_0 \in \mathcal{G}_\triangleright(\hat{\mathcal{X}}_K)$ ,  $g_0 = \pi(g_0)$ , and  $\mathcal{R}_\theta(g_0) = 0$ . Since  $g_n$  is a solution to (9), we have  $\mathcal{L}_n(g_n) + \mathcal{R}_\theta(g_n) \leq \mathcal{L}_n(g_0) = \frac{1}{n} \sum_{i \in [n]} (Y_i - y_0)^2$ . Then,  $\mathcal{E}_\gamma$  implies the claim by Theorem 9. ■

To bound the inner product terms in (17), we apply the concentration inequality of Theorem 11. For this, we need the following bound on the parameter space of DCF estimators:

**Lemma 14** *Suppose that event  $\mathcal{E}_\gamma$  and (7) hold. Then,  $g_n \in \bar{\mathcal{G}}_\triangleright(\hat{\mathcal{X}}_K)$ , where*

$$\bar{\mathcal{G}}_\triangleright(\hat{\mathcal{X}}_K) \doteq \left\{ g \in \mathcal{G}_\triangleright(\hat{\mathcal{X}}_K) : \max_{k \in [K]} \max \left\{ |b_{g,k} - y_0|, \sqrt{d} R_{\mathcal{X}_n} \|\mathbf{w}_{g,k}\| \right\} \leq \beta_1 \right\},$$

and  $\beta_1$  is a constant satisfying  $r_\sigma \sqrt{d} \leq \beta_1 = \Theta(\sqrt{dn}(r_\sigma + \lambda_* r_\rho))$ .

**Proof** Since  $\hat{\mathcal{X}}_K \subseteq \mathcal{X}_n$ , we can define  $i_k \in [n]$  for each  $k \in [K]$  such that  $(\hat{\mathbf{X}}_k, Y_{i_k}) \in \mathcal{D}_n$ . By the definition of  $g_n$ , we have  $b_{n,k} = g_n(\hat{\mathbf{X}}_k)$ . Hence, by Theorem 13, and  $\theta_1 \geq R_{\mathcal{X}_n}^2(d/n)$  from (7), it follows that

$$\frac{1}{n} \max_{k \in [K]} \left\{ (b_{n,k} - Y_{i_k})^2, d R_{\mathcal{X}_n}^2 (\|\mathbf{w}_{n,k}\| - \theta_0)_+^2 \right\} \leq \mathcal{L}_n(g_n) + \mathcal{R}_\theta(g_n) \leq (r_\sigma + 2\lambda_* r_\rho)^2.$$

Then, by the triangle inequality and Theorem 9,  $\mathcal{E}_\gamma$  implies  $|b_{n,k} - y_0| \leq (1 + \sqrt{n})(r_\sigma + 2\lambda_* r_\rho)$ . Further, we also get  $\sqrt{d} R_{\mathcal{X}_n} \|\mathbf{w}_{n,k}\| \leq \sqrt{d} R_{\mathcal{X}_n} \theta_0 + \sqrt{n}(r_\sigma + 2\lambda_* r_\rho)$ . Therefore,  $\beta_1$  can be chosen as claimed as  $R_{\mathcal{X}_n} \theta_0 = O(R_{\mathcal{Y}_n} \ln(n))$  by (7),  $\ln(n) \leq \sqrt{n}$ , and  $R_{\mathcal{Y}_n} = O(r_\sigma + \lambda_* r_\rho)$  from Theorem 9. ■

The bound  $\beta_1$  in Theorem 14 scales with  $\sqrt{n}$ , making it too loose to directly bound the Lipschitz constant of  $f_n$  for establishing the near-minimax rate. However, we can still use the bounded class  $\bar{\mathcal{G}}_\triangleright(\hat{\mathcal{X}}_K)$  in the concentration inequality of Theorem 11 to upper bound the inner product terms of (17) in the next result. The key observation is that  $\beta_1$  appears only inside the logarithmic term.

**Lemma 15** *Let  $g_*$  be an approximation of  $f_* \in \mathcal{F}_{\lambda_*, \mathcal{X}_*}$  as in Theorem 4. For any  $\delta \in (0, \beta_1]$ , define event  $\mathcal{E}_{\gamma, \delta}$  as*

$$\begin{aligned} \mathcal{E}_{\gamma, \delta} \doteq \mathcal{E}_\gamma \cap \Big\{ & \langle f_n - g_n, f_* - y \rangle_n = O\left(\sigma(\|f_n - g_n\|_n + \delta) \sqrt{dK \ln(\beta_1/(\delta\gamma))}/n + r_\sigma \delta\right), \\ & \langle g_* - f_*, f_* - y \rangle_n = O\left(\sigma\|g_* - f_*\|_n \sqrt{\ln(1/\gamma)/n}\right), \\ & \langle g_* - g_n, f_* - y \rangle_n = O\left(\sigma(\|g_* - g_n\|_n + \delta) \sqrt{dK \ln(\beta_1/(\delta\gamma))}/n + r_\sigma \delta\right) \Big\}. \end{aligned}$$

Then,  $\mathbb{P}\{\mathcal{E}_{\gamma, \delta}\} \geq 1 - 5\gamma$ .

**Proof** Define the metric  $\psi$  by  $\psi(h, \hat{h}) \doteq \max_{k \in [K]} |b_{h,k} - b_{\hat{h},k}| + 2R_{\mathcal{X}_n} \sqrt{1+d} \|\mathbf{w}_{h,k} - \mathbf{w}_{\hat{h},k}\|$  for all  $h, \hat{h} \in \mathcal{H}$ , where  $\mathcal{H} \in \{\mathcal{F}_\triangleright(\hat{\mathcal{X}}_K), \mathcal{G}_\triangleright(\hat{\mathcal{X}}_K)\}$ . Note that  $\tau_{\phi_\triangleright}^2 \leq 1 + d$  by definition, and  $\|\mathbf{X}_i - \hat{\mathbf{X}}_k\| \leq 2R_{\mathcal{X}_n}$  holds for all  $i \in [n]$  and  $k \in [K]$  by the triangle inequality. Therefore, by Theorem 12, we have  $\|h - \hat{h}\|_n \leq \max_{i \in [n]} |h(\mathbf{X}_i) - \hat{h}(\mathbf{X}_i)| \leq \psi(h, \hat{h})$ . Additionally, note that  $\psi(g, \hat{g}) = O(\beta_1)$  and  $\psi(\pi(g), \pi(\hat{g})) = O(\beta_1)$  for all  $g, \hat{g} \in \bar{\mathcal{G}}_\triangleright(\hat{\mathcal{X}}_K)$ .

Let  $\mathcal{H}_1 \doteq \{f - g : g \in \bar{\mathcal{G}}_\triangleright(\hat{\mathcal{X}}_K), f = \pi(g)\}$ , where  $f_n - g_n \in \mathcal{H}_1$  by  $f_n = \pi(g_n)$  and Theorem 14. For all  $h, \hat{h} \in \mathcal{H}_1$  with  $h = f - g$  and  $\hat{h} = \hat{f} - \hat{g}$ , define the metric  $\psi_1(h, \hat{h}) \doteq \psi(f, \hat{f}) + \psi(g, \hat{g})$ . By using Theorem 10, the bound  $\beta_1$  on the parameter magnitudes within  $\bar{\mathcal{G}}_\triangleright(\hat{\mathcal{X}}_K)$ , and since the parameters of the functions  $g$  and  $\pi(g)$  are the same, we have  $\ln N_{\psi_1}(\mathcal{H}_1, \delta) = O(dK \ln(\beta_1/\delta))$  for all  $\delta \in (0, \beta_1]$ . Then, the first inequality in  $\mathcal{E}_{\gamma, \delta}$  holds with probability at least  $1 - \gamma$  by Theorem 11, using  $\mathcal{H}_n \leftarrow \mathcal{H}_1$ , and  $\psi_n \leftarrow \psi_1$ .

Define the singleton set  $\mathcal{H}_2 \doteq \{g_* - f_*\}$ . Then, the second inequality in  $\mathcal{E}_{\gamma, \delta}$  holds with probability at least  $1 - \gamma$  by Theorem 11, using  $\mathcal{H}_n \leftarrow \mathcal{H}_2$ , the zero constant function for  $\psi_n$ , and taking the limit  $\delta \rightarrow 0$ .

Let  $\mathcal{H}_3 \doteq \{g_* - g : g \in \bar{\mathcal{G}}_\triangleright(\hat{\mathcal{X}}_K)\}$ , where  $g_* - g_n \in \mathcal{H}_3$  by Theorem 14. For all  $h, \hat{h} \in \mathcal{H}_3$  with  $h = g_* - g$  and  $\hat{h} = g_* - \hat{g}$ , define  $\psi_3(h, \hat{h}) \doteq \psi(g, \hat{g})$ . By Theorem 10, and the bound  $\beta_1$  on the parameter magnitudes within  $\bar{\mathcal{G}}_\triangleright(\hat{\mathcal{X}}_K)$ , we have  $\ln N_{\psi_3}(\mathcal{H}_3, \delta) = O(dK \ln(\beta_1/\delta))$  for all  $\delta \in (0, \beta_1]$ . Then, the third inequality in  $\mathcal{E}_{\gamma, \delta}$  holds with probability at least  $1 - \gamma$  by Theorem 11, using  $\mathcal{H}_n \leftarrow \mathcal{H}_3$ , and  $\psi_n \leftarrow \psi_3$ .

Finally, the result follows by combining the three cases with Theorem 9.  $\blacksquare$

We now bound  $\|g_n - g_*\|_n$  and  $\lambda_{f_n}$  by combining the result of Theorem 15 with the “basic inequality” (16). This, in turn, yields a bound on the approximation error  $E_{\text{approx}}$  via (17).

**Lemma 16** *Let  $g_*$  be an approximation of  $f_* \in \mathcal{F}_{\lambda_*, \mathcal{X}_*}$  as in Theorem 4. Set  $\delta_n \doteq r_\sigma \sqrt{dK/n}$ , and suppose that  $\mathcal{E}_{\gamma, \delta_n}$  and (7) hold. Then, for some  $\lambda_0 > 0$ , the following bounds hold:*

$$\begin{aligned} \lambda_{f_n}^2 &\leq \lambda_0^2 = \Theta(\theta_0^2 + \tau_\triangleright^2 \lambda_*^2 + \sigma^2 \ln(\beta_2/\gamma)), \\ \|g_n - g_*\|_n^2 &= O\left(\frac{dK}{n} \left(\tau_\triangleright^2 (1 + r_\rho^2) \lambda_*^2 + \sigma^2 \ln(\beta_2/\gamma)\right)\right), \\ E_{\text{approx}} &= O\left(\frac{dK}{n} (1 + r_\rho^2) \lambda_0^2\right), \end{aligned}$$

where  $\beta_2 \doteq n\beta_1/(r_\sigma\sqrt{d})$  satisfies  $n \leq \beta_2 = \Theta(n^{3/2}(1 + \lambda_*\rho/\sigma))$ .

**Proof** Notice that  $\delta_n \in (0, \beta_1]$  by definition since  $K \leq n$ . Using  $ab = (a/c)(cb) \leq a^2/(2c^2) + b^2(c^2/2)$  for all  $a, b \in \mathbb{R}$  and  $c > 0$ , we obtain from Theorem 15 and  $\beta_1/\delta_n = O(\beta_2)$  that

$$\begin{aligned} \langle g_* - g_n, f_* - y \rangle_n &= \frac{1}{8} \|g_* - g_n\|_n^2 + O\left(\delta_n^2 + \frac{dK\sigma^2}{n} \ln(\beta_1/(\delta_n\gamma)) + r_\sigma\delta_n\right) \\ &= \frac{1}{8} \|g_* - g_n\|_n^2 + O\left(\frac{dK\sigma^2}{n} \ln(\beta_2/\gamma)\right). \end{aligned} \quad (20)$$

From Theorem 4, we have  $\|g_* - f_*\|_n^2 = O(d\lambda_*^2\epsilon_n^2(\hat{\mathcal{X}}_K))$  by  $\tilde{\tau}_\triangleright^2 = O(d)$ , and  $\mathcal{R}_\theta(g_*) = O(\theta_1\tau_\triangleright^2\lambda_*^2)$  by  $\theta_1 \geq K\theta_2$  from (7). Recall that  $\theta_1 = \Theta(\max\{1, R_{\mathcal{X}_n}^2\}dK/n)$  from (7). Further,  $\epsilon_n^2(\hat{\mathcal{X}}_K) = O(R_{\mathcal{X}_n}^2 K/n)$  from Theorem 8. Therefore, using  $\mathcal{R}_\theta(g_n) \geq \theta_1(\lambda_{f_n} - \theta_0)_+^2$ ,  $\lambda_{f_n} = \lambda_{g_n}$ ,  $\tau_\triangleright \geq 1$ , and combining (20) with the “basic inequality” (16), we obtain

$$\frac{1}{4} \|g_* - g_n\|_n^2 + \max\{1, R_{\mathcal{X}_n}^2\} \frac{dK}{n} (\lambda_{f_n} - \theta_0)_+^2 = O\left(\frac{dK}{n} \left(\tau_\triangleright^2 \max\{1, R_{\mathcal{X}_n}^2\} \lambda_*^2 + \sigma^2 \ln(\beta_2/\gamma)\right)\right). \quad (21)$$

Then, the claims for  $\lambda_{f_n}^2$  and  $\|g_n - g_*\|_n^2$  follow from (21) after rearranging terms, and applying  $R_{\mathcal{X}_n} = O(r_\rho)$  which follows from  $\mathcal{E}_{\gamma, \delta_n} \subset \mathcal{E}_\gamma$  by Theorem 9.

Similarly to the derivation of (20), Theorem 15 yields for  $\delta_n$  that

$$\langle f_n - g_n, f_* - y \rangle_n + \langle g_* - f_*, f_* - y \rangle_n = O\left(\|f_n - g_n\|_n^2 + \|g_* - f_*\|_n^2 + \frac{dK}{n} \sigma^2 \ln(\beta_2/\gamma)\right). \quad (22)$$

Since  $\|f_n - g_n\|_n^2 = O(\lambda_{f_n}^2 \epsilon_n^2(\hat{\mathcal{X}}_K))$  by Theorem 3, and  $\|g_* - f_*\|_n^2 = O(\tilde{\tau}_\triangleright^2 \lambda_*^2 \epsilon_n^2(\hat{\mathcal{X}}_K))$  by Theorem 4, we upper bound the approximation error  $E_{\text{approx}}$  by combining (17) with (20), the bound on  $\|g_* - g_n\|_n^2$ , and (22) as

$$E_{\text{approx}} = O\left((\lambda_{f_n}^2 + \tilde{\tau}_\triangleright^2 \lambda_*^2) \epsilon_n^2(\hat{\mathcal{X}}_K) + \theta_1 \tau_\triangleright^2 \lambda_*^2 + \frac{dK}{n} \left(\tau_\triangleright^2 (1 + r_\rho^2) \lambda_*^2 + \sigma^2 \ln(\beta_2/\gamma)\right)\right), \quad (23)$$

which proves the claim on  $E_{\text{approx}}$ , as  $\tilde{\tau}_\triangleright^2 = O(d)$  by definition, and  $\lambda_{f_n}^2 \epsilon_n^2(\hat{\mathcal{X}}_K) = O(\frac{K}{n} r_\rho^2 \lambda_0^2)$  by  $\lambda_{f_n} \leq \lambda_0$ ,  $\epsilon_n^2(\hat{\mathcal{X}}_K) = O(R_{\mathcal{X}_n}^2 K/n)$  by Theorem 8, and  $R_{\mathcal{X}_n} = O(r_\rho)$ . The bounds on  $\beta_2$  hold by definition.  $\blacksquare$

In the proof of Theorem 16, for bounding  $\lambda_{f_n}$ , we used that  $\theta_1$  in (7) scales with  $\max\{1, R_{\mathcal{X}_n}\}$  rather than just  $R_{\mathcal{X}_n}$ . This choice is necessary because, in the latter case, we cannot ensure that  $\sigma/R_{\mathcal{X}_n}$  remains upper bounded in the setting of Theorem 1.

Notice that the bound  $\beta_1$  on the slope parameters from Theorem 14 is improved by the bound on  $\lambda_{f_n}$  from Theorem 16, replacing the earlier  $\sqrt{n}$  dependence with  $\theta_0 + \sqrt{\ln(n)}$ . This improvement will be important for applying our concentration inequality in the random design setting and obtaining the near-minimax rate.

## 4.4.3 APPLYING THE CONCENTRATION INEQUALITY IN THE RANDOM DESIGN

After bounding the approximation error  $E_{\text{approx}} = \|f_n^+ - y\|_n^2 - \|f_* - y\|_n^2$  of the DCF estimator  $f_n^+$  to the regression function  $f_*$ , it remains to bound the first term in (11), namely  $\|f_n^+ - f_*\|_*^2 - \tilde{c}_0 E_{\text{approx}}$ , in order to complete the proof of Theorem 1. To this end, we use the following concentration inequality, which builds on the Bernstein inequality and leverages the results of Balázs et al. (2016), as applied here to the product of subgaussian random variables.

**Lemma 17** *Let  $\mathcal{F}$  be a finite, nonempty set, and  $n \in \mathbb{N}$ . For each  $f \in \mathcal{F}$  and  $i \in [n] \cup \{0\}$ , let  $Z_{f,i}, W_{f,i}$  be real-valued, subgaussian random variables satisfying  $\mathbb{E}[e^{Z_{f,i}^2/\omega^2}] \leq 2$  and  $\mathbb{E}[e^{W_{f,i}^2/\nu^2}] \leq 2$  for some  $\omega, \nu > 0$ . Suppose that, for each fixed  $f \in \mathcal{F}$ , the random variables  $Z_{f,0}W_{f,0}, Z_{f,1}W_{f,1}, \dots, Z_{f,n}W_{f,n}$  are i.i.d., and that there exist constants  $\alpha, \mu > 0$  such that  $\mu^2 \leq \mathbb{E}[Z_{f,0}^2] \leq \alpha \mathbb{E}[Z_{f,0}W_{f,0}]$  for all  $f \in \mathcal{F}$ . Then, for all  $\gamma \in (0, 1)$ , it holds with probability at least  $1 - \gamma$  that*

$$\max_{f \in \mathcal{F}} \left\{ \mathbb{E}[Z_{f,0}W_{f,0}] - \frac{2}{n} \sum_{i=1}^n Z_{f,i}W_{f,i} \right\} \leq 16(\omega + 2\alpha\nu)\nu \frac{\ln(3\omega/\mu)}{n} \ln(|\mathcal{F}|/\gamma).$$

**Proof** See Section 4.4.4. ■

The condition  $\mathbb{E}[Z_{f,0}^2] \leq \alpha \mathbb{E}[Z_{f,0}W_{f,0}]$  in Theorem 17 is related to the widely used Bernstein condition (Bartlett and Mendelson, 2006, Definition 2.6).

In order to apply Theorem 17, we need to show that  $f_n^+$  belongs to a (non-random) function class with a bounded covering number. From Theorem 5 we already know that  $\lambda_{f_n^+} \leq (1 + \theta_3)\lambda_{f_n}$ , where  $\lambda_{f_n}$  is further bounded in Theorem 16 as  $\lambda_{f_n} \leq \lambda_0$ . The next result provides the missing upper bound on the magnitudes of the bias parameters of  $f_n^+$ . This bound follows from properties of  $f_n^+$  ensured by the final step in the definition of the DCF estimator (5). Unlike in Section 4.4.2, here the bound on the bias terms cannot be allowed to scale polynomially with the sample size  $n$ .

**Lemma 18** *Suppose that event  $\mathcal{E}_{\gamma, \delta_n}$  and (7) hold. Then, there exists  $\beta_0 > 0$  such that  $\max_{k \in \mathcal{I}_n^+} |b_{f_n^+, k} - y_0| \leq \beta_0$ , and  $\beta_0 = \Theta(\tau_{\phi_\triangleright}(1 + r_\rho)\theta_3\lambda_0)$ . Further,  $\mathcal{L}_n(f_n^+) = O((1 + r_\rho^2)\theta_3^2\lambda_0^2)$ .*

**Proof** For each  $k \in \mathcal{I}_n^+$ , let  $i_k \in [n]$  be such that  $f_n^+(\mathbf{X}_{i_k}) = b_{f_n^+, k} + \mathbf{w}_{f_n^+, k}^\top \phi_\triangleright(\mathbf{X}_{i_k}, \hat{\mathbf{X}}_k)$ , which always exists by the definition of  $\mathcal{I}_n^+$  in (5). The triangle inequality and  $\mathcal{E}_{\gamma, \delta_n} \subset \mathcal{E}_\gamma$  yield  $\|\mathbf{X}_{i_k} - \hat{\mathbf{X}}_k\| \leq 2r_\rho$ . Moreover, Theorems 5 and 16 imply  $\lambda_{f_n^+} \leq (1 + \theta_3)\lambda_0$ . Then, by the triangle and Cauchy-Schwartz inequalities, and Theorem 6, we get

$$\begin{aligned} |b_{f_n^+, k} - y_0| &\leq |f_n^+(\mathbf{X}_{i_k}) - y_0| + \|\mathbf{w}_{f_n^+, k}\| \|\phi_\triangleright(\mathbf{X}_{i_k}, \hat{\mathbf{X}}_k)\| \\ &\leq |f_n^+(\mathbf{X}_{i_k}) - y_0| + 2r_\rho \tau_{\phi_\triangleright}(1 + \theta_3)\lambda_0. \end{aligned} \quad (24)$$

Additionally, from (5), we have  $f_n^+(\mathbf{X}_i) = C_n^+ + \hat{f}_n^+(\mathbf{X}_i)$  for all  $i \in [n]$ , which yields

$$\frac{1}{n} \sum_{i=1}^n f_n^+(\mathbf{X}_i) = C_n^+ + \sum_{i=1}^n \hat{f}_n^+(\mathbf{X}_i) = \bar{Y}. \quad (25)$$

Since  $\lambda_{f_n^+} \leq (1 + \theta_3)\lambda_0$ , we have that  $f_n^+$  is  $((1 + \theta_3)\lambda_0)$ -Lipschitz w.r.t.  $\|\cdot\|$  over  $\mathbb{R}^d$ , that is  $f_n^+ \in \mathcal{F}_{(1+\theta_3)\lambda_0, \mathbb{R}^d}$ . Using this, the triangle and Jensen's inequalities, (25), and Theorem 9,  $\mathcal{E}_\gamma$  implies

$$\begin{aligned} |f_n^+(\mathbf{X}_{i_k}) - y_0| &\leq \left| f_n^+(\mathbf{X}_{i_k}) - \frac{1}{n} \sum_{i=1}^n f_n^+(\mathbf{X}_i) \right| + |\bar{Y} - y_0| \\ &\leq 2(1 + \theta_3)r_\rho\lambda_0 + (r_\sigma + 2\lambda_*r_\rho). \end{aligned} \quad (26)$$

We prove the bound  $\beta_0$  on the bias by combining (24) and (26), and using  $\theta_3 \geq 1$  with  $\max\{\lambda_*, r_\sigma\} = O(\lambda_0)$ .

By the triangle inequality and Theorem 9, event  $\mathcal{E}_\gamma$  also implies  $|f_n^+(\mathbf{X}_{i_k}) - Y_{i_k}| \leq |f_n^+(\mathbf{X}_{i_k}) - y_0| + O(r_\sigma + \lambda_0 r_\rho)$ . The bound on  $\mathcal{L}_n(f_n^+)$  then follows from (26) by squaring and averaging over all  $i_k \in [n]$ .  $\blacksquare$

Having bounded the parameter magnitudes of  $f_n^+$ , we now construct a bounded, non-random function class that always contains a function uniformly approximating  $f_n^+$  over the entire space  $\mathbb{R}^d$  to the required accuracy. This construction relies on the fact that  $\phi_\triangleright(\cdot, \cdot)$  is Lipschitz in its second argument (Theorem 6).

We also need the notion of tuples: the set of all tuples of size  $k \in \mathbb{N}$  with elements from a set  $\mathcal{X} \subseteq \mathbb{R}^d$  is defined as  $\mathcal{T}_k(\mathcal{X}) \doteq \{\langle \tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_k \rangle : \tilde{\mathbf{x}}_k \in \mathcal{X}, k \in [K]\}$ . We extend the definition of  $\mathcal{F}_\triangleright(\mathcal{X})$  to allow  $\tilde{\mathcal{X}}$  to be a tuple, so that multiple function parameters may be associated with the same center.

The following result presents the construction mentioned above:

**Lemma 19** *Suppose the conditions of Theorem 18 hold, and let  $k_*$  be as in Theorem 8. Furthermore, let  $\mathcal{X}_\rho \doteq \{\mathbf{x} \in \mathcal{X}_* : \|\mathbf{x} - \mathbb{E}[\mathbf{X}]\| \leq r_\rho\}$ ,  $\eta > 0$ , and  $\hat{\mathcal{X}}_{\rho, \eta}$  be an  $\eta$ -cover of  $\mathcal{X}_\rho$  w.r.t.  $\|\cdot\|$ . Define the following (non-random) function class:*

$$\overline{\mathcal{F}}_{\triangleright, \eta} \doteq \bigcup_{k \in [k_*]} \bigcup_{\tilde{\mathcal{X}} \in \mathcal{T}_k(\hat{\mathcal{X}}_{\rho, \eta})} \overline{\mathcal{F}}_\triangleright(\tilde{\mathcal{X}})$$

where  $\overline{\mathcal{F}}_\triangleright(\tilde{\mathcal{X}}) \doteq \{f \in \mathcal{F}_\triangleright(\tilde{\mathcal{X}}) : |b_{f,k} - y_0| \leq \beta_0, \|\mathbf{w}_{f,k}\| \leq (1 + \theta_3)\lambda_0, k \in [\tilde{\mathcal{X}}]\}$  for all  $\tilde{\mathcal{X}} \in \mathcal{T}_k(\hat{\mathcal{X}}_{\rho, \eta})$  and  $k \in \mathbb{N}$ , restricting  $\mathcal{F}_\triangleright(\tilde{\mathcal{X}})$  to functions with bounded parameter magnitudes. Then, there exists  $f \in \overline{\mathcal{F}}_{\triangleright, \eta}$  such that  $\max_{\mathbf{x} \in \mathbb{R}^d} |f_n^+(\mathbf{x}) - f(\mathbf{x})| \leq (1 + \theta_3)\lambda_0\lambda_{\phi_\triangleright}\eta$ .

**Proof** Let  $K_+ \doteq |\mathcal{I}_n^+|$  and write  $\mathcal{I}_n^+ \doteq \{j_1, \dots, j_{K_+}\}$  with some indices  $j_1, \dots, j_{K_+} \in [K]$ . By Theorem 9,  $\mathcal{E}_\gamma$  implies  $\hat{\mathcal{X}}_K \subseteq \mathcal{X}_n \subset \mathcal{X}_\rho$ . Since  $K \leq k_*$  by Theorem 8, it follows that  $1 \leq K_+ \leq K \leq k_*$ , and since  $\hat{\mathcal{X}}_{\rho, \eta}$  is an  $\eta$ -cover of  $\mathcal{X}_\rho$  w.r.t.  $\|\cdot\|$ , we can select a tuple  $\tilde{\mathcal{X}} \doteq \{\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{K_+}\} \in \mathcal{T}_{K_+}(\hat{\mathcal{X}}_{\rho, \eta})$  such that  $\overline{\mathcal{F}}_\triangleright(\tilde{\mathcal{X}}) \subset \overline{\mathcal{F}}_{\triangleright, \eta}$  and  $\|\hat{\mathbf{X}}_{j_k} - \tilde{\mathbf{x}}_k\| \leq \eta$  for all  $k \in [K_+]$ . The reason for using tuples is that we cannot guarantee that distinct elements from  $\hat{\mathcal{X}}_{\rho, \eta}$  can be associated with all of  $\hat{\mathbf{X}}_{j_1}, \dots, \hat{\mathbf{X}}_{j_{K_+}}$ .

Recall that we have  $\lambda_{f_n^+} \leq (1 + \theta_3)\lambda_{f_n} \leq (1 + \theta_3)\lambda_0$  from Theorems 5 and 16, and  $\max_{k \in \mathcal{I}_n^+} |b_{f_n^+, k} - y_0| \leq \beta_0$  from Theorem 18. Using this, we define function  $f \in \overline{\mathcal{F}}_\triangleright(\tilde{\mathcal{X}})$  by setting  $b_{f,k} \doteq b_{f_n^+, j_k}$  and  $\mathbf{w}_{f,k} \doteq \mathbf{w}_{f_n^+, j_k}$  for all  $k \in [K_+]$ . Then, for all  $\mathbf{x} \in \mathbb{R}^d$ , we have by

the Cauchy-Schwartz inequality and the  $\lambda_{\phi_{\triangleright}}$ -Lipschitzness of  $\phi_{\triangleright}(\mathbf{x}, \cdot)$  from Theorem 6 that

$$|f_n^+(\mathbf{x}) - f(\mathbf{x})| \leq \max_{k \in [K_+]} \left| \left( \phi_{\triangleright}(\mathbf{x}, \hat{\mathbf{X}}_{j_k}) - \phi_{\triangleright}(\mathbf{x}, \tilde{\mathbf{x}}_k) \right)^\top \mathbf{w}_{f,k} \right| \leq \max_{k \in [K_+]} \lambda_{\phi_{\triangleright}} \|\hat{\mathbf{X}}_{j_k} - \tilde{\mathbf{x}}_k\| \|\mathbf{w}_{f,k}\|,$$

which proves the result as  $\|\hat{\mathbf{X}}_{j_k} - \tilde{\mathbf{x}}_k\| \leq \eta$  and  $\|\mathbf{w}_{f,k}\| \leq \lambda_{f_n^+} \leq (1 + \theta_3)\lambda_0$  for all  $k \in [K_+]$ . ■

As the concentration inequality in Theorem 17 requires a finite function set, next we construct a cover of  $\overline{\mathcal{F}}_{\triangleright, \eta}$ .

**Lemma 20** *Suppose that the conditions of Theorem 19 hold, and that the cover  $\hat{\mathcal{X}}_{\rho, \eta}$  of  $\mathcal{X}_\rho$  w.r.t.  $\|\cdot\|$  is of minimal cardinality. For all  $k_0 \in \mathbb{N}$  and  $\tilde{\mathcal{X}} \in \mathcal{T}_{k_0}(\mathbb{R}^d)$ , define a metric between any  $f, \hat{f} \in \mathcal{F}_{\triangleright}(\tilde{\mathcal{X}})$  as*

$$\psi_{k_0}(f, \hat{f}) \doteq \max_{k \in [k_0]} |b_{f,k} - b_{\hat{f},k}| + r_\rho \|\mathbf{w}_{f,k} - \mathbf{w}_{\hat{f},k}\|.$$

Let  $\eta \in (0, r_\rho/2]$  and  $\delta \in (0, \beta_0]$ . For all  $k \in \mathbb{N}$  and  $\tilde{\mathcal{X}} \in \mathcal{T}_k(\hat{\mathcal{X}}_{\rho, \eta})$ , define  $\hat{\overline{\mathcal{F}}}_{\triangleright, \delta}(\tilde{\mathcal{X}})$  to be a  $\delta$ -cover of  $\overline{\mathcal{F}}_{\triangleright}(\tilde{\mathcal{X}})$  w.r.t.  $\psi_k$  of minimal cardinality. Finally, let

$$\hat{\overline{\mathcal{F}}}_{\triangleright, \eta, \delta} \doteq \bigcup_{k \in [k_*]} \bigcup_{\tilde{\mathcal{X}} \in \mathcal{T}_k(\hat{\mathcal{X}}_{\rho, \eta})} \hat{\overline{\mathcal{F}}}_{\triangleright, \delta}(\tilde{\mathcal{X}}).$$

Then, for every function  $f \in \overline{\mathcal{F}}_{\triangleright, \eta}$  there exists  $\hat{f} \in \hat{\overline{\mathcal{F}}}_{\triangleright, \eta, \delta}$  which satisfies  $|f(\mathbf{x}) - \hat{f}(\mathbf{x})| \leq 2\delta\tau_{\phi_{\triangleright}}(1 + \|\mathbf{x} - \mathbb{E}[\mathbf{X}]\|/r_\rho)$  for all  $\mathbf{x} \in \mathbb{R}^d$ . Additionally,  $\ln |\hat{\overline{\mathcal{F}}}_{\triangleright, \eta, \delta}| = O(dk_* \ln(r_\rho\beta_0/(\eta\delta)))$ .

**Proof** Note that the result of Theorem 12 extends straightforwardly to  $\mathcal{F}_{\triangleright}(\tilde{\mathcal{X}})$ , where  $\tilde{\mathcal{X}} = \langle \tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{k_0} \rangle$  is a tuple. Therefore, we have for all  $f, \hat{f} \in \mathcal{F}_{\triangleright}(\tilde{\mathcal{X}})$  and for all  $\mathbf{x} \in \mathbb{R}^d$  that

$$|f(\mathbf{x}) - \hat{f}(\mathbf{x})| \leq \psi_{k_0}(f, \hat{f}) \left( 1 + \max_{k \in [k_0]} \tau_{\phi_{\triangleright}} \|\mathbf{x} - \tilde{\mathbf{x}}_k\|/r_\rho \right). \quad (27)$$

Fix  $f \in \overline{\mathcal{F}}_{\triangleright, \eta}$  arbitrarily, and let  $\tilde{\mathcal{X}} \doteq \langle \tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{k_0} \rangle \in \mathcal{T}_{k_0}(\hat{\mathcal{X}}_{\rho, \eta})$  such that  $f \in \overline{\mathcal{F}}_{\triangleright}(\tilde{\mathcal{X}})$ . By the definition of  $\delta$ -cover w.r.t.  $\psi_{k_0}$ , we choose  $\hat{f} \in \hat{\overline{\mathcal{F}}}_{\triangleright, \delta}(\tilde{\mathcal{X}}) \subset \hat{\overline{\mathcal{F}}}_{\triangleright, \eta, \delta}$  to be such that  $\psi_{k_0}(f, \hat{f}) \leq \delta$ . Now fix any  $\mathbf{x} \in \mathbb{R}^d$ . Since  $\tilde{\mathbf{x}}_k \in \mathcal{X}_\rho$  for all  $k \in [k_0]$ , we have  $\|\tilde{\mathbf{x}}_k - \mathbb{E}[\mathbf{X}]\| \leq r_\rho$ , and the triangle inequality yields  $\max_{k \in [k_0]} \|\mathbf{x} - \tilde{\mathbf{x}}_k\| \leq \|\mathbf{x} - \mathbb{E}[\mathbf{X}]\| + r_\rho$ . Therefore, (27) and  $\tau_{\phi_{\triangleright}} \geq 1$  imply the claimed upper bound on  $|f(\mathbf{x}) - \hat{f}(\mathbf{x})|$ .

Define  $\mathcal{T}_{\eta, k} \doteq \mathcal{T}_k(\hat{\mathcal{X}}_{\rho, \eta})$  for all  $k \in [k_*]$ . By Theorem 7,  $|\mathcal{T}_{\eta, k}| = |\hat{\mathcal{X}}_{\rho, \eta}|^k = O((r_\rho/\eta)^{dk})$ . Further, for all  $\tilde{\mathcal{X}} \in \mathcal{T}_{\eta, k}$ , the bounds on the bias and slope parameters of any  $f, \hat{f} \in \overline{\mathcal{F}}_{\triangleright}(\tilde{\mathcal{X}})$  imply  $\psi_k(f, \hat{f}) \leq 2(\beta_0 + (1 + \theta_3)r_\rho\lambda_0) = O(\beta_0)$ . Therefore,  $N_{\psi_k}(\overline{\mathcal{F}}_{\triangleright}(\tilde{\mathcal{X}}), \delta) = O((\beta_0/\delta)^{(1+d_{\triangleright})k})$  by Theorem 10. Note that  $\sum_{k \in [k_*]} t^k = (t^{k_*+1} - t)/(t - 1) < 2t^{k_*}$  holds for any  $t \geq 2$ . Then, using  $r_\rho\beta_0/(\eta\delta) \geq 2$  for all  $\eta \in (0, r_\rho/2]$  and  $\delta \in (0, \beta_0]$ , and  $d \leq d_{\triangleright}$ , we obtain

$$|\hat{\overline{\mathcal{F}}}_{\triangleright, \eta, \delta}| = \sum_{k=1}^{k_*} \sum_{\tilde{\mathcal{X}} \in \mathcal{T}_{\eta, k}} N_{\psi_k}(\overline{\mathcal{F}}_{\triangleright}(\tilde{\mathcal{X}}), \delta) = O\left(\sum_{k=1}^{k_*} |\mathcal{T}_{\eta, k}| (\beta_0/\delta)^{(1+d_{\triangleright})k}\right) = O\left(\left(\frac{r_\rho\beta_0}{\eta\delta}\right)^{(1+d_{\triangleright})k_*}\right),$$



which proves the claimed upper bound on  $|\hat{\mathcal{F}}_{\triangleright, \eta, \delta}|$  with  $1 + d_{\triangleright} = O(d)$ .  $\blacksquare$

Finally, we are ready to combine the results and prove Theorem 1.

**Proof** [of Theorem 1] Define  $\eta \doteq (r_\rho/2)(K/n)$  and choose  $\delta \in (0, \beta_0]$  to be such that  $\delta = \Theta((1 + r_\rho)\lambda_0(K/n))$ , which is always possible since  $\beta_0^2 = \Theta(\tau_{\phi_\triangleright}^2(1 + r_\rho)^2\theta_3^2\lambda_0^2)$ ,  $\theta_3 \geq 1$ ,  $\tau_{\phi_\triangleright} \geq 1$ , and  $K \leq n$ . Let  $\hat{f}_n^{++} \in \overline{\mathcal{F}}_{\triangleright, \eta}$  be the approximation to the DCF estimator  $f_n^+$  from Theorem 19, and  $\hat{f}_n^{++} \in \hat{\mathcal{F}}_{\triangleright, \eta, \delta}$  be the approximation to  $\hat{f}_n^{++}$  from Theorem 20. Define  $c_{\eta, \delta} \doteq (1 + \theta_3)\lambda_0\lambda_{\phi_\triangleright}\eta + 4\delta\tau_{\phi_\triangleright}$ , and notice that  $\mathbb{E}[\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|^2/r_\rho^2] \leq \ln(2) < 1$  holds by (8) and the Jensen's inequality. Then,  $\|f_n^+ - \hat{f}_n^{++}\|_* \leq c_{\eta, \delta}$  follows from the triangle inequality, and Theorems 19 and 20. Furthermore, by using  $(a + b)^2 \leq 2(a^2 + b^2)$  for all  $a, b \in \mathbb{R}$ ,  $\theta_3 \geq 1$ , and  $\max\{\lambda_{\phi_\triangleright}, \tau_{\phi_\triangleright}\}^2 = O(d)$ , we obtain

$$\|f_n^+ - f_*\|_*^2 \leq 2\|\hat{f}_n^{++} - f_*\|_*^2 + 2c_{\eta, \delta}^2 = 2\|\hat{f}_n^{++} - f_*\|_*^2 + O\left(\frac{dK}{n}(1 + r_\rho^2)\theta_3^2\lambda_0^2\right). \quad (28)$$

Notice that if  $\|\hat{f}_n^{++} - f_*\|_*^2 \leq \beta_0^2/n$ , then the term  $\|\hat{f}_n^{++} - f_*\|_*^2$  in (28) becomes negligible, since  $\tau_{\phi_\triangleright}^2 = O(d)$ . Hence, we can assume, without loss of generality, that  $\hat{f}_n^{++} \in \hat{\mathcal{F}}_{\triangleright, \eta, \delta, \beta_0} \doteq \{f \in \hat{\mathcal{F}}_{\triangleright, \eta, \delta} : \|f - f_*\|_*^2 > \beta_0^2/n\}$ .

Combining the error decomposition (11) using  $\tilde{c}_0 \doteq 4$ , the bound of Theorem 16 on the approximation error  $E_{\text{approx}} = \|f_n^+ - y\|_n^2 - \|f_* - y\|_n^2$ , and (28) yields

$$\|f_n^+ - f_*\|_*^2 = 2(\|\hat{f}_n^{++} - f_*\|_*^2 - 2E_{\text{approx}}) + O\left(\frac{dK}{n}(1 + r_\rho^2)\theta_3^2\lambda_0^2\right). \quad (29)$$

Event  $\mathcal{E}_\gamma$  implies  $\max_{i \in [n]} \|\mathbf{X}_i - \mathbb{E}[\mathbf{X}]\|^2/r_\rho^2 \leq 1$ . Hence,  $\|f_n^+ - \hat{f}_n^{++}\|_n \leq c_{\eta, \delta}$  follows from the triangle inequality, and Theorems 19 and 20. Then, using the Cauchy-Schwartz inequality, and the bound on  $\|f_n^+ - y\|_n^2$  from Theorem 18, we obtain

$$\begin{aligned} -\|f_n^+ - y\|_n^2 &= \|f_n^+ - \hat{f}_n^{++}\|_n^2 - 2\langle f_n^+ - \hat{f}_n^{++}, f_n^+ - y \rangle_n - \|\hat{f}_n^{++} - y\|_n^2 \\ &\leq -\|\hat{f}_n^{++} - y\|_n^2 + c_{\eta, \delta}^2 + 2\|f_n^+ - \hat{f}_n^{++}\|_n\|f_n^+ - y\|_n \\ &\leq -\|\hat{f}_n^{++} - y\|_n^2 + c_{\eta, \delta}^2 + 2c_{\eta, \delta}(1 + r_\rho)\theta_3\lambda_0. \end{aligned} \quad (30)$$

For any function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , define  $Z_{f,0} \doteq f(\mathbf{X}) - f_*(\mathbf{X})$ ,  $W_{f,0} \doteq f(\mathbf{X}) + f_*(\mathbf{X}) - 2Y$  and  $Z_{f,i} \doteq f(\mathbf{X}_i) - f_*(\mathbf{X}_i)$ ,  $W_{f,i} \doteq f(\mathbf{X}_i) + f_*(\mathbf{X}_i) - 2Y_i$  for all  $i \in [n]$ . Then, we have  $\mathbb{E}[Z_{f,0}W_{f,0}] = \|f - y\|_*^2 - \|f_* - y\|_*^2$  and  $\frac{1}{n} \sum_{i \in [n]} Z_{f,i}W_{f,i} = \|f - y\|_n^2 - \|f_* - y\|_n^2$ , since  $a^2 - b^2 = (a - b)(a + b)$  for all  $a, b \in \mathbb{R}$ . Note that  $\|f - f_*\|_*^2 = \|f - y\|_*^2 - \|f_* - y\|_*^2$  for any function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  (e.g., Györfi et al., 2002, Section 1.1), which can be also expressed as  $\mathbb{E}[Z_{f,0}^2] = \mathbb{E}[Z_{f,0}W_{f,0}]$ . Therefore, combining (29) and (30) implies

$$\begin{aligned} \|f_n^+ - f_*\|_*^2 &\leq 2\left(\|\hat{f}_n^{++} - y\|_n^2 - \|f_* - y\|_n^2 - 2(\|\hat{f}_n^{++} - y\|_n^2 - \|f_* - y\|_n^2)\right) \\ &\quad + O\left(\frac{dK}{n}(1 + r_\rho^2)\theta_3^2\lambda_0^2\right) \\ &= O\left(\max_{f \in \hat{\mathcal{F}}_{\triangleright, \eta, \delta, \beta_0}} \left\{ \mathbb{E}[Z_{f,0}W_{f,0}] - \frac{2}{n} \sum_{i=1}^n Z_{f,i}W_{f,i} \right\} + \frac{dK}{n}(1 + r_\rho^2)\theta_3^2\lambda_0^2\right). \end{aligned} \quad (31)$$

Note that  $Z_{f,0}W_{f,0}, \dots, Z_{f,n}W_{f,n}$  are i.i.d. random variables, and  $\mathbb{E}[Z_{f,0}^2] \geq \beta_0^2/n$ , for all  $f \in \hat{\mathcal{F}}_{\triangleright, \eta, \delta, \beta_0}$ .

Take any  $f \in \hat{\mathcal{F}}_{\triangleright, \eta, \delta, \beta_0}$ , and let  $\tilde{\mathcal{X}} \doteq \{\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_{k_0}\} \in \mathcal{T}_{k_0}(\hat{\mathcal{X}}_{\rho, \eta})$  be the centers associated with  $f$  for some  $k_0 \in [k_*]$ , that is  $f \in \hat{\mathcal{F}}_{\triangleright, \delta}(\tilde{\mathcal{X}})$ . Then, for any  $\mathbf{x} \in \mathcal{X}_*$ , using the triangle and Cauchy-Schwartz inequalities, Theorem 6,  $y_0 = f_*(\mathbf{x}_0)$ ,  $\mathbf{x}_0 \in \mathcal{X}_*$ ,  $f_* \in \mathcal{F}_{\lambda_*, \mathcal{X}_*}$ , and  $\|\tilde{\mathbf{x}}_k - \mathbb{E}[\mathbf{X}]\| \leq r_\rho$  since  $\tilde{\mathbf{x}}_k \in \hat{\mathcal{X}}_{\rho, \eta} \subseteq \mathcal{X}_\rho$ , we get

$$\begin{aligned} |f(\mathbf{x}) - f_*(\mathbf{x})| &\leq \max_{k \in [k_0]} |b_{f,k} - y_0| + \|\phi_{\triangleright}(\mathbf{x}, \tilde{\mathbf{x}}_k)\| \|\mathbf{w}_{f,k}\| + |y_0 - f_*(\mathbf{x})| \\ &\leq \beta_0 + \tau_{\phi_{\triangleright}}(\|\mathbf{x} - \mathbb{E}[\mathbf{X}]\| + r_\rho)\lambda_0 + \lambda_*\|\mathbf{x} - \mathbf{x}_0\|. \end{aligned} \quad (32)$$

Since  $\|\mathbf{x} - \mathbf{x}_0\| \leq 2\|\mathbf{x} - \mathbb{E}[\mathbf{X}]\|$  for all  $\mathbf{x} \in \mathcal{X}_*$  by the definition of  $\mathbf{x}_0$  in Section 4.4.1, and  $\mathbb{E}[e^{\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\|^2/r_\rho^2}] \leq 2$  from (8), (32) yields that the random variable  $Z_{f,0}$  satisfies  $\mathbb{E}[e^{Z_{f,0}^2/\omega^2}] \leq 2$  for some constant  $\omega > 0$  such that  $\omega = \Theta(\beta_0 + \tau_{\phi_{\triangleright}}r_\rho\lambda_0) = O(\beta_0)$ . Similarly, using Theorem 9,  $\mathcal{E}_\gamma$  implies  $\mathbb{E}[e^{Z_{f,i}^2/\omega^2}] \leq 2$  for all  $i \in [n]$ . Additionally, by writing  $W_{f,0} = Z_{f,0} + 2(f_*(\mathbf{X}) - Y)$  and  $W_{f,i} = Z_{f,i} + 2(f_*(\mathbf{X}_i) - Y_i)$  for all  $i \in [n]$ , we have  $\mathbb{E}[e^{W_{f,i}^2/\nu^2}] \leq 2$  for some  $\nu > 0$  satisfying  $\nu = \Theta(\omega + r_\sigma) = O(\beta_0)$ , for all  $i \in [n] \cup \{0\}$ .

Then, by applying Theorem 17 with  $\alpha = 1$ ,  $\mu = \beta_0/\sqrt{n}$ ,  $\omega \leq \nu = O(\beta_0)$ , and using  $|\hat{\mathcal{F}}_{\triangleright, \eta, \delta, \beta_0}| \leq |\hat{\mathcal{F}}_{\triangleright, \eta, \delta}|$  with the bound from Theorem 20, we get with probability at least  $1 - \gamma$  that

$$\begin{aligned} \max_{f \in \hat{\mathcal{F}}_{\triangleright, \eta, \delta, \beta_0}} \left\{ \mathbb{E}[Z_{f,0}W_{f,0}] - \frac{2}{n} \sum_{i=1}^n Z_{f,i}W_{f,i} \right\} &= O\left( \nu^2 \frac{\ln(n)}{n} dk_* \ln\left( \frac{r_\rho\beta_0}{\eta\delta\gamma} \right) \right) \\ &= O\left( \frac{dk_*}{n} \beta_0^2 \ln(n) \ln(dn/\gamma) \right), \end{aligned} \quad (33)$$

where in the last step we simplified using  $\eta\delta = \Omega(r_\rho(1 + r_\rho)\lambda_0/n^2)$ ,  $\tau_{\phi_{\triangleright}} = O(\sqrt{d})$ ,  $\theta_3 = O(\ln(n)) = O(\sqrt{n})$  from (7), so  $(r_\rho\beta_0)/(\eta\delta) = O(n^2\beta_0/((1 + r_\rho)\lambda_0)) = O(n^2\sqrt{dn})$ .

We finally prove Theorem 1 by combining (31) and (33), together with the bound  $K \leq k_* = O(n^{d_*/(2+d_*)})$  from Theorem 8. We conclude the proof by appropriately rescaling  $\gamma$ , and simplifying the bound by using  $\beta_0^2 = O(\tau_{\phi_{\triangleright}}^2(1 + r_\rho^2)\theta_3^2\lambda_0^2)$  from Theorem 18,  $\lambda_0^2 = O(\theta_0^2 + \tau_{\triangleright}^2\lambda_*^2 + \sigma^2 \ln(\beta_2/\gamma))$  from Theorem 16,  $\theta_0^2 = O((r_\sigma^2 + \lambda_*^2) \ln^2(n))$  from (7) and Theorem 9,  $\beta_2 = O(\beta_{\ln}^2)$ , and  $\tau_{\phi_{\triangleright}}^2\tau_{\triangleright}^2 = O(1 + d\mathbb{I}\{\triangleright \neq 2\})$  from Theorems 4 and 6. ■

#### 4.4.4 PROOF OF THE CONCENTRATION INEQUALITY

In this section, we present the deferred proof of Theorem 17. To this end, we employ the following variant of the Bernstein inequality, applied to products of subgaussian random variables:

**Lemma 21** *Let  $Z$  and  $W$  be real-valued random variables satisfying  $\mathbb{E}[e^{Z^2/\omega^2}] \leq 2$  and  $\mathbb{E}[e^{W^2/\nu^2}] \leq 2$  for some constants  $\omega, \nu > 0$ , and  $\mathbb{E}[Z^2] > 0$ . Define the kurtosis of  $Z$  by  $\mathbb{K}[Z] \doteq \mathbb{E}[Z^4]/\mathbb{E}[Z^2]^2$ , and let  $c \geq 4\ln(4\sqrt{\mathbb{K}[Z]})$ . Then, the following hold:*

- (a)  $\mathbb{E}[|ZW|^k] \leq (k!/2)\mathbb{E}[Z^2](2c\nu^2)(c\omega\nu)^{k-2}$  for all integers  $k \geq 2$ ,
- (b)  $\mathbb{E}[e^{s(\mathbb{E}[ZW]-ZW)}] \leq \exp\left(\frac{c\nu^2 s^2 \mathbb{E}[Z^2]}{1-c\omega\nu s}\right)$  for all  $s \in (0, 1/(c\omega\nu))$ .

**Proof** Part (a) was proved in Lemma A.5 of Balázs (2016). Part (b) follows from Theorem 2.10 of Boucheron et al. (2013), using part (a).  $\blacksquare$

Next, we combine Theorem 21 with the ideas of Balázs et al. (2016) to prove Theorem 17.

**Proof** [Proof of Theorem 17] Introduce the shorthand notation  $Z_f \doteq Z_{f,0}$  and  $W_f \doteq W_{f,0}$ . By Lemma A.2 of Balázs (2016), we have  $\mathbb{E}[Z_f^4] \leq 2(2/e)^2\omega^4$ . Since  $\mathbb{E}[Z_f^2] \geq \mu^2$ , it follows that  $\mathbb{K}[Z_f] \leq 2(\omega/\mu)^4$  for all  $f \in \mathcal{F}$ . Define  $c \doteq 8 \ln(3\omega/\mu)$ , which satisfies  $c \geq \max_{f \in \mathcal{F}} 4 \ln(4\sqrt{\mathbb{K}[Z_f]})$ . Then, by applying Theorem 21 with  $\mathbb{E}[Z_f^2] \leq \alpha \mathbb{E}[Z_f W_f]$ , we obtain for all  $s \in (0, 1)$ ,  $i \in [n]$ , and  $f \in \mathcal{F}$  that

$$\mathbb{E}\left[e^{s(\mathbb{E}[Z_f W_f]-Z_{f,i}W_{f,i})/(c\omega\nu)}\right] \leq \exp\left(\frac{c\nu^2 s^2 \alpha \mathbb{E}[Z_f W_f]}{(1-s)(c\omega\nu)^2}\right). \quad (34)$$

Let  $s \in (0, 1)$  and  $t > 0$  be constants to be chosen later. Thereby, applying the union and Chernoff bounds, using the independence of  $Z_f W_f, Z_{f,1} W_{f,1}, \dots, Z_{f,n} W_{f,n}$  for each fixed  $f \in \mathcal{F}$ , and applying (34), we obtain

$$\begin{aligned} & \mathbb{P}\left\{\max_{f \in \mathcal{F}} \left\{\mathbb{E}[Z_f W_f] - \frac{2}{n} \sum_{i=1}^n Z_{f,i} W_{f,i}\right\} > \frac{2tc\omega\nu}{sn}\right\} \\ & \leq \mathbb{P}\left\{\max_{f \in \mathcal{F}} \frac{s}{2c\omega\nu} \sum_{i=1}^n \mathbb{E}[Z_f W_f] - 2Z_{f,i} W_{f,i} > t\right\} \\ & \leq e^{-t} \sum_{f \in \mathcal{F}} \mathbb{E}\left[e^{\frac{s}{2c\omega\nu} (\sum_{i \in [n]} \mathbb{E}[Z_f W_f] - 2Z_{f,i} W_{f,i})}\right] \\ & = e^{-t} \sum_{f \in \mathcal{F}} e^{-\frac{sn \mathbb{E}[Z_f W_f]}{2c\omega\nu}} \prod_{i=1}^n \mathbb{E}\left[e^{s(\mathbb{E}[Z_f W_f]-Z_{f,i}W_{f,i})/(c\omega\nu)}\right] \\ & \leq e^{-t} \sum_{f \in \mathcal{F}} \exp\left(\frac{sn \mathbb{E}[Z_f W_f]}{c\omega\nu} \left(-\frac{1}{2} + \frac{s\nu\alpha}{(1-s)\omega}\right)\right) \\ & = \gamma, \end{aligned}$$

where we set  $s \doteq \omega/(\omega + 2\alpha\nu)$  and  $t \doteq \ln(|\mathcal{F}|/\gamma)$  for the last line. This proves the claim.  $\blacksquare$

## 5 Approximation of DC functions

An important part of our analysis is understanding the approximation rate of the chosen function representation for the estimator to the underlying Lipschitz regression function. To strengthen the connection between our work and the existing literature, we establish uniform

approximation results for other delta-convex function classes that have been proposed to extend convex regression techniques to the more general Lipschitz setting of (1).

Define the uniform norm of a function  $f : \mathcal{X} \rightarrow \mathbb{R}$  on  $\mathcal{X} \subseteq \mathbb{R}^d$  by  $\|f\|_{\infty, \mathcal{X}} \doteq \sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x})|$ .

### 5.1 Max-min-affine functions

For all  $k_0, l_0 \in \mathbb{N}$ , define the set of max-min-affine functions by

$$\mathcal{M}_{k_0, l_0} \doteq \left\{ m : \mathbb{R}^d \rightarrow \mathbb{R} \mid m(\mathbf{x}) \doteq \max_{k \in [k_0]} \min_{l \in [l_0]} b_{k, l} + \mathbf{x}^\top \mathbf{w}_{k, l}, \right. \\ \left. \mathbf{x} \in \mathbb{R}^d, b_{k, l} \in \mathbb{R}, \mathbf{w}_{k, l} \in \mathbb{R}^d, k \in [k_0], l \in [l_0] \right\}.$$

Recall from (2) that  $v_{f, k}$  denotes the parameter of  $f \in \mathcal{F}_\infty(\hat{\mathcal{X}})$  associated with the norm term, for some finite  $\hat{\mathcal{X}} \subset \mathbb{R}^d$  and any  $k \in [|\hat{\mathcal{X}}|]$ . We then define the class in which this parameter is restricted to be nonpositive by  $\mathcal{F}_{\infty-}(\hat{\mathcal{X}}) \doteq \{f \in \mathcal{F}_\infty(\hat{\mathcal{X}}) : v_{f, k} \leq 0, k \in [|\hat{\mathcal{X}}|]\}$ . The approximation in Theorem 2 belongs to  $\mathcal{F}_{\infty-}(\hat{\mathcal{X}})$ , thereby achieving a uniform approximation rate for max-min-affine functions, as established in the next result.

**Corollary 22** *Let  $\mathcal{X} \subset \mathbb{R}^d$  and suppose there exist  $r, t > 0$  such that  $N_{\|\cdot\|}(\mathcal{X}, \epsilon) \leq (r/\epsilon)^t$  for all  $\epsilon \in (0, r]$ . Let  $f \in \mathcal{F}_{\lambda, \mathcal{X}}$  for some Lipschitz constant  $\lambda > 0$ . Then, for all  $k_0 \in \mathbb{N}$  and  $l_0 \geq 2d$ , there exists  $m \in \mathcal{M}_{k_0, l_0}$  such that  $\|f - m\|_{\infty, \mathcal{X}} \leq (1 + \sqrt{d})\lambda r k_0^{-1/t}$  and  $m \in \mathcal{F}_{\sqrt{d}\lambda, \mathbb{R}^d}$ .*

**Proof** Write the max-norm  $\|\cdot\|_\infty$  in max-linear form as  $\|\mathbf{x}\|_\infty = \max_{j \in [d], s \in \{-1, 1\}} s \mathbf{e}_j^\top \mathbf{x}$  for all  $\mathbf{x} \in \mathbb{R}^d$ , where  $\mathbf{e}_1, \dots, \mathbf{e}_d$  are the canonical basis vectors of  $\mathbb{R}^d$ . Then, functions in  $\mathcal{M}_{k_0, l_0}$  with  $l_0 \geq 2d$  can use the internal minimization to represent the negated max-norm  $-\|\cdot\|_\infty$ , thereby allowing implementation of any  $f \in \mathcal{F}_{\infty-}(\hat{\mathcal{X}})$  with any  $\hat{\mathcal{X}} \doteq \{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_{k_0}\}$  as

$$f(\mathbf{x}) = \max_{k \in [k_0]} b_{f, k} + \mathbf{u}_{f, k}^\top (\mathbf{x} - \hat{\mathbf{x}}_k) + v_{f, k} \|\mathbf{x} - \hat{\mathbf{x}}_k\|_\infty = \max_{k \in [k_0]} \min_{\substack{j \in [d], \\ s \in \{-1, 1\}}} b_{f, k} + (\mathbf{u}_{f, k} + v_{f, k} s \mathbf{e}_j)^\top (\mathbf{x} - \hat{\mathbf{x}}_k)$$

for all  $\mathbf{x} \in \mathbb{R}^d$ . Hence, the function  $\hat{f} \in \mathcal{F}_{\infty-}(\hat{\mathcal{X}})$  from Theorem 2 belongs to  $\mathcal{M}_{k_0, l_0}$ .

Set  $\epsilon \doteq r k_0^{-1/t}$ , which ensures that  $N_{\|\cdot\|}(\mathcal{X}, \epsilon) \leq (r/\epsilon)^t = k_0$ . Choose  $\hat{\mathcal{X}} \subset \mathbb{R}^d$  such that  $|\hat{\mathcal{X}}| = k_0$  and it contains an  $\epsilon$ -cover of  $\mathcal{X}$  w.r.t.  $\|\cdot\|$ . The claims then follows from Theorem 2 using  $\mathcal{X}$ ,  $\hat{\mathcal{X}}$ ,  $\epsilon$ ,  $t_0 = 1$ ,  $t_1 = \sqrt{d}$ , and  $f \in \mathcal{F}_{\lambda, \mathcal{X}}$ .  $\blacksquare$

For any bounded set  $\mathcal{X}$ , the covering condition of Theorem 22 is satisfied with  $t = d$  by Theorem 7. This yields the approximation rate  $k_0^{-1/d}$ , which is known to be optimal (DeVore et al., 1989, Theorem 4.2).

An appealing property of the class  $\mathcal{M}_{k_0, l_0}$  is that it does not depend on the choice of center points  $\hat{\mathcal{X}}$ , unlike  $\mathcal{F}_{\infty-}(\hat{\mathcal{X}})$  or  $\mathcal{F}_\infty(\hat{\mathcal{X}})$ . In fact,  $\mathcal{F}_{\infty-}(\hat{\mathcal{X}}) \subset \mathcal{M}_{k_0, l_0}$  for any  $\hat{\mathcal{X}} \subset \mathbb{R}^d$  with  $k_0 \geq |\hat{\mathcal{X}}|$  and  $l_0 \geq 2d$ . However,  $\mathcal{M}_{k_0, l_0}$  uses at least  $d$  times more parameters than  $\mathcal{F}_\infty(\hat{\mathcal{X}})$ , specifically at least  $2d|\hat{\mathcal{X}}|(d+1)$  versus  $|\hat{\mathcal{X}}|(d+2) + |\hat{\mathcal{X}}|d$ . Moreover, we are not aware of any tractable algorithm for solving the nonconvex ERM problem over the full class  $\mathcal{M}_{k_0, l_0}$ . Only heuristic methods have been proposed (e.g., Bagirov et al., 2010, 2022).

The DCF algorithm (Algorithm 2) addresses this gap in the presented nonparametric setting. Using  $\triangleright = \infty$  and additional linear constraints  $v_{f,k} \leq 0$  for all  $k \in [K]$ , expressed as  $\mathbf{w}_k^\top = [\mathbf{u}_k^\top \ v_k]$  with  $\mathbf{u}_k \in \mathbb{R}^d$  and  $v_k \leq 0$  in (3), DCF computes an estimator  $f_n \in \mathcal{F}_{\infty-}(\hat{\mathcal{X}}_K)$  in polynomial-time, where the set  $\hat{\mathcal{X}}_K$  is computed by AFPC. This  $f_n$  estimator achieves the near-minimax rate of Theorem 1 and can be converted to an equivalent representation  $m_n \in \mathcal{M}_{K,2d}$ . The final refinement step (4) is performed directly over  $\mathcal{M}_{K,2d}$ , initialized at  $m_n$ , and the resulting estimator continues to satisfy the near-minimax rate of Theorem 1. Our proof adapts to this case (and in fact simplifies) since  $\mathcal{M}_{K,2d}$  depends not on the random covariates  $\mathcal{X}_n$ , but only on  $K \leq k_*$ . This leads to a substantial simplification of Theorems 19 and 20, as we only need to construct a cover of the (non-random) class  $\mathcal{M}_{k_*,2d}$ , which follows straightforwardly from Theorem 10 after bounding the parameter space. The convergence rate bound for  $m_n$  scales with an additional factor of  $d$  relative to that of Theorem 1, due to the larger number of parameters in  $m_n$  compared to  $f_n$ . Finally, the symmetrization of this max-min-affine estimator can be carried out in a similar way as for the other DCF variants, as described in Section 6.1.

## 5.2 Approximation of smooth functions

We now present an approximation result analogous to Theorem 2 for smooth functions, which we use in Section 5.3 to establish uniform approximation results for certain delta-convex classes.

For some  $\nu > 0$ , we say that a function  $f : \mathcal{X} \rightarrow \mathbb{R}$  is  $\nu$ -smooth on  $\mathcal{X}$  w.r.t.  $\|\cdot\|$  if it is differentiable on  $\mathcal{X} \subseteq \mathbb{R}^d$ , and its gradient  $\nabla f : \mathcal{X} \rightarrow \mathbb{R}^d$  is  $\nu$ -Lipschitz on  $\mathcal{X}$  w.r.t.  $\|\cdot\|$ , that is  $\|\nabla f(\mathbf{x}) - \nabla f(\hat{\mathbf{x}})\| \leq \nu \|\mathbf{x} - \hat{\mathbf{x}}\|$  holds for all  $\mathbf{x}, \hat{\mathbf{x}} \in \mathcal{X}$ . Denote the class of  $\nu$ -smooth functions on  $\mathcal{X}$  w.r.t.  $\|\cdot\|$  by  $\mathcal{F}_{\nu,\mathcal{X}}^\nabla$ .

Then consider the following uniform approximation bound for smooth functions:

**Theorem 23** *Let  $\mathcal{X}_\epsilon \subseteq \mathcal{X}$  be an  $\epsilon$ -cover of a convex set  $\mathcal{X} \subset \mathbb{R}^d$  w.r.t.  $\|\cdot\|$ . Let  $f \in \mathcal{F}_{\nu,\mathcal{X}}^\nabla$  for some constant  $\nu > 0$ , and define  $\tilde{f}_1(\mathbf{x}) \doteq \max_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} f(\hat{\mathbf{x}}) + \nabla f(\hat{\mathbf{x}})^\top (\mathbf{x} - \hat{\mathbf{x}}) - \nu \|\mathbf{x} - \hat{\mathbf{x}}\|^2$  for all  $\mathbf{x} \in \mathbb{R}^d$ . Then,  $0 \leq f(\mathbf{x}) - \tilde{f}_1(\mathbf{x}) \leq 2\nu\epsilon^2$  for all  $\mathbf{x} \in \mathcal{X}$ .*

**Proof** Choose  $\mathbf{x} \in \mathcal{X}$  arbitrarily. Because  $\mathcal{X}$  is a convex set, Taylor's theorem and  $f \in \mathcal{F}_{\nu,\mathcal{X}}^\nabla$  yield for all  $\hat{\mathbf{x}} \in \mathcal{X}_\epsilon$  that  $f(\mathbf{x}) = f(\hat{\mathbf{x}}) + \nabla f(t_{\hat{\mathbf{x}}}\mathbf{x} + (1 - t_{\hat{\mathbf{x}}})\hat{\mathbf{x}})^\top (\mathbf{x} - \hat{\mathbf{x}})$  for some  $t_{\hat{\mathbf{x}}} \in [0, 1]$ . Then by the Cauchy-Schwartz inequality, we get

$$\begin{aligned} \tilde{f}_1(\mathbf{x}) - f(\mathbf{x}) &= \max_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} f(\hat{\mathbf{x}}) - f(\mathbf{x}) + \nabla f(\hat{\mathbf{x}})^\top (\mathbf{x} - \hat{\mathbf{x}}) - \nu \|\mathbf{x} - \hat{\mathbf{x}}\|^2 \\ &\leq \max_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} (t_{\hat{\mathbf{x}}} - 1) \nu \|\mathbf{x} - \hat{\mathbf{x}}\|^2 \\ &\leq 0. \end{aligned}$$

For the other side, we have  $\min_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} \|\mathbf{x} - \hat{\mathbf{x}}\| \leq \epsilon$  by the  $\epsilon$ -covering property. Then by Taylor's theorem,  $f \in \mathcal{F}_{\nu,\mathcal{X}}^\nabla$ , and the Cauchy-Schwartz inequality, we get

$$\begin{aligned} f(\mathbf{x}) - \tilde{f}_1(\mathbf{x}) &= \min_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} f(\mathbf{x}) - f(\hat{\mathbf{x}}) - \nabla f(\hat{\mathbf{x}})^\top (\mathbf{x} - \hat{\mathbf{x}}) + \nu \|\mathbf{x} - \hat{\mathbf{x}}\|^2 \\ &\leq \min_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} (t_{\hat{\mathbf{x}}} + 1) \nu \|\mathbf{x} - \hat{\mathbf{x}}\|^2 \\ &\leq 2\nu\epsilon^2, \end{aligned}$$

which proves the claim.  $\blacksquare$

The function  $\tilde{f}_1$  of Theorem 23 provides a lower approximation of  $f$ . Similarly, an upper approximation is given by  $\check{f}_1$ , defined as  $\check{f}_1(\mathbf{x}) \doteq \min_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} f(\hat{\mathbf{x}}) + \nabla f(\hat{\mathbf{x}})^\top (\mathbf{x} - \hat{\mathbf{x}}) + \nu \|\mathbf{x} - \hat{\mathbf{x}}\|^2$  for all  $\mathbf{x} \in \mathbb{R}^d$ .

Note that  $\tilde{f}_1$  and  $\check{f}_1$  use the quadratic feature  $\|\cdot\|^2$ , in contrast to the norm feature  $\|\cdot\|$  used by the functions  $\tilde{f}$  and  $\check{f}$  in Section 4.1. The next result shows that the quadratic feature  $\|\cdot\|^2$  can also be used to approximate non-smooth Lipschitz functions.

**Theorem 24** *Let  $\mathcal{X}_\epsilon \subseteq \mathcal{X}$  be an  $\epsilon$ -cover of a set  $\mathcal{X} \subset \mathbb{R}^d$  w.r.t.  $\|\cdot\|$ . Suppose that  $f \in \mathcal{F}_{\lambda, \mathcal{X}}$  for some Lipschitz constant  $\lambda > 0$ , and define  $\tilde{f}_0(\mathbf{x}) \doteq \max_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} f(\hat{\mathbf{x}}) - (\lambda/\epsilon) \|\mathbf{x} - \hat{\mathbf{x}}\|^2$  for all  $\mathbf{x} \in \mathbb{R}^d$ . Then,  $-\lambda\epsilon/4 \leq f(\mathbf{x}) - \tilde{f}_0(\mathbf{x}) \leq 2\lambda\epsilon$  for all  $\mathbf{x} \in \mathcal{X}$ .*

**Proof** Choose  $\mathbf{x} \in \mathcal{X}$  arbitrarily. By  $f \in \mathcal{F}_{\lambda, \mathcal{X}}$ , we have

$$\tilde{f}_0(\mathbf{x}) - f(\mathbf{x}) = \max_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} f(\hat{\mathbf{x}}) - f(\mathbf{x}) - (\lambda/\epsilon) \|\mathbf{x} - \hat{\mathbf{x}}\|^2 \leq \max_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} \lambda \|\mathbf{x} - \hat{\mathbf{x}}\| \left(1 - \frac{\|\mathbf{x} - \hat{\mathbf{x}}\|}{\epsilon}\right) \leq \frac{\lambda\epsilon}{4}.$$

For the other side, we have  $\min_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} \|\mathbf{x} - \hat{\mathbf{x}}\| \leq \epsilon$  by the  $\epsilon$ -covering property. The claim then follows from  $f \in \mathcal{F}_{\lambda, \mathcal{X}}$  as  $f(\mathbf{x}) - \tilde{f}_0(\mathbf{x}) = \min_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} f(\mathbf{x}) - f(\hat{\mathbf{x}}) + (\lambda/\epsilon) \|\mathbf{x} - \hat{\mathbf{x}}\|^2 \leq 2\lambda\epsilon$ .  $\blacksquare$

Again, the max-concave approximation  $\tilde{f}_0$  has an analogous min-convex variant  $\check{f}_0$ , defined as  $\check{f}_0(\mathbf{x}) \doteq \min_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} f(\hat{\mathbf{x}}) + (\lambda/\epsilon) \|\mathbf{x} - \hat{\mathbf{x}}\|^2$  for all  $\mathbf{x} \in \mathbb{R}^d$ . Figure 6 illustrates the approximations of this section on the smooth and non-smooth examples from Figure 5.

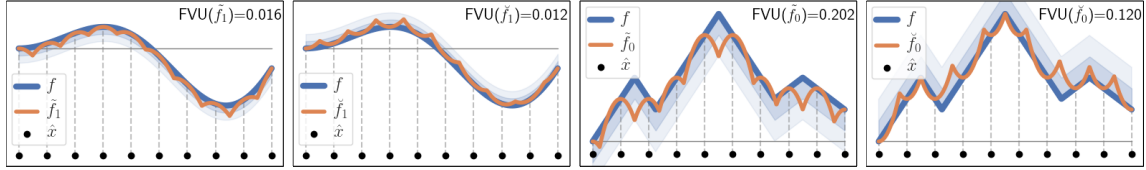


Figure 6: Approximation of a smooth function (left two plots) by  $\tilde{f}_1$  and  $\check{f}_1$  of Theorem 23, and of a non-smooth Lipschitz function (right two plots) by  $\tilde{f}_0$  and  $\check{f}_0$  of Theorem 24. The setting and notation are the same as in Figure 5, except that in the left two plots, the shaded areas indicate distances of  $\nu\epsilon^2$  and  $2\nu\epsilon^2$  from  $f$ .

Although it is straightforward to modify the DCF algorithm to use  $\|\cdot\|^2$  instead of  $\|\cdot\|$  in the function representation of  $\mathcal{F}_2(\hat{\mathcal{X}}_K)$ , the near-minimax guarantee of Theorem 1 does not carry over. Our proof of Theorem 1 does not extend to this case due to the challenge that the approximation functions  $\tilde{f}_0$  and  $\check{f}_1$  are only locally Lipschitz, and the coefficient of the quadratic feature  $\|\cdot\|^2$  in  $\tilde{f}_0$  grows as  $\lambda/\epsilon$ , which scales polynomially in  $n$  since  $\epsilon \approx n^{-1/(2+d_*)}$ . Addressing this issue remains an open direction for future research.

### 5.3 Weakly max-affine and delta-max-affine functions

The functions  $\tilde{f}_0$  and  $\check{f}_1$  of Theorems 23 and 24 are weakly convex in the sense of Vial (1983); that is, there exist  $s_0, s_1 \geq 0$  such that  $\tilde{f}_0 + s_0 q$  and  $\check{f}_1 + s_1 q$  are convex functions, where  $q$  is a symmetric convex quadratic function given by  $q(\mathbf{x}) \doteq \|\mathbf{x}\|^2$  for all  $\mathbf{x} \in \mathbb{R}^d$ .



Define the class of max-affine functions with at most  $k_0 \in \mathbb{N}$  hyperplanes by

$$\mathcal{M}_{k_0} \doteq \left\{ m : \mathbb{R}^d \rightarrow \mathbb{R} \mid m(\mathbf{x}) \doteq \max_{k \in [k_0]} b_k + \mathbf{w}_k^\top \mathbf{x}, \mathbf{x} \in \mathbb{R}^d, b_k \in \mathbb{R}, \mathbf{w}_k \in \mathbb{R}^d, k \in [k_0] \right\}.$$

Let the class of weakly max-affine functions be  $\mathcal{M}_{k_0}^w \doteq \{f \mid f \doteq m - sq, m \in \mathcal{M}_{k_0}, s \in \mathbb{R}\}$ . Furthermore, consider the closely related class of delta-max-affine functions, defined by  $\mathcal{M}_{k_0}^\Delta \doteq \{f \mid f \doteq m_1 - m_2, m_1, m_2 \in \mathcal{M}_{k_0}\}$ . The next result provides uniform approximation bounds for both of these classes. For convenience, let  $\mathcal{F}_{\infty, \mathcal{X}} \doteq \mathcal{F}_{\infty, \mathcal{X}}^\nabla \doteq \{f \mid f : \mathcal{X} \rightarrow \mathbb{R}\}$ .

**Corollary 25** *Let  $\mathcal{X} \subset \mathbb{R}^d$  be a convex set, and suppose there exist  $r, t > 0$  such that  $N_{\|\cdot\|}(\mathcal{X}, \epsilon) \leq (r/\epsilon)^t$  for all  $\epsilon \in (0, r]$ . Let  $f \in \mathcal{F}_{\lambda, \mathcal{X}} \cap \mathcal{F}_{\nu, \mathcal{X}}^\nabla$  for some constants  $\lambda, \nu \in (0, \infty]$ . Then, for all  $k_0 \in \mathbb{N}$ , there exist functions  $f_w \in \mathcal{M}_{k_0}^w$  and  $f_\Delta \in \mathcal{M}_{k_0}^\Delta$  such that*

$$\|f_w - f\|_{\infty, \mathcal{X}} \leq 2\epsilon \min\{\lambda, \nu\epsilon\}, \quad \|f_\Delta - f\|_{\infty, \mathcal{X}} \leq 3\epsilon \min\{\lambda, \nu\epsilon\}, \quad \epsilon \doteq rk_0^{-1/t}.$$

**Proof** The choice  $\epsilon = rk_0^{-1/t}$  ensures that  $N_{\|\cdot\|}(\mathcal{X}, \epsilon) \leq (r/\epsilon)^t = k_0$ , which allows us to choose  $\hat{\mathcal{X}} \subset \mathbb{R}^d$  with  $|\hat{\mathcal{X}}| = k_0$  such that it contains an  $\epsilon$ -cover of  $\mathcal{X}$  w.r.t.  $\|\cdot\|$ . Define the following weakly max-affine functions for all  $\mathbf{x} \in \mathbb{R}^d$ :

$$\begin{aligned} \tilde{f}_0(\mathbf{x}) &\doteq m_0(\mathbf{x}) - (\lambda/\epsilon)q(\mathbf{x}), \quad m_0(\mathbf{x}) \doteq \max_{\hat{\mathbf{x}} \in \hat{\mathcal{X}}} f(\hat{\mathbf{x}}) - (\lambda/\epsilon)\|\hat{\mathbf{x}}\|^2 + 2(\lambda/\epsilon)\hat{\mathbf{x}}^\top \mathbf{x}, \\ \tilde{f}_1(\mathbf{x}) &\doteq m_1(\mathbf{x}) - \nu q(\mathbf{x}), \quad m_1(\mathbf{x}) \doteq \max_{\hat{\mathbf{x}} \in \hat{\mathcal{X}}} f(\hat{\mathbf{x}}) - \nu(\|\hat{\mathbf{x}}\|^2 - \nabla f(\hat{\mathbf{x}})^\top \hat{\mathbf{x}}) + (\nabla f(\hat{\mathbf{x}}) + 2\nu\hat{\mathbf{x}})^\top \mathbf{x}. \end{aligned}$$

Clearly,  $\tilde{f}_0, \tilde{f}_1 \in \mathcal{M}_{k_0}^w$  and they coincide with the functions constructed in Theorems 23 and 24, respectively. Therefore, the claimed upper bound on  $\|f_w - f\|_{\infty, \mathcal{X}}$  follows directly.

Next, we approximate the quadratic function  $q$  by a max-affine function formed as the maximum of the first-order Taylor approximations of  $q$  at the points in  $\hat{\mathcal{X}}$ . Define  $\hat{m}(\mathbf{x}) \doteq \max_{\hat{\mathbf{x}} \in \hat{\mathcal{X}}} -\|\hat{\mathbf{x}}\|^2 + 2\hat{\mathbf{x}}^\top \mathbf{x}$  for all  $\mathbf{x} \in \mathbb{R}^d$ . Then, by the  $\epsilon$ -covering property of  $\hat{\mathcal{X}}$ , we have  $q(\mathbf{x}) - \hat{m}(\mathbf{x}) = \min_{\hat{\mathbf{x}} \in \hat{\mathcal{X}}} \|\mathbf{x} - \hat{\mathbf{x}}\|^2 \in [0, \epsilon^2]$  for all  $\mathbf{x} \in \mathcal{X}$ . Let  $\delta \doteq \min\{\lambda, \nu\epsilon\}$  and  $m_w \in \{m_0, m_1\}$  such that  $f_w \doteq m_w - (\delta/\epsilon)q \in \{\tilde{f}_0, \tilde{f}_1\}$  satisfies  $\|f_w - f\|_{\infty, \mathcal{X}} \leq 2\epsilon\delta$ . Define  $f_\Delta \doteq m_w - (\delta/\epsilon)\hat{m}$ , where  $f_\Delta \in \mathcal{M}_{k_0}^\Delta$ . The second claim then follows by using the triangle inequality as  $\|f_\Delta - f\|_{\infty, \mathcal{X}} \leq \|f_\Delta - f_w\|_{\infty, \mathcal{X}} + \|f_w - f\|_{\infty, \mathcal{X}} \leq (\delta/\epsilon)\epsilon^2 + 2\epsilon\delta = 3\epsilon\delta$ .  $\blacksquare$

For estimator design in regression, weakly max-affine functions were studied by Sun and Yu (2019), although without establishing convergence rates. Using delta-max-affine functions, Siahkamari et al. (2020) proposed an estimator and proved a suboptimal convergence rate for the case in which the regression function is delta-convex. Our proof techniques do not apply to either of these function classes, for the reasons explained at the end of Section 5.2. Nevertheless, Theorem 25 may be useful for extending these developments and designing near-minimax estimators for smooth regression functions, which lies beyond the scope of this work.

## 6 Variants of DCF

As mentioned in Section 2.2, one can train over any of the function classes  $\mathcal{F}_\triangleright(\cdot)$ ,  $\mathcal{F}_\triangleright^-(\cdot)$ , or  $\mathcal{F}_\triangleright^\Delta(\cdot)$  with  $\triangleright \in \{1, 2, \infty, +\}$  using DCF (Algorithm 2). While all these settings enjoy

the near-minimax guarantee of Theorem 1, they can differ significantly in empirical performance, as shown in Section 3. We further illustrate this difference on Figure 7 for the 1-dimensional examples discussed earlier in Figures 5 and 6. The plots show that  $\mathcal{F}_{\triangleright}(\cdot)$

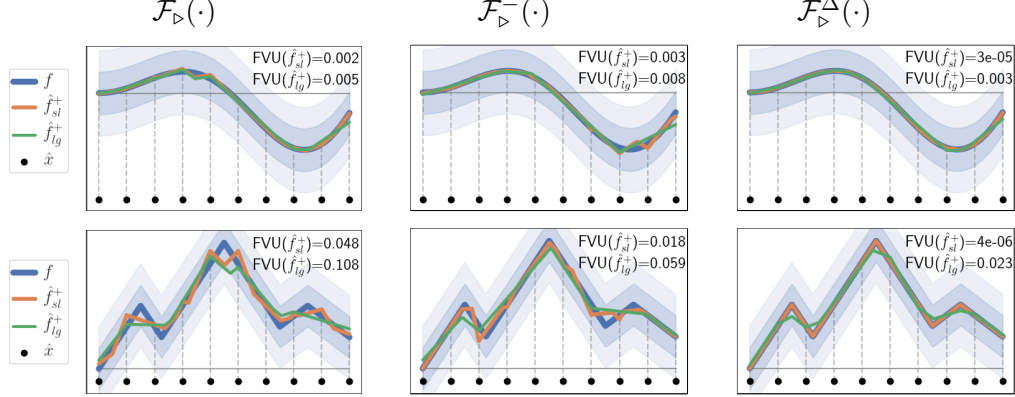


Figure 7: DCF approximations, each column showing the result for  $\mathcal{F}_{\triangleright}(\cdot)$ ,  $\mathcal{F}_{\triangleright}^{-}(\cdot)$ ,  $\mathcal{F}_{\triangleright}^{\Delta}(\cdot)$ , respectively. We used the norms  $\triangleright \in \{1, 2, \infty\}$  which are equivalent for  $d = 1$ . The settings and the notations are the same as on Figure 5. DCF uses the same parameters as in Section 3, with  $\theta_2 = (R_{\mathcal{X}_n}/n)^2$  for  $\hat{f}_{sl}^+$  and  $\theta_2 = R_{\mathcal{X}_n}^2/n$  for  $\hat{f}_{lg}^+$ .

struggles to approximate the concave regions, while  $\mathcal{F}_{\triangleright}^{-}(\cdot)$  struggles with convex regions. In contrast, the symmetric representation in  $\mathcal{F}_{\triangleright}^{\Delta}(\cdot)$  overcomes both issues and achieves significantly better approximation accuracy. In all cases, the accuracy is an order of magnitude better than that of the worst-case optimal approximation functions from Theorem 2 in Figure 5. The plots also illustrate how the choice of the regularizer  $\theta_2$  controls the “smoothness” of the estimator. In these noise-free settings, smaller values of  $\theta_2$  yield better results; however, in the noisy problems discussed in Section 3, such choices can lead to overfitting.

Recall the definition  $\mathcal{F}_{\triangleright}^{-}(\hat{\mathcal{X}}_K) \doteq \{-f : f \in \mathcal{F}_{\triangleright}(\hat{\mathcal{X}}_K)\}$  from Section 2.2. Training over  $\mathcal{F}_{\triangleright}^{-}(\hat{\mathcal{X}}_K)$  requires only a minor modification to DCF: one can flip the sign of the response variables  $Y_1, \dots, Y_n$  during training, and then flip the sign of the final estimator  $f_n^+$  at the end. The case of  $\mathcal{F}_{\triangleright}^{\Delta}$  is slightly more involved, and we discuss it in detail in the next section.

## 6.1 Symmetric representations

We now describe the modifications to DCF (Algorithm 2) needed to work with the symmetric class  $\mathcal{F}_{\triangleright}^{\Delta}(\hat{\mathcal{X}}_K) \doteq \{f \mid f \doteq f_1 - f_2, f_1, f_2 \in \mathcal{F}_{\triangleright}(\hat{\mathcal{X}}_K)\}$ .

As before, we compute a single set of center points  $\hat{\mathcal{X}}_K$  using AFPC (Algorithm 1). The main difference lies in modifying (3) to use two sets of parameters,  $\langle (b_k, \mathbf{w}_k) : k \in [K] \rangle$  and

$\langle (b'_k, \mathbf{w}'_k) : k \in [K] \rangle$ , as shown in the following:

$$\begin{aligned} \min_{\substack{z \in \mathbb{R}, \\ b_1, \dots, b_K \in \mathbb{R}, \\ \mathbf{w}_1, \dots, \mathbf{w}_K \in \mathbb{R}^{d_\triangleright}, \\ b'_1, \dots, b'_K \in \mathbb{R}, \\ \mathbf{w}'_1, \dots, \mathbf{w}'_K \in \mathbb{R}^{d_\triangleright}}} \theta_1 z^2 + \sum_{k \in [K]} \theta_2 (\|\mathbf{w}_k\|^2 + \|\mathbf{w}'_k\|^2) \\ + \frac{1}{n} \sum_{i \in [n]} \mathbb{I}\{\mathbf{X}_i \in \mathcal{C}_k(\hat{\mathcal{X}}_K)\} \left( b_k - b'_k + \phi_\triangleright(\mathbf{X}_i, \hat{\mathbf{X}}_k)^\top (\mathbf{w}_k - \mathbf{w}'_k) - y_i \right)^2 \end{aligned}$$

such that for all  $k, l \in [K] : b_k \geq b_l + \phi_\triangleright(\hat{\mathbf{X}}_k, \hat{\mathbf{X}}_l)^\top \mathbf{w}_l, \quad \|\mathbf{w}_k\| \leq z + \theta_0,$   
 $b'_k \geq b'_l + \phi_\triangleright(\hat{\mathbf{X}}_k, \hat{\mathbf{X}}_l)^\top \mathbf{w}'_l, \quad \|\mathbf{w}'_k\| \leq z + \theta_0.$

(35)

Let the solution of (35) be  $(z_n, \langle (b_{n,k}, \mathbf{w}_{n,k}, b'_{n,k}, \mathbf{w}'_{n,k}) : k \in [K] \rangle)$ , and define the (initial) estimator by  $f_n \doteq f_{n,1} - f_{n,2}$ , where  $f_{n,1}(\mathbf{x}) \doteq \max_{k \in [K]} b_{n,k} + \phi_\triangleright(\mathbf{x}, \hat{\mathbf{X}}_k)^\top \mathbf{w}_{n,k}$ , and  $f_{n,2}(\mathbf{x}) \doteq \max_{k \in [K]} b'_{n,k} + \phi_\triangleright(\mathbf{x}, \hat{\mathbf{X}}_k)^\top \mathbf{w}'_{n,k}$ , for all  $\mathbf{x} \in \mathbb{R}^d$ . Clearly,  $f_{n,1}, f_{n,2} \in \mathcal{F}_\triangleright(\hat{\mathcal{X}}_K)$ , and we have  $f_n \in \mathcal{F}_\triangleright^\Delta(\hat{\mathcal{X}}_K)$ .

Next, the final step refines the estimator  $f_n$  to  $\hat{f}_n^+ \in \mathcal{F}_\triangleright^\Delta(\hat{\mathcal{X}}_K)$  that satisfies (4), where the regularizer

$$\mathcal{R}_{c_0, c_1, c_2}(f) \doteq c_1 \max_{k \in [K], j \in [2]} (\|\mathbf{w}_{f_j, k}\| - c_0)_+^2 + c_2 \sum_{k \in [K], j \in [2]} \|\mathbf{w}_{f_j, k}\|^2 \quad (36)$$

is defined for all  $c_0, c_1, c_2 \geq 0$ , and for all  $f \doteq f_1 - f_2 \in \mathcal{F}_\triangleright^\Delta(\hat{\mathcal{X}}_K)$  with  $f_1, f_2 \in \mathcal{F}_\triangleright(\hat{\mathcal{X}}_K)$ . Here, we use  $\lambda_f \doteq \max_{k \in [K], j \in [2]} \|\mathbf{w}_{f_j, k}\|$  to define  $\mathcal{R}_n(\cdot)$  in (4).

Suppose the refined estimator  $\hat{f}_n^+$  is expressed as  $\hat{f}_n^+ \doteq \hat{f}_{n,1}^+ - \hat{f}_{n,2}^+$  with some functions  $\hat{f}_{n,1}^+, \hat{f}_{n,2}^+ \in \mathcal{F}_\triangleright(\hat{\mathcal{X}}_K)$ . Then, the final estimator  $f_n^+ \doteq \hat{f}_{n,1}^+ - \hat{f}_{n,2}^+$  is defined for all  $\mathbf{x} \in \mathbb{R}^d$  as

$$f_{n,j}^+(\mathbf{x}) \doteq \left(\frac{3}{2} - j\right) C_n^+ + \max_{k \in \mathcal{I}_{n,j}^+} b_{\hat{f}_{n,j}^+, k} - C_\Delta^+ + \phi(\mathbf{x}, \hat{\mathbf{X}}_k)^\top \mathbf{w}_{\hat{f}_{n,j}^+, k}, \quad j \in [2], \quad (37)$$

where  $\mathcal{I}_{n,j}^+ \doteq \mathcal{I}_n(\hat{f}_{n,j}^+)$ ,  $C_n^+$  and  $\mathcal{I}_n(\cdot)$  are defined as in (5), and the average bias term  $C_\Delta^+$  is given by  $C_\Delta^+ \doteq \sum_{j \in [2]} \frac{1}{|\mathcal{I}_{n,j}^+|} \sum_{k \in \mathcal{I}_{n,j}^+} b_{\hat{f}_{n,j}^+, k}$  for all  $j \in [2]$ .

Let  $b_{j,k}^+ \doteq \left(\frac{3}{2} - j\right) C_n^+ + b_{\hat{f}_{n,j}^+, k} - C_\Delta^+$  be the bias parameters used in defining  $f_n^+$ , for all  $j \in [2]$  and  $k \in \mathcal{I}_{n,j}^+$ . Define the average bias terms as  $\bar{b}_{n,j}^+ \doteq \frac{1}{|\mathcal{I}_{n,j}^+|} \sum_{k \in \mathcal{I}_{n,j}^+} b_{j,k}^+$  for all  $j \in [2]$ . Note that subtracting the constant  $C_\Delta^+$  from all the bias parameters leaves the function  $f_n^+$  unchanged and ensures that  $\bar{b}_{n,1}^+ + \bar{b}_{n,2}^+ = 0$ .

We claim that the theoretical guarantee of Theorem 1 extends to DCF estimators based on the symmetric set  $\mathcal{F}_\triangleright^\Delta(\cdot)$ . The derivation in Section 4 can be straightforwardly adapted to this case, except for one detail concerning the bounding of the unregularized bias parameters in Theorems 14 and 18. The techniques presented in Section 4.4 yield bounds on differences between pairs of bias parameters, rather than on individual ones. To decouple these bounds, we apply the following result.

**Lemma 26** *Let  $n, m \in \mathbb{N}$ , and  $a_1, \dots, a_n, b_1, \dots, b_m \in \mathbb{R}$ . Define  $\bar{a} \doteq (1/n) \sum_{i=1}^n a_i$  and  $\bar{b} \doteq (1/m) \sum_{j=1}^m b_j$ . Suppose that  $\bar{a} + \bar{b} = 0$ , and that there exist  $c \in \mathbb{R}$  and  $\beta > 0$  such that  $\max_{i \in [n], j \in [m]} |a_i - b_j - c| \leq \beta$ . Then,  $\max \{ \max_{i \in [n]} |a_i - c/2|, \max_{j \in [m]} |b_j + c/2| \} \leq 3\beta/2$ .*

**Proof** By Jensens's inequality, we get  $|\bar{a} - \bar{b} - c| \leq \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m |a_i - b_j - c| \leq \beta$ . Combining this with  $\bar{a} = -\bar{b}$  yields  $|\bar{a} - c/2| \leq \beta/2$  and  $|\bar{b} + c/2| \leq \beta/2$ . Then, by using the reverse triangle and Jensen's inequalities, we get for all  $i \in [n]$  that  $|a_i - c/2| - |\bar{b} + c/2| \leq |a_i - \bar{b} - c| \leq \beta$ , which implies  $\max_{i \in [n]} |a_i - c/2| \leq 3\beta/2$ . The other claim,  $\max_{j \in [m]} |b_j + c/2| \leq 3\beta/2$ , follows analogously.  $\blacksquare$

Similarly to the proof of Theorem 18, one can show that  $|b_{1,k}^+ - b_{2,l}^+ - y_0| \leq \beta_0$  for all  $k \in \mathcal{I}_{n,1}^+$  and  $l \in \mathcal{I}_{n,2}^+$ . Then, we separate these bounds using Theorem 26 with  $\bar{b}_{n,1}^+ + \bar{b}_{n,2}^+ = 0$ .

The bias parameters of  $f_n$  can be centered to satisfy  $0 = \sum_{k \in [K]} b_{n,k} + b'_{n,k}$  without changing the function. Analogously to the proof of Theorem 14, we get  $|b_{n,k} - b'_{n,k} - y_0| = O(\beta_1)$  for all  $k \in [K]$ . To extend this to all pairs, we use the constraints in (35). Specifically, for each pair  $k, l \in [K]$ , we have the constraints  $b_k \geq b_l + \phi_{\triangleright}(\hat{\mathbf{X}}_k, \hat{\mathbf{X}}_l)^\top \mathbf{w}_k$  and  $b_l \geq b_k + \phi_{\triangleright}(\hat{\mathbf{X}}_l, \hat{\mathbf{X}}_k)^\top \mathbf{w}_l$ , which imply  $|b_{n,k} - b_{n,l}| = O(\tau_{\phi_{\triangleright}} R_{\mathcal{X}_n} \max_{k' \in [K]} \|\mathbf{w}_{n,k'}\|) = O(\beta_1)$ . The same reasoning applies to  $\{b'_{n,k} : k \in [K]\}$ . We then use the triangle inequality to bound the distance between all pairs as  $|b_{n,k} - b'_{n,l} - y_0| = O(\beta_1)$  for all  $k, l \in [K]$ , and invoke Theorem 26. The remaining details are straightforward and omitted for brevity.

## 6.2 Adapting to convex shape-restricted regression

We now consider the setting of convex (shape-restricted) regression, where the  $\lambda_*$ -Lipschitz regression function  $f_*$  is known to be convex. The goal is to estimate  $f_*$  with a convex function.

Let  $\mathcal{F}_{\lambda, \mathcal{X}}^{\text{cvx}} \doteq \{f \in \mathcal{F}_{\lambda, \mathcal{X}} : f \text{ is convex}\}$  denote the set of  $\lambda$ -Lipschitz, convex functions over a convex set  $\mathcal{X} \subseteq \mathbb{R}^d$ . Let  $\nabla f(\mathbf{x})$  denote an arbitrary but fixed subgradient of  $f$  at  $\mathbf{x} \in \mathbb{R}^d$ . We consider the statistical model (1), modified so that  $f_* \in \mathcal{F}_{\lambda_*, \mathcal{X}_*}^{\text{cvx}}$ , for an unknown Lipschitz constant  $\lambda_* > 0$  and an unknown convex domain  $\mathcal{X}_* \subseteq \mathbb{R}^d$ . In this setting, DCF (Algorithm 2) can be applied by restricting  $\mathcal{F}_{\triangleright}(\hat{\mathcal{X}}_K)$  to convex functions, where the center points  $\hat{\mathcal{X}}_K \doteq \{\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_K\}$  are computed by AFPC (Algorithm 1).

The max-affine representation is by far the most widely used in convex regression. DCF can be readily adapted to this case by disabling the “norm feature” and restricting  $\mathcal{F}_{\triangleright}(\hat{\mathcal{X}}_K)$  to the class of max-affine functions as

$$\mathcal{M}_K = \left\{ f \in \mathcal{F}_{\triangleright}(\hat{\mathcal{X}}_K) : \mathbf{w}_{f,k}^\top = [\mathbf{u}_{f,k}^\top \ v_{f,k}], \mathbf{u}_{f,k} \in \mathbb{R}^d, v_{f,k} = 0, k \in [K] \right\},$$

which holds for all  $\triangleright \in \{1, 2, \infty\}$ .

To adapt the proof of Theorem 1 to the convex regression setting, we replace Theorem 2 with the following approximation result, which slightly extends the result of Balázs (2022).

**Theorem 27** *Let  $\mathcal{X}_\epsilon$  be an  $\epsilon$ -cover of a convex set  $\mathcal{X} \subset \mathbb{R}^d$ . Suppose  $f \in \mathcal{F}_{\lambda, \mathcal{X}}^{\text{cvx}}$  for some Lipschitz constant  $\lambda > 0$ , and let  $\hat{m}(\mathbf{x}) \doteq \max_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} f(\hat{\mathbf{x}}) + \nabla f(\hat{\mathbf{x}})^\top (\mathbf{x} - \hat{\mathbf{x}})$  for all  $\mathbf{x} \in \mathbb{R}^d$ . Then, for all  $\mathbf{x} \in \mathcal{X}$  and  $\tilde{\mathbf{x}} \in \mathcal{X}_\epsilon$ ,*

$$0 \leq f(\mathbf{x}) - \hat{m}(\mathbf{x}) \leq 2\lambda\epsilon, \quad \hat{m}(\tilde{\mathbf{x}}) = f(\tilde{\mathbf{x}}), \quad \hat{m} \in \mathcal{F}_{\lambda, \mathbb{R}^d}^{\text{cvx}} \cap \mathcal{M}_{|\mathcal{X}_\epsilon|}.$$

**Proof** The convexity of  $f$  directly implies  $0 \leq f - \hat{m}$ . Moreover, since  $f \in \mathcal{F}_{\lambda, \mathcal{X}}$ , we get for all  $\mathbf{x} \in \mathcal{X}$  that

$$f(\mathbf{x}) - \hat{m}(\mathbf{x}) = \min_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} f(\mathbf{x}) - f(\hat{\mathbf{x}}) - \nabla f(\hat{\mathbf{x}})^\top (\mathbf{x} - \hat{\mathbf{x}}) \leq 2\lambda \min_{\hat{\mathbf{x}} \in \mathcal{X}_\epsilon} \|\mathbf{x} - \hat{\mathbf{x}}\|, \quad (38)$$

which implies  $f(\mathbf{x}) - \hat{m}(\mathbf{x}) \leq 2\lambda\epsilon$  for all  $\mathbf{x} \in \mathcal{X}$  by the  $\epsilon$ -covering property of  $\mathcal{X}_\epsilon$ . Furthermore, choosing  $\mathbf{x} = \tilde{\mathbf{x}} \in \mathcal{X}_\epsilon$  in (38), we get  $f(\tilde{\mathbf{x}}) - \hat{m}(\tilde{\mathbf{x}}) \leq 0$ , implying the second claim. The third claim follows because the max function is 1-Lipschitz, and  $\|\nabla f(\hat{\mathbf{x}})\| \leq \lambda$  for all  $\hat{\mathbf{x}} \in \mathcal{X}_\epsilon$  as  $\mathcal{X}_\epsilon \subseteq \mathcal{X}$  and  $f \in \mathcal{F}_{\lambda, \mathcal{X}}$ . ■

Since Theorem 27 delivers the same  $O(\lambda\epsilon)$  approximation accuracy as Theorem 2, one may apply DCF with the class of max-affine functions  $\mathcal{M}_K$  to achieve the near-minimax rate of Theorem 1 in the convex regression setting. This matches the theoretical guarantees of the APCNLS algorithm of Balázs (2022), but unlike APCNLS, DCF does not require knowledge of the Lipschitz constant  $\lambda_*$ . Moreover, the optimization problem (3) in DCF uses  $K^2$  constraints, which provides a substantial reduction in computational burden compared to the  $nK$  constraints used in APCNLS.

Motivated by the experimental results of Section 3, it may be beneficial to use a richer function class than the max-affine one. In particular, one can allow  $v_{f,k} \geq 0$  for all  $k \in [K]$  instead of enforcing  $v_{f,k} = 0$  as in (6.2), which still ensures that the set  $\mathcal{F}_\triangleright(\hat{\mathcal{X}}_K)$  is restricted to convex functions for all  $\triangleright \in \{1, 2, \infty\}$ . Moreover, one can restrict the set  $\mathcal{F}_+(\hat{\mathcal{X}}_K)$  to convex functions using the following simple result:

**Lemma 28** *Fix  $\hat{x}, u, v \in \mathbb{R}$ , and define  $f(x) \doteq (x - \hat{x})_+ u + (\hat{x} - x)_+ v$  for all  $x \in \mathbb{R}$ . Then  $f$  is convex over  $\mathbb{R}$  if and only if  $u \geq -v$ .*

**Proof** Write  $f(x) = (x - \hat{x})_+(u + v) - ((x - \hat{x})_+ - (\hat{x} - x)_+)v = (x - \hat{x})_+(u + v) - (x - \hat{x})_+ v$  which is convex over  $\mathbb{R}$  if  $u + v \geq 0$  and concave otherwise by the convexity of  $(\cdot)_+$ . ■

Then, by Theorem 28, the above mentioned restriction to convex functions can be formalized by using an extra linear constraint as

$$\mathcal{F}_+^{\text{cvx}}(\hat{\mathcal{X}}_K) \doteq \left\{ f \in \mathcal{F}_+(\hat{\mathcal{X}}_K) : \mathbf{w}_{f,k}^\top = [\mathbf{u}_{f,k}^\top \ \mathbf{v}_{f,k}^\top], \ \mathbf{u}_{f,k} \geq -\mathbf{v}_{f,k}, \ \mathbf{u}_{f,k}, \mathbf{v}_{f,k} \in \mathbb{R}^d, \ k \in [K] \right\}.$$

A detailed analysis of when these extended convex function classes provide performance gains compared to max-affine functions is left for future research.

## 7 Conclusions

We introduced the polynomial-time DCF algorithm, which decomposes the nonparametric estimation of a Lipschitz function into three stages: a partitioning step, a convex initial fitting step over the resulting partition, and an optional refinement step applied to the initial solution. As shown in Theorem 1, DCF achieves the adaptive near-minimax rate in our setting, capturing the doubling dimension of the covariate space without relying on external model selection techniques (e.g., Bartlett et al., 2002).

Our empirical results show that DCF, when equipped with an appropriately chosen regularization parameter  $\theta_2$ , can be competitive with state-of-the-art methods and can outperform other theoretically justified algorithms. However, its sensitivity to  $\theta_2$ , together with its computationally intensive training procedure, highlights an important direction for future research.

## Acknowledgments

We thank Csaba Szepesvári for many helpful comments on an early draft of the paper. This work was funded by G&G (Mariana Gema and the author).

## References

- Necdet Serhat Aybat and Zi Wang. A parallelizable dual smoothing method for large scale convex regression problems. *arXiv preprint arXiv:1608.02227*, 2016.
- Adil Bagirov, Conny Clausen, and Michael Kohler. An algorithm for the estimation of a regression function by continuous piecewise linear functions. *Computational Optimization and Applications*, 45(1):159–179, 2010.
- Adil M. Bagirov, Sona Taheri, Napsu Karmita, Nargiz Sultanova, and Soodabeh Asadi. Robust piecewise linear L1-regression via nonsmooth DC optimization. *Optimization Methods and Software*, 37(4):1289–1309, 2022.
- Gábor Balázs. *Convex Regression: Theory, Practice, and Applications*. PhD thesis, University of Alberta, 2016.
- Gábor Balázs. Adaptively partitioning max-affine estimators for convex regression. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 860–874. PMLR, 2022.
- Gábor Balázs, András György, and Csaba Szepesvári. Near-optimal max-affine estimators for convex regression. In Guy Lebanon and S. V. N. Vishwanathan, 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, 2015. PMLR.
- Gábor Balázs, András György, and Csaba Szepesvári. Chaining bounds for empirical risk minimization. *arXiv preprint arXiv:1609.01872v1*, 2016.
- Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.
- Peter L. Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48(1):85–113, 2002.
- Jose Blanchet, Peter W Glynn, Jun Yan, and Zhengqing Zhou. Multivariate distributionally robust convex regression under absolute error loss. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

- Stephen Boyd and Lieven Vandenbergh. *Convex Optimization*. Cambridge University Press, 2004.
- Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. Association for Computing Machinery, 2016. doi: 10.1145/2939672.2939785.
- Wenyu Chen and Rahul Mazumder. Subgradient regularized multivariate convex regression at scale. *SIAM Journal on Optimization*, 34(3):2350–2377, 2024.
- Sanjoy Dasgupta and Yoav Freund. Random projection trees and low dimensional manifolds. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, page 537–546, New York, NY, USA, 2008. Association for Computing Machinery.
- Ronald A. DeVore, Ralph Howard, and Charles Micchelli. Optimal nonlinear approximation. *Manuscripta Mathematica*, 63(4):469–478, 1989.
- Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.
- Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout Networks. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1319–1327. PMLR, 2013.
- Paul J. Goulart and Yuwen Chen. Clarabel: An interior-point solver for conic programs with quadratic objectives. *arXiv preprint arXiv:2405.12762*, 2024.
- A. Gupta, R. Krauthgamer, and J.R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 534–543, 2003.
- László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.
- Qiyang Han and Jon A. Wellner. Multivariate convex regression: Global risk bounds and adaptation. *arXiv preprint arXiv:1601.06844v1*, 2016.
- Philip Hartman. On functions representable as a difference of convex functions. *Pacific Journal of Mathematics*, 9(3):707–713, 1959.
- J.-B Hiriart-Urruty. Extension of Lipschitz functions. *Journal of Mathematical Analysis and Applications*, 77(2):539–554, 1980.
- Jean-Baptiste Hiriart-Urruty. Generalized differentiability, duality and optimization for problems dealing with differences of convex functions. *Convexity and Duality in Optimization*, pages 37–70, 1985.



- Dorit S. Hochbaum and David B. Shmoys. A best possible heuristic for the  $k$ -center problem. *Mathematics of Operations Research*, 10(2):180–184, 1985.
- Samory Kpotufe. *The curse of dimension in nonparametric regression*. PhD thesis, University of California, 2010.
- Samory Kpotufe and Sanjoy Dasgupta. A tree-based regressor that adapts to intrinsic dimension. *Journal of Computer and System Sciences*, 78(5):1496–1515, 2011.
- S.R. Kulkarni and S.E. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41(4):1028–1039, 1995.
- Gil Kur, Fuchang Gao, Adityanand Guntuboyina, and Bodhisattva Sen. Convex regression in multidimensions: Suboptimality of least squares estimators. *The Annals of Statistics*, 52(6):2791–2815, 2024.
- Eunji Lim. On convergence rates of convex regression in multiple dimensions. *INFORMS Journal on Computing*, 26(3):616–628, 2014.
- Eunji Lim. Convex regression with a penalty. *arXiv preprint arXiv:2509.19788v1*, 2025.
- E. J. McShane. Extension of range of functions. *Bulletin of the American Mathematical Society*, 40(12):837–842, 1934.
- E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.
- Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.
- Yurii Nesterov and Arkadii Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics, 1994.
- Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2006.
- Sergei Ovchinnikov. Max-min representation of piecewise linear functions. *Contributions to Algebra and Geometry*, 43(1):297–302, 2002.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.
- Carlos Ramos-Carreño, José L. Torrecilla, Miguel Carbajo Berrocal, Pablo Marcos Manchón, and Alberto Suárez. scikit-fda: A Python Package for Functional Data Analysis. *Journal of Statistical Software*, 109(2):1–37, 2024.
- Ali Siahkamari, Aditya Gangrade, Brian Kulis, and Venkatesh Saligrama. Piecewise linear regression via a difference of convex functions. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8895–8904. PMLR, 2020.

- Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.
- Sun Sun and Yaoliang Yu. Least squares estimation of weakly convex functions. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2271–2280. PMLR, 2019.
- Alejandro Toriello and Juan Pablo Vielma. Fitting piecewise linear continuous functions. *European Journal of Operational Research*, 219(1):86–95, 2012.
- Sara van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.
- Jean-Philippe Vial. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8(2):231–259, 1983.
- Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.
- Geoffrey S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372, 1964.
- C. T. Zahn. Black box maximization of circular coverage. *Journal of Research of the National Bureau of Standards – B. Mathematics and Mathematical Physics*, 66B(4):181–216, 1962.