An Efficient Minimax Optimal Estimator For Multivariate Convex Regression

Gil Kur GILKUR@MIT.EDU
Massachusetts Institute of Technology

Eli Putterman PUTTERMAN@MAIL.TAU.AC.IL
Tel Aviv University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We study the computational aspects of the task of multivariate convex regression in dimension $d \geq 5$. We present the first computationally efficient minimax optimal (up to logarithmic factors) estimators for the tasks of L-Lipschitz and Γ-bounded convex regression under polytopal support. This work is the first to show the existence of efficient minimax optimal estimators for non-Donsker classes whose corresponding Least Squares Estimators are provably minimax suboptimal. The proof of the correctness of these estimators uses a variety of tools from different disciplines, among them empirical process theory, stochastic geometry, and potential theory.

Keywords: Multivariate Convex Regression; Minimax Optimality; Non-Donsker Regime;

1. Introduction and Main Results

In this paper, we consider the following well-specified regression model in the random design setting:

$$Y = f^*(X) + \xi$$

where $f^* : \Omega \to \mathbb{R}$ lies in a known function class $\mathcal{F}$, X is drawn from a *known* distribution $\mathbb{P}$ on Ω, and ξ is a zero-mean noise with a finite variance σ^2.

In this task, given $(\mathcal{F}, \mathbb{P})$ and n i.i.d. observations $\mathcal{D} := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, we aim to estimate the underlying function f^* as well as possible with respect to the classical minimax risk (Tsybakov, 2003b). More precisely, define an estimator as a computable function that for any realization of the input $\mathcal{D}$, outputs some measurable function on Ω (denote by $\mathcal{M}(\Omega)$ the class of such functions). The minimax risk of such an estimator $\bar{f} : \mathcal{D} \to \mathcal{M}(\Omega)$ is defined as

$$\mathcal{R}_n(\bar{f}, \mathbb{P}, \mathcal{F}) := \sup_{f^* \in \mathcal{F}} \mathbb{E}_{\mathcal{D}} \int (f^* - \bar{f})^2 d\mathbb{P},$$

and the minimax rate of $\mathcal{F}$ is defined by $\mathcal{M}_n(\mathbb{P}, \mathcal{F}) := \inf_{\bar{f}} \mathcal{R}_n(\bar{f}, \mathbb{P}, \mathcal{F})$. Our goal is to find an estimator that is *minimax optimal* up to logarithmic factors in n, i.e. its risk satisfies

$$\mathcal{R}_n(\bar{f}, \mathbb{P}, \mathcal{F}) = \tilde{O}_{\mathbb{P},\mathcal{F}}(\mathcal{M}_n(\mathbb{P}, \mathcal{F})) := O_{\mathbb{P},\mathcal{F}}(\mathcal{M}_n(\mathbb{P}, \mathcal{F}) \cdot \log(n)^{O_{\mathbb{P},\mathcal{F}}(1)}),$$

where $O_{\mathbb{P},\mathcal{F}}$ denotes equality up to a multiplicative constant that depends only on $\mathbb{P}, \mathcal{F}$. We would also like our estimator to be *efficiently computable*, which for our purposes means that its runtime $R_{\hat{f}}(n)$ is polynomial in the number of samples: $R_{\hat{f}}(n) = O_{\mathbb{P},\mathcal{F}}(n^{O_{\mathbb{P},\mathcal{F}}(1)})$.

In this paper, we take $\mathcal{F}$ to be one of the following two function classes, which are subsets of the class of convex functions on a domain:

1. $\mathcal{F}_L(P)$, the class of convex L-Lipschitz functions supported on a convex polytope $P \subseteq B_d$, where B_d denotes the unit (Euclidean) ball in dimension d.

2. $\mathcal{F}^{\Gamma}(P)$, the class of convex functions on a convex polytope $P \subset B_d$ with range contained in $[-\Gamma, \Gamma]$.

These tasks are known as L-Lipschitz convex regression (Seijo and Sen, 2011) and Γ-bounded convex regression (Han and Wellner, 2016), respectively. For reasons which will become apparent later, we always take $d \geq 5$. In our work, we further assume that $\mathbb{P}$ satisfies the following:

Assumption 1 $\mathbb{P}$ *is uniformly bounded on its support by some positive constants* $c(d), C(d)$ *that only depend on* d, *i.e.* $c(d) \leq \frac{d\mathbb{P}}{dx}(x) \leq C(d)$, *for all* $x \in P \subset B_d$.

Convex regression tasks have been a central concern in the "shape-constrained" statistics literature (Devroye and Lugosi, 2012), and have innumerable applications in a variety of disciplines, from economic theory (Varian, 1982) to operations research (Powell and Topaloglu, 2003) and more (Balázs, 2016). In general, convexity is extensively studied in pure mathematics (Artstein-Avidan et al., 2015), computer science (Lovász and Vempala, 2007), and optimization (Boyd et al., 2004). We remark that there is a density-estimation counterpart of the convex regression problem, known as log-concave density estimation (Samworth, 2018; Cule et al., 2010), and these two tasks are closely related (Kur et al., 2019; Kim and Samworth, 2016).

Due to the appearance of convex regression in various fields, it has been studied from many perspectives and by many different communities. For example, in the mathematical statistics literature the minimax rates of convex regression tasks and the risk of the maximum likelihood estimator (MLE) are the main areas of interest; an incomplete sample of works treating this problem is (Guntuboyina, 2012; Guntuboyina and Sen, 2013; Gardner, 1995; Gao and Wellner, 2017; Kur et al.; Han, 2019; Brunel, 2013; Diakonikolas et al., 2018, 2016; Carpenter et al., 2018; Kur et al., 2019; Balázs et al., 2015). In operations research, work has focused on the algorithmic aspects of convex regression, i.e., finding scalable and efficient algorithms; see, e.g., (Ghosh et al., 2021; Brunel, 2016; O'Reilly and Chandrasekaran, 2021; Soh and Chandrasekaran, 2021; Balázs, 2016; Mazumder et al., 2019; Chen and Mazumder, 2020; Bertsimas and Mundru, 2021; Siahkamari et al., 2021; Simchowitz et al., 2018; Hannah and Dunson, 2012; Blanchet et al., 2019; Lin et al., 2020; Chen et al., 2021; Balázs, 2022). Initially, convex regression was mostly studied in the univariate case, which is now considered to be well-understood. Multivariate convex regression has only begun to be explored in recent years, and is still an area of active research.

The naïve algorithm for any variant of the convex regression task is the least squares estimator (LSE), which is also the MLE under Gaussian noise, defined by

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

where $\mathcal{F} = \mathcal{F}_L(\Omega)$ or $\mathcal{F} = \mathcal{F}^{\Gamma}(P)$ in our convex regression tasks. From a computational point of view, the LSE can be formulated as a quadratic programming problem with $O(n^2)$ constraints, and is thus efficiently computable in our terms (Seijo and Sen, 2011; Han and Wellner, 2016); however, the LSE has been seen empirically not to be scalable for large number of samples (Chen and Mazumder, 2020).

From a statistical point of view, the minimax rates of both of our convex regression tasks are $\Theta_{d,L,\sigma}(n^{-\frac{4}{d+4}})$ and $\Theta_{d,\Gamma,\sigma,}(C(P) \cdot n^{-\frac{4}{d+4}})$ for all $d \geq 1$ (Gao and Wellner, 2017; Bronshtein, 1976; Yang and Barron, 1999). The LSE is minimax optimal *only* in low dimension, when $d \leq 4$ (Birgé and Massart, 1993), while for $d \geq 5$ it attains a suboptimal risk of $\tilde{\Theta}_d(n^{-\frac{2}{d}})$ (Kur et al., 2020). The poor statistical performance of the LSE for $d \geq 5$ has also been verified empirically (Gardner et al., 2006; Ghosh et al., 2021). There are known minimax optimal estimators when $d \geq 5$, yet all of them are computationally inefficient. Moreover, all of them are based on some sort of discretization of the relevant function classes, i.e., they consider some ϵ-nets (see Definition 14 below). In our tasks, these algorithms require examining nets of cardinality $\Omega_d(\exp(\Theta_d(n^{\Theta(1)})))$, and are thus perforce inefficient (Rakhlin et al., 2017; Guntuboyina, 2012).

The empirically-observed poor performance of the LSE and the computational intractability of known minimax optimal estimators have motivated the study of efficient algorithms for convex regression with better statistical properties than the LSE; an incomplete list of relevant works appears above. However, previously studied algorithms are either provably minimax suboptimal or do not provide any statistical guarantees at all with respect to the minimax risk. We would however like to mention the "adaptive partitioning" estimator constructed in Hannah and Dunson (2013), which is the first provable computationally efficient estimator for convex regression which has been shown to be *consistent* in the L_∞ norm. The authors' approach is somewhat related to our proposed algorithm, but it is unknown whether their algorithm is minimax optimal.

Our main results are the existence of computationally efficient minimax optimal estimators for the task of multivariate Lipschitz convex regression and bounded convex regression under polytopal support. Specifically, we prove the following results:

Theorem 1 *Let $d \geq 5$ and $n \geq d+1$. Then, under Assumption 1, for the task of L-Lipschitz convex regression on a convex polytope $P \subset B_d$, there exists an efficient estimator, $\widehat{f_L}$, with runtime of at most $O_d(n^{O(d)})$ such that*

$$\mathcal{R}_n(\widehat{f_L}, \mathcal{F}_L(P), \mathbb{P}) \leq O_d((\sigma + L)^2 n^{-\frac{4}{d+4}} \log(n)^{h(d)} + C(P) n^{-\frac{d}{d+4}} \log(n)^{2 \cdot h(d)}), \tag{2}$$

where $h(d) \leq 3d$ and $C(P)$ is a constant that only depends on the number of flags of the polytope P (see (Reitzner et al., 2019) for the definition).

Theorem 1 gives a minimax optimal estimator in many natural cases; e.g., it applies when the polytope P is assumed to have only $C(d)$ vertices or facets, where $C(d)$ is a constant that only depends on d; this class includes, for instance, the unit cube, the simplex, and the ℓ_1 ball. Then by (McMullen, 1970), the second term in our bound is of order $\tilde{O}_d(n^{-\frac{d}{d+4}})$, which is strictly smaller than the minimax rate $\Theta_d(n^{-\frac{4}{d+4}})$.

Remark 2 *In the journal version of this manuscript, we remove the redundant term that depends on P. Specifically, we show that*

$$\mathcal{R}_n(\widehat{f}_L, \mathcal{F}_L(\Omega), \mathbb{P}) \leq O_d((\sigma + L)^2 n^{-\frac{4}{d+4}} \log(n)^{h(d)})$$

for any convex domain $\Omega \subset B_d$. The proof for an arbitrary convex body Ω involves a more involved and technical detour through stochastic geometry, and is therefore omitted in this version.

Our second main result, concerning bounded convex regression, is proved in the same way as Theorem 1, using the entropy bounds of Gao and Wellner (2017).

Theorem 3 *Let $d \geq 5$ and $n \geq d+1$. Then, under Assumption 1, for the task of Γ-bounded convex regression on the polytope $P \subset B_d$, there exists an efficient estimator, $\widehat{f}^\Gamma$, with runtime of at most $O_d(n^{O(d)})$ such that*

$$\mathcal{R}_n(\widehat{f}^\Gamma, \mathcal{F}^\Gamma(P), \mathbb{P}) \leq O_d(C_1(P)(\sigma + \Gamma)^2 n^{-\frac{4}{d+4}} \log(n)^{h(d)}),$$

where $h(d) \leq 3d$ and $C_1(P)$ is a constant that only depends on P.

As we mentioned earlier, for both of these two tasks the minimax rate is of order $n^{-\frac{4}{d+4}}$, so up to polylogarithmic factors in n, the above estimators are minimax optimal. We note that in Theorem 3, the dependence of the constants on the polytope P is unavoidable. This follows from the results of (Gao and Wellner, 2017; Han and Wellner, 2016), in which the authors showed the geometry of the support of the measure $\mathbb{P}$ affects the minimax rate of bounded convex regression. For example, in the extreme case $\Omega = B_d$, the minimax rate is of order $n^{-\frac{2}{d+1}}$, which is asymptotically larger than the error rate for polytopes; thus, if we take a sequence of polytopes which approaches B_d, the sequence of constants $C(P_n)$ will necessarily blow up.

Remark 4 *In the journal version of this paper, we show that the dependence on the domain P in Theorem 3 is an artifact of the use of the $L_2(\mathbb{P})$ metric (the L_2 entropy numbers of the bounded convex functions depends on the domain). As we show there, for any convex domain Ω, the minimax rate of bounded convex regression in the $L_1(\mathbb{P})$ metric is of order $\Theta_{d,\Gamma,\sigma}(n^{-\frac{2}{d+4}})$ (that is, unlike the L_2-entropy numbers, the L_1-entropy numbers do not depend on the domain), and there exists an efficient minimax optimal estimator attaining this rate. Thus, in the L_1 setting, the minimax rate for convex regression is* universal.

We consider our results as mainly a proof-of-concept for the existence of efficient estimators for the task of convex regression when $d \geq 5$. Due to their high polynomial runtime, in practice our estimators would probably not work well. However, as we mentioned above, the other minimax optimal estimators in the literature are computationally inefficient, and they all require consideration of some net of exponential size in n; our estimator is conceptually quite different. We hope that insights from our algorithm can be used to construct a practical estimator with the same desirable statistical properties. From a purely theoretical point of view, our estimators are the first known minimax optimal efficient estimators for non-Donsker classes for which their LSE are provably minimax suboptimal in L_2 (see Definition 16 and Remark 17 below). Prior to this work, there were efficient optimal estimators for non-Donsker classes such that their corresponding LSE (or MLE) is provably efficient and optimal (such as log-concave density estimation and isotonic regression, cf. Kur et al. (2019); Han et al. (2019); Han (2019); Pananjady and Samworth (2022)). Our work should be contrasted with these earlier works. We show that it is possible to overcome the suboptimality of the LSE with an efficient optimal algorithm in the non-Donsker regime - a result that was unknown before this work.

We prove Theorem 1 in Section 2. The proof of Theorem 3 uses the same method as that of Theorem 1, along with the main result of (Gao and Wellner, 2017, Thm 1.1). We sketch the requisite modifications to the proof in Section C. We conclude this section with the following remarks:

Remark 5

1. *We conjecture that the estimators of Theorems 1 and Theorem 3 are minimax-optimal up to constants that only depend on* d, σ*, i.e the.* $\log(n)$ *factors are unnecessary.*

2. *Our estimators' runtime is of order* $O_d(n^{O(d)})$*, which is much worse than the* $O_d(n^{O(1)})$ *runtime of the suboptimal convex LSE.*

3. *When* $d \geq 5$*, one can show that when* f^* *is a max* k*-affine function (restricted to* $P \subset B_d$*), i.e.* $f^*\mathbb{1}_P(x) = \max_{1 \leq i \leq k} a_i^\top x + b_i$*, our estimator attains a parametric rate, i.e.*

$$\mathbb{E}\int (\widehat{f}^\Gamma - f^*)^2 d\mathbb{P} \leq \tilde{O}_d\left(\frac{C(P,k)}{n}\right).$$

   When $d \leq 4$*, Han and Wellner (2016) showed that* Γ*-bounded convex LSE, that is defined in Eq. (1) with* $\mathcal{F} = \mathcal{F}^\Gamma(P)$*, attains a parametric rate as well. However, when* $d \geq 5$*, the LSE attains a non-parametric error of* $\tilde{\Theta}_d(C(P,k)n^{-4/d})$ *(Kur et al. (2020); for a more general result see Kur and Rakhlin (2021)). Therefore, our algorithm has the proper adaptive rates when* $d \geq 5$*; see Ghosh et al. (2021) for more details.*

4. *An interesting property of our estimator is that the random design setting, i.e. the fact that data points* $X_1, \ldots, X_n$ *are drawn from* $\mathbb{P}$ *rather than fixed, is essential to its success, a phenomenon not often observed when studying shape-constrained estimators. Usually these estimators also perform well on a "nice enough" fixed design set, for example when* $\Omega = [-1/2, 1/2]^d$ *and* $X_1, \ldots, X_n$ *are the regular grid points.*

2. The Proposed Estimator of Theorem 1

2.1. Notations and preliminaries

Throughout this text, $C, C_1, C_2 \in (1, \infty)$ and $c, c_1, c_2, \ldots \in (0, 1)$ are positive absolute constants that may change from line to line. Similarly, $C(d), C_1(d), C_2(d), \ldots \in (1, \infty)$ and $c(d), c_1(d), c_2(d), \ldots \in (0, 1)$ are positive constants that only depend on d that may change from line to line. We also often use expressions such as $g(n) \leq O_d(f(n))$ to mean that there exists $C_d \geq 0$ such that $g(n) \leq C_d f(n)$ for all n.

For any probability measure $\mathbb{Q}$ and $m \geq 0$, we introduce the notation $\mathbb{Q}_m$ for the random empirical measure of $Z_1, \ldots, Z_m \underset{i.i.d.}{\sim} \mathbb{Q}$, i.e. $\mathbb{Q}_m = m^{-1} \sum_{i=1}^{m} \delta_{Z_i}$. Also, given a subset $A \subset \Omega$ of positive measure, we let $\mathbb{P}_A$ denote the conditional probability measure on A. For a positive integer k, $[k]$ denotes $\{1, \ldots, k\}$.

Definition 6 *A simplex in $\mathbb{R}^d$ is the convex hull of $d + 1$ points $v_1, \ldots, v_{d+1} \in \mathbb{R}^d$ which do not all lie in any hyperplane.*

Definition 7 *A convex function $f : \Omega \to \mathbb{R}$ is defined to be k-simplicial if there exists $\triangle_1, \ldots, \triangle_k \subset \mathbb{R}^{\dim(\Omega)}$ simplices such that $\Omega = \bigcup_{i=1}^{k} \triangle_i$ and for each $1 \leq i \leq k$, we have that $f : \triangle_i \to \mathbb{R}$ is affine.*

Note that the definition is more restrictive than the usual definition of a k-max affine function (see Remark 3), since the affine pieces of a k-max affine function are not constrained to be simplices.

The following result from empirical process theory is a corollary of the peeling device (van de Geer, 2000, Ch. 5), (Bousquet, 2002) and Bronshtein's entropy bound (Bronshtein, 1976).

Lemma 8 *Let $d \geq 5$, $m \geq C^d$ and $\mathbb{Q}$ be a probability measure on $\Omega' \subset B_d$. Suppose $Z_1, \ldots, Z_m$ are drawn independently from $\mathbb{Q}$; then with probability at least $1 - C_1(d) \exp(-c_1(d)\sqrt{m})$, the following holds uniformly for all $f, g \in \mathcal{F}_L(\Omega')$:*

$$2^{-1} \int_{\Omega'} (f - g)^2 d\mathbb{Q} - CL^2 m^{-\frac{4}{d}} \leq \int (f - g)^2 d\mathbb{Q}_m \leq 2 \int_{\Omega'} (f - g)^2 d\mathbb{Q} + CL^2 m^{-\frac{4}{d}}. \quad (3)$$

Next, we introduce a statistical estimator with the high-probability guarantees we shall need, based on a recent result that is presented in (Mourtada et al., 2021, Prop. 1). Its statistical aspects are proven in the seminal works (Tsybakov, 2003a; Lugosi and Mendelson, 2019), and guarantees on its runtime are given in (Hopkins, 2018; Depersin and Lecué, 2019; Hopkins et al., 2020).

Lemma 9 *Let $m \geq d + 1$, $d \geq 1$, $\delta \in (0, 1)$ and $\mathbb{Q}$ be a probability measure that is supported on $\Omega' \subset B_d$ with* a known *covariance matrix Σ. Consider the regression model $W = f^*(Z) + \xi$, where $\|f^*\|_\infty \leq L$, and let $Z_1, \ldots, Z_m \underset{i.i.d.}{\sim} \mathbb{Q}$. Then, there exists an estimator $\hat{f}_{R,\delta}$ that has an* *input of* $(\Sigma, \{(Z_i, W_i)\}_{i=1}^m)$ *and runtime of* $\tilde{O}_d(m)$ *and outputs an* L*-Lipschitz affine function that satisfies with probability at least* $1-\delta$

$$\int (\hat{f}_{R,\delta}(x) - w^*(x))^2 d\mathbb{Q}(x) \leq \frac{C(\sigma+L)^2(d+\log(1/\delta))}{m},$$

where $w^* = \operatorname{argmin}_{w \text{ affine}} \int (w - f^*)^2 d\mathbb{Q}$.

2.2. Proof of Theorem 1

The first ingredient in our estimator is our new approximation theorem for convex functions:

Theorem 10 *Let* $\Omega \subset B_d$ *be a convex polytope,* $f \in \mathcal{F}_L(\Omega)$*, and* k *an integer greater than* $(Cd)^{d/2}$*, for some large enough* $C \geq 0$*. Then, there exists a convex set* $\Omega_k \subset \Omega$ *and a* k*-simplicial convex function* $f_k : \Omega_k \to \mathbb{R}$ *such that*

$$\mathbb{P}(\Omega \setminus \Omega_k) \leq C(\Omega) k^{-\frac{d+2}{d}} \log(k)^{d-1}. \tag{4}$$

and

$$\int_{\Omega_k} (f_k - f)^2 d\mathbb{P} \leq L^2 \cdot O_d(k^{-\frac{4}{d}} + C(\Omega) k^{-1} \log(k)^{d-1}). \tag{5}$$

Note that both Ω_k, as well as f_k, depend on f^*. The bound of Eq. (5) is in fact tight, up to a constant that only depends on d, cf. (Ludwig et al., 2006). Also note that f_k is not necessarily an L-Lipschitz function, i.e., it may be an "improper" approximation to f. We remark that the constant $C(\Omega)$ depends on the flag number of the polytope Ω; for more details see (Reitzner et al., 2019). As we mentioned earlier, we assume that the number of vertices or facets of Ω is bounded by C_d, so the definition of the flag number and the upper bound theorem of McMullen (McMullen, 1970) implies that we can assume $C(\Omega) \leq C_1(d)$.

For simplicity, we shall assume that $L = \sigma = 1$. Also, since our function is 1-Lipschitz and $\Omega \subset B_d$, we may also assume that $\|f^*\|_\infty \leq 1$. Finally, we may and do assume that $\mathbb{P} = U(\Omega)$, since we can always simulate $\Theta_d(n)$ uniform samples using the method of rejection sampling given samples from any distribution $\mathbb{P}$ which satisfies Assumption 1 (cf. (Devroye, 1986)).

Fix $n \geq (Cd)^{d/2}$, $f^* \in \mathcal{F}_1(\Omega)$, and set $k(n) := n^{\frac{d}{d+4}}$. Let $f_{k(n)} : \Omega_{k(n)} \to \mathbb{R}$ be the convex function whose existence is guaranteed by Theorem 10 for $f = f^*$. We have

$$\int (f^* - f_{k(n)})^2 d\mathbb{P} \leq O_d(n^{-\frac{4}{d+4}}), \tag{6}$$

and there exist $\triangle_1, \ldots, \triangle_{k(n)} \subset \Omega$ simplices such that $f_{k(n)}\big|_{\triangle_i}$ is affine on each i.

If we were given the decomposition of Ω into pieces on which f^* is near-affine, it would be relatively simple to estimate f^*, as we show in Appendix D below. We recommend reading it, to get some intuition for our approach, before attempting the description and correctness proof for our "full" estimator below.

To overcome the fact that we do not know the simplices $\triangle_i$ on which f_k is affine, we need another lemma, which says that if we randomly sample a set of n points $\{X_{n+1}, \ldots, X_{2n}\}$ from Ω, there exists a collection of at most $\tilde{O}_d(k(n))$ simplices covering "most" of Ω_k, such that the vertices of each simplex belong to $\{X_{n+1}, \ldots, X_{2n}\}$ and f_k is affine on each simplex in the collection:

Lemma 11 *Let $n \geq d+1$, and $\triangle_1, \ldots, \triangle_{k(n)}$ that are defined above, and let $X_{n+1}, \ldots, X_{2n} \underset{i.i.d.}{\sim} \mathbb{P}$. Then, with probability at least $1 - n^{-1}$, there exist $k(n)$ disjoint sets $\mathcal{S}_X^1, \ldots \mathcal{S}_X^{k(n)}$ of simplices with disjoint interiors such that*

1. *The vertices of each simplex in $\bigcup_{i=1}^{k(n)} \mathcal{S}_X^i$ lie in $\{X_{n+1}, \ldots, X_{2n}\}$. Moreover, for each $1 \leq i \leq k(n)$, we have that $|\mathcal{S}_X^i| \leq O_d(\log(n\mathbb{P}(\triangle_i))^{d-1})$.*
2. *For each $1 \leq i \leq k(n)$, we have that $\bigcup \mathcal{S}_X^i \subset \triangle_i$, and*

$$\mathbb{P}(\bigcup \mathcal{S}_X^i) \geq \mathbb{P}(\triangle_i) - \min\left\{O_d\left(\frac{\log(n)\log(n\mathbb{P}(\triangle_i))^{d-1}}{n}\right), \mathbb{P}(\triangle_i)\right\}.$$

The proof of this lemma appears in subsection B.2. Essentially, this lemma states that we can triangulate "most" of each simplex $\triangle_i$ with "few" simplices whose vertices lie among the points $X_{n+1}, \ldots, X_{2n}$ which fall in $\triangle_i$, so long as $\triangle_i$ is large enough. From now on, we condition on the high-probability event of Lemma 11.

Note that if we were given the set of simplices $\mathcal{S}_X := \bigcup_{i=1}^{k(n)} \mathcal{S}_X^i$, we could use the same strategy as in Appendix D to obtain a minimax optimal estimator for this task as well. Unfortunately, we do not know how to identify the simplices of $\mathcal{S}_X$, but we do know that they belong to the collection of all simplices with vertices in $\{X_{n+1}, \ldots, X_{2n}\}$,

$$\mathcal{S} := \{\mathrm{conv}\{X_{n+i} : i \in S\} : S \subset [n], |S| = d+1\}. \tag{7}$$

Note that $|\mathcal{S}| = O_d(n^{d+1})$, which is polynomial in n.

Instead of trying to identify the simplices $\mathcal{S}_X \subset \mathcal{S}$ on which f is close to being linear, our algorithm finds a function $\hat{f}$ which, on *every* simplex $\triangle \in \mathcal{S}$, is "not much farther" from the best linear approximation to f on $\triangle$ then f is. Since f itself is close to its best linear approximation on each simplex in $\mathcal{S}_X$, $\hat{f}$ will be close to f on $\bigcup \mathcal{S}_X$, which is most of Ω.

We restate this a bit more precisely: if $\hat{f} : \Omega \to \mathbb{R}$ is a convex Lipschitz function such that

$$\forall \triangle \in \mathcal{S} : \int (\hat{f} - w_\triangle^*)^2 d\mathbb{P}_\triangle \leq \tilde{O}_d\left(\int (f^* - w_\triangle^*)^2 d\mathbb{P}_\triangle + (\mathbb{P}(\triangle)n)^{-1}\right) \tag{8}$$

where $w_\triangle^* = \inf_{w \text{ affine}} \int (f^* - w)^2 d\mathbb{P}_\triangle$, then $\hat{f}$ satisfies $\int_\Omega (\hat{f} - f^*)^2 d\mathbb{P} \leq \tilde{O}_d(n^{-\frac{4}{d+4}})$, and the RHS is the minimax optimal rate. (The idea of the proof is to use the fact that f^* is close to $w_\triangle^*$ for each $\triangle \in \mathcal{S}_X$, along with the triangle inequality, and then sum over all simplices in $\mathcal{S}_X$; the full justification is given at (17)-(20) below.) In the remainder of this section, we will describe an efficient algorithm which constructs a function $\hat{f}$ which comes "close enough" to satisfying (8) that it manages to attain the minimax optimal rate.

We now begin the description of our algorithm. In the notation of (8), for each simplex $\triangle \in \mathcal{S}$, we estimate $w^*_\triangle$ by applying Lemma 9, with the data points of $\mathcal{D}^1 = \{(X_i, Y_i)\}_{i=1}^{n/2}$ that lie in $\triangle$ as input; denote the regressor we obtain by $\hat{w}_\triangle$.

Next, we shall need to estimate (with high probability) the squared error of the regressor on each simplex in $\mathcal{S}$, i.e. $\ell^2_\triangle := \|f^* - \hat{w}_\triangle\|^2_{L^2(\triangle)}$, up to a polylogarithmic multiplicative factor, using the data points in $\mathcal{D}^2 = \{(X_i, Y_i)\}_{i=n/2+1}^{n}$ that lie in $\triangle$. Letting $w^*_\triangle$ to be defined as above, we have

$$\ell^2_\triangle = \|w^*_\triangle - \hat{w}_\triangle\|^2_{L^2(\triangle)} + \|f^* - w^*_\triangle\|^2_{L^2(\triangle)};$$

the first term is called the (squared) estimation error and the second is called the (squared) approximation error. By Lemma 9, with probability at least $1 - n^{-2d}$ the estimation error will be at most $Cd\log(n)/(\mathbb{P}(\triangle)n)$, which is no more than a $O(d\log(n))$ factor times the expected estimation error. However, f^* may not be affine on $\triangle$, and the squared approximation error may be significantly larger than the squared estimation error. When this occurs, the estimation of $\ell^2_\triangle$ by noisy samples is challenging, even in the (unrealistic) setting of sub-Gaussian noise with known variance σ^2. Indeed, it would be natural to estimate the approximation error by the (centered) empirical mean of the squared loss, namely

$$\frac{1}{\mathbb{P}(\triangle)n} \sum_{(X,Y)\in\mathcal{D}_2, X\in\triangle} (Y - \hat{w}_\triangle(X))^2 - \sigma^2.$$

However, the additive deviation of this estimate is of order $\Omega_d(n_\triangle^{-\frac{1}{2}})$, where $n_\triangle \approx \mathbb{P}(\triangle)n$ is the number of data points falling in $\triangle$, and therefore when $\ell^2_\triangle$ is in the range $[O_d(n_\triangle^{-1}), \Omega_d(n_\triangle^{-\frac{1}{2}})]$ we will not be able to estimate $\ell^2_\triangle$ even *up to a multiplicative constant*, which is what our algorithm requires in order to succeed.

To overcome this problem necessitates constructing a new procedure to estimate $\ell^2_\triangle$, that is the $L_2^2(\mathbb{P}_\triangle)$-norm of the *convex* function $f^* - \hat{w}_\triangle$, up to a multiplicative constant, with an *additive* deviation of $\tilde{O}((\mathbb{P}(\triangle)n)^{-1})$. We proceed in several steps. First, we develop a new estimator for the L_1 norm of any convex function g:

Lemma 12 *Let K be any convex body in $\mathbb{R}^d$. Let $\delta \in (0,1)$ and $f : K \to \mathbb{R}$ be a convex function, and suppose that m i.i.d. samples are drawn from the regression model $Y = f(Z) + \xi$, where $Z \sim U(K)$. There exists an estimator $\bar{f}^\delta_m$ taking these samples as input which satisfies, with probability at least $1 - 3\max\{\delta, e^{-cm}\}$,*

$$\|f\|_{L_1(U(K))} \le \bar{f}^\delta_m \le C(d,K)\|f\|_{L_1(U(K))} + C_1(d,K)\sqrt{\frac{\log(2/\delta)}{m}} \cdot (\|f\|_{L_2(U(K))} + \sigma),$$

where $C(d,K), C_1(d,K)$ are constants that only depend on the convex set K and the dimension d.

The estimator of Lemma 12 is invariant with respect to affine transformations of the domain. Thus, the constants $C(d, K), C_1(d, K)$ are the same for all K in the class of affine images of a fixed convex body in $\mathbb{R}^d$, such as the class of simplices.

The estimator of the last lemma gives an optimal error rate, with respect to the number of samples m, for the $L_1(U(\triangle))$-norm of any convex function g (with no restriction on its uniform norm or Lipschitz constant). In the next step, we aim to find the $L_2(U(\triangle))$ norm of a convex function g. In order to obtain a similar result as in the previous lemma, it will be essential for both the statistical guarantees and the computational aspects of the proposed estimator to assume that the domain of g is a simplex and that $\|g\|_\infty \leq L$. Under these conditions, we construct the following estimator:

Lemma 13 *Let $\delta \in (0, 1)$, and $\triangle \subset \Omega$ be a simplex. Consider the regression model $Y = g(Z)+\xi$, where $Z \sim \mathbb{P}_\triangle$ and $\|g\|_\infty \leq L$. Furthermore, assume that $\int_\triangle g^2 d\mathbb{P} \geq CdL^2 \log(n)^2/n$. Then, there exists an estimator $\hat{f}_{E,\delta}$ that runs in time $O_d((\mathbb{P}\triangle n)^{O(1)})$ and with probability at least $1 - \log(n) \max\{\delta, n^{-4d}\}$ satisfies*

$$\|g\|^2_{L_2(\triangle)} \leq \hat{f}_{E,\delta} \leq C(d) \log(n)^{2d-1} \left(\|g\|^2_{L_2(\triangle)} + (L + \sigma)^2 \frac{\log(2/\delta)}{\mathbb{P}(\triangle)n} \right),$$

where $C(d)$ is a constant that only depends on d.

Our estimator outputs $\|g\|^2_{L_2(\triangle)}$ up to a multiplicative factor of $\tilde{\Theta}_d(1)$. We apply Lemma 13 with $g = f^*|_\triangle - \hat{w}_\triangle$, and $\delta = n^{-2d}$, and the data points of $\mathcal{D}_2$ that fall in $\triangle$, and we denote the output of the estimator by $\hat{\ell}^2_\triangle$. Note the definition of $\hat{\ell}^2_\triangle$ implies we must know some upper bound on L and σ (up to multiplicative constants that only depend on d). Both can be found using standard methods.

Given our regressors $\hat{w}_\triangle$ and squared error estimates $\hat{\ell}^2_\triangle$, we proceed to solve the quadratic program which encodes the conditions $\|\tilde{f} - \hat{w}_\triangle\|^2_{L^2(\triangle)} \leq \hat{\ell}^2_\triangle$ for all simplices with large enough volume. (We rely on the fact that the L^2-norm on each simplex can be approximated by the empirical L^2-norm, again using Lemma 8.) This program is feasible, since f^* itself is a solution. $\tilde{f}$ is close to f^* on every simplex in our collection and in particular on the simplices restricted to which f^* is near-affine (which we don't know how to identify), which allows us to conclude that $\int_{\tilde{\Omega}_{k(n)}} (\tilde{f}-f^*)^2 d\mathbb{P} \leq \tilde{O}_d(n^{-\frac{4}{d+4}})$ with high probability, where $\tilde{\Omega}_{k(n)}$ is the union of the simplices in Lemma 11.

So we have constructed a function $\tilde{f}$ which closely approximates f^* on $\tilde{\Omega}_{k(n)}$. $\tilde{\Omega}_{k(n)}$ is not known to us, but as we shall see, $\Omega \backslash \tilde{\Omega}_{k(n)}$ has asymptotically negligible volume, so the function $\min\{\tilde{f}, 1\}$ turns out to be a minimax optimal improper estimator (up to logarithmic factors) of f^* on all of Ω. In order to transform this improper estimator to a proper estimator, i.e., one whose output is a convex 1-Lipschitz function, we use a standard procedure (denoted by MP), as described in Appendix E below. This concludes the sketch of our algorithm.

Pseudocode for the algorithm is given in Algorithm 1 below. In its formulation, note that the procedure $\hat{f}_{R,\delta(n)}$ is described in Lemma 9, $\hat{f}_{E,\delta(n)}$ is described in Lemma 13, and MP is described in Appendix E.

Algorithm 1 A Minimax Optimal For L-Lipschitz Multivariate Convex Regression

```
Require: $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$
Ensure: A random $\widehat{f}_L \in \mathcal{F}_L(\Omega)$ s.t. w.h.p. $\|\widehat{f}_L - f^*\|_{\mathbb{P}}^2 \le \tilde{O}_d((L+\sigma)^2 n^{-\frac{4}{d+4}})$.
  Draw $X_{n+1}, \dots, X_{2n} \underset{i.i.d.}{\sim} \mathbb{P}$
  $\mathcal{S} \leftarrow \{\mathrm{conv}\{X_{n+i} : i \in S\} : S \subset [n], |S| = d+1\}$
  Part I:
  $\mathcal{D}^1 \leftarrow \{(X_i, Y_i)\}_{i=1}^{n/2}$
  $\mathcal{D}^2 \leftarrow \{(X_i, Y_i)\}_{i=n/2+1}^{n}$
  $\delta(n) \leftarrow n^{-(d+2)}$
  for $\triangle_1, \dots, \triangle_i, \dots \in \mathcal{S}$ do
    $\hat{w}_i \leftarrow \hat{f}_{R,\delta(n)}(\{(X,Y) \in \mathcal{D}^1 : X \in \triangle_i\})$.
    $\hat{\ell}_i \leftarrow \min(4, \hat{f}_{E,\delta(n)}(\{(X, Y - \hat{w}_i(X)) : (X,Y) \in \mathcal{D}^2, X \in \triangle_i\}))$
      end
  Part II:
  for $i \in 1, \dots |\mathcal{S}|$ do
    Draw $Z_{i,1}, \dots, Z_{i,n} \sim \mathbb{P}_{\triangle_i}$
    Define an inequality constraint $I_j := \frac{1}{n}\sum_{j=1}^n (f(Z_{i,j}) - \hat{w}_i^\top (Z_{i,j}, 1))^2 \le \hat{\ell}_i^2 + CL^2\sqrt{\frac{d\log(n)}{n}}$.
      end
  Construct $\tilde{f} \in \mathcal{F}_L(\Omega)$ satisfying the constraints $I_1, I_2 \dots, I_{|S|}$ (cf. Eqs. (13)-(15))
  return $MP(\min\{\tilde{f}, L\})$,
```

We now turn to the proof that Algorithm 1 succeeds with high probability. In the analysis, we assume for simplicity that $L = \sigma = 1$. Let $\mathcal{S}$ be as defined in Algorithm 1, and let $\mathcal{S}^T := \{\triangle : \triangle \in \mathcal{S}, \int_\triangle g^2 d\mathbb{P} \ge Cd\log(n)^2/n\}$, for some sufficiently large C. In particular, we have $\mathbb{P}(S) \ge C_1(C)d\log(n)/n$ for all $S \in \mathcal{S}^T$. We first note that our samples may be assumed to be close to uniformly distributed on the simplices in $\mathcal{S}^T$. Indeed, by standard concentration bounds, we have

$$\forall \triangle \in \mathcal{S}^T, j \in \{1,2,3\} : \quad \frac{1}{2} \le \frac{\mathbb{P}_n^{(j)}(\triangle)}{\mathbb{P}(\triangle)} \le 2, \tag{9}$$

with probability $1 - 3n^{-3d}$, where $\mathbb{P}_n^{(1)} = \frac{2}{n}\sum_{i=1}^{n/2}\delta_{X_i}$, $\mathbb{P}_n^{(2)} = \frac{2}{n}\sum_{i=n/2+1}^{n}\delta_{X_i}$ and $\mathbb{P}_n^{(3)} = \frac{1}{n}\sum_{i=n+1}^{2n}\delta_{X_i}$ (see Lemma 20 in sub-Section B.2). From now on, we condition on the intersection of the events of (9) and Lemma 11.

The first step in the algorithm is to apply the estimator of Lemma 9 for each $\triangle_i \in \mathcal{S}^T$ with $\mathbb{Q} := \mathbb{P}_{\triangle_i}$, and $\delta = n^{-(d+2)}$, using those points among of $\mathcal{D}^1$ that fall in $\triangle_i$. (By the preceding paragraph, under our conditioning, we may assume $\mathbb{P}(\triangle_i)n$ of the points in $X_1, \dots, X_{\frac{n}{2}}$ fall in each $\triangle_i$, up to absolute constants. We will silently use the same argument several more times below.) By the lemma and a union bound, we know that the following event has probability at least $1 - n^{-1}$:

$$\forall 1 \le i \le |\mathcal{S}^T| \quad \int_{\triangle_i} (\hat{w}_i(x) - f^*(x))^2 d\mathbb{P}_{\triangle_i} \le \frac{2Cd\log(n)}{\mathbb{P}(\triangle_i)n} + \int_{\triangle_i} (w_i^*(x) - f^*(x))^2 d\mathbb{P}_{\triangle_i}, \tag{10}$$

where $w_i^* = \operatorname{argmin}_{w \text{ affine}} \int_{\triangle_i} (w(x) - f^*)^2 d\mathbb{P}_{\triangle_i}$. We condition also on the event of (10). Next, we apply Lemma 13 (with $\delta = n^{-(d+2)}$) on each $\triangle_i$, with $g = f^* - \hat{w}_i$, and using those points among of $\mathcal{D}^2$ that fall in $\triangle_i$, and obtain that

$$\forall 1 \leq i \leq |\mathcal{S}| : \quad \int_{\triangle_i} (\hat{w}_i - f^*)^2 d\mathbb{P}_{\triangle_i} \leq \hat{\ell}_i^2, \tag{11}$$

with $\hat{\ell}_i^2$ as defined in Algorithm 1. Note that for $\triangle \in \mathcal{S} \setminus \mathcal{S}^T$, taking $\hat{w}_i = 0$ suffices, since f^* is bounded by 1, the loss is bounded by 4. Finally, we further condition on the event of the last equation.

We proceed to explain and analyze Part II of Algorithm 1. We first claim that conditioned on (11), the function f^* satisfies the constraints $I_1, I_2, \ldots$ defined in the algorithm with probability at least $1 - n^{-1}$. Indeed, for each $1 \leq i \leq |\mathcal{S}|$, $\|(f^* - \hat{w}_i)^2\|_{L^\infty(\triangle_i)} \leq 4$, so by Hoeffding's inequality and (11) we know that with probability at least $1 - n^{-(d+2)}$, we have that

$$\begin{aligned} \frac{1}{n}\sum_{j=1}^{n} (f^*(Z_{i,j}) - \hat{w}_i(Z_{i,j}))^2 &\leq \int_{\triangle_i} (f^* - \hat{w}_i)^2 d\mathbb{P}_{\triangle_i} + \sqrt{\frac{Cd\log(n)}{n}} \\ &\leq \hat{\ell}_i^2 + \sqrt{\frac{Cd\log(n)}{n}}. \end{aligned} \tag{12}$$

Taking a union over i, we know that (12) holds for all i with probability at least $1 - n^{-1}$.

We also note (for later use) that applying Lemma 8 to the measures $\mathbb{P}_{\triangle_i}$ and using a union bound, it holds with probability at least $1 - Cn^d e^{-c\sqrt{n}}$ that for all i, the empirical measure $\mathbb{P}_{\triangle_i, n} = \frac{1}{n}\sum_{j=1}^{n} \delta_{Z_{i,j}}$ on $\triangle_i$ approximates $\mathbb{P}_{\triangle_i}$ in the sense of (3). We condition on the intersection of these two events as well.

We now explain how to algorithmically construct $\tilde{f} \in \mathcal{F}_L(\Omega)$ satisfying all the constraints I_j. The idea is to mimic the computation of the convex LSE (Seijo and Sen, 2011), by considering the values of the unknown function $y_{i,j} = \tilde{f}(Z_{i,j})$ and the subgradients $\xi_{i,j} \in \partial f(Z_{i,j})$ at each $Z_{i,j}$ as variables. More precisely, we search for $y_{i,j} \in \mathbb{R}$ and $\xi_{i,j} \in \mathbb{R}^d$ satisfying the following set of constraints (here $L = 1$):

$$\forall i \leq |\mathcal{S}| : \quad \frac{1}{n}\sum_{j=1}^{n} (y_{i,j} - \hat{w}_i(Z_{i,j}))^2 \leq \hat{\ell}_i^2 + \sqrt{\frac{Cd\log(n)}{n}} \tag{13}$$

$$\forall (i,j) \in [|\mathcal{S}|] \times [n] : \quad \|\xi_{i,j}\|^2 \leq L^2 \tag{14}$$

$$\forall (i_1, j_1), (i_2, j_2) \in [|\mathcal{S}|] \times [n] : \quad y_{i_2,j_2} \geq \langle \xi_{i_1,j_1}, Z_{i_2,j_2} - Z_{i_1,j_1} + y_{i_1,j_1} \rangle. \tag{15}$$

For any feasible solution $(y_{i,j}, \xi_{i,j})_{i,j}$ of (13)-(15), define the affine functions $a_{i,j}(x) = y_{i,j} + \langle \xi_{i,j}, x - Z_{i,j} \rangle$. We claim that the function $\tilde{f} = \max_{i,j} a_{i,j}$ is a 1-Lipschitz convex function which satisfies the constraints I_j. Indeed, (15) guarantees that $\tilde{f}(Z_{i,j}) = y_{i,j}$ for each i, so the I_j are satisfied due to (13); moreover, the $a_{i,j}$ are convex and 1-Lipschitz (the latter because of (14)), so $\tilde{f}$ is convex as a maximum of convex functions and 1-Lipschitz as a maximum of 1-Lipschitz functions.

Conditioned on (12) there exists a feasible solution to the problem (13)-(15), namely that obtained by taking $y_{i,j} = f^*(Z_{i,j})$, $\xi_{i,j} \in \partial f^*(Z_{i,j})$ (where $\partial f(x)$ denotes the subgradient set of a convex function f at the point x). Moreover, the constraints in (13)-(15) are either linear or convex and quadratic in $f(Z_{i,j}), u_{i,j}$, and hence the problem can be solved efficiently. For instance, it can be expressed as a second-order cone program (SOCP) with $O_d(n^{2d+2})$ variables and constraints, which can be solved in time $O_d(n^{O(d)})$ (see, e.g., Ben-Tal and Nemirovski (2001)).

Next, recall that under our conditions on the $Z_{i,j}$, we have for each i that

$$\int_{\triangle_i} (\tilde{f}(x) - f^*(x))^2 d\mathbb{P}_{\triangle_i} \leq \frac{1}{n}\sum_{j=1}^{n}(\tilde{f}(Z_{i,j}) - f^*(Z_{i,j}))^2 + Cn^{-\frac{4}{d}}, \tag{16}$$

since both f^* and $\tilde{f}$ lie in $\mathcal{F}_1(\Omega)$. Recall also that under our conditioning, for each i, the constraint

$$\frac{1}{n}\sum_{j=1}^{n}(y_{i,j} - \hat{w}_i(Z_{i,j}))^2 \leq \hat{\ell}_i^2 + \sqrt{\frac{Cd\log(n)}{n}}$$

holds whether we take $y_{i,j} = \tilde{f}(Z_{i,j})$ or $y_{i,j} = f^*(Z_{i,j})$. Using this bound along with the inequality $(f^* - \tilde{f})^2 \leq 2(\tilde{f} - \hat{w}_i)^2 + 2(\hat{w}_i - f^*)^2$ in (16), we obtain

$$\int_{\triangle_i} (\tilde{f}(x) - f^*(x))^2 d\mathbb{P}_{\triangle_i} \leq 2\hat{\ell}_i^2 + \sqrt{\frac{Cd\log(n)}{n}} + Cn^{-\frac{4}{d}} \leq 2\hat{\ell}_i^2 + C'n^{-\frac{4}{d+4}}, \tag{17}$$

where we used our assumption of $d \geq 5$. Now, recalling that ℓ_i^2 denotes the LSE error $\|f^* - \hat{w}_i\|^2_{L^2(\triangle_i)}$, which is bounded by (10), we have

$$\hat{\ell}_i^2 \leq C_d \log(n)^{3d}\ell_i^2 \leq C_d\log(n)^{2d-1}\left(\|f^* - w_i^*\|^2_{L^2(\triangle_i)} + \frac{C(d)\log n}{\mathbb{P}(\triangle_i)n}\right).$$

Substituting in (17), we obtain for any $\triangle_i$ that

$$\int_{\triangle_i} (\tilde{f}(x) - f^*(x))^2 d\mathbb{P}_{\triangle_i} \leq C(d)\left(\log(n)^{2d-1}\|f^* - w_i^*\|^2_{L^2(\triangle_i)} + \frac{\log(n)^{2d}}{\mathbb{P}(\triangle_i)n} + n^{-\frac{4}{d+4}}\right). \tag{18}$$

Now recall that $f_{k(n)}$ is our $k(n)$-simplicial approximation to f, and that $f_{k(n)}\big|_{S_{i,j}}$ is affine for each i and $S_{i,j} \in \mathcal{S}_X^i$, where the sets $\mathcal{S}_X^i$ are defined in Lemma 11. Define $\tilde{\Omega}_{k(n)} := \bigcup\bigcup_{i=1}^{k(n)} \mathcal{S}_X^i$. Recall that by definition, $\|f^* - w_m^*\|^2 = \inf_{w \text{ affine}} \|f^* - w\|^2_{L^2(S_m)}$ for any $S_m \in \mathcal{S}$, in particular for those S_m which belong to one of the $\mathcal{S}_X^i$. Hence, multiplying (18) by $\mathbb{P}(\triangle_i)$ and summing over all the $\triangle_i$ belonging to any of the $\mathcal{S}_X^i$, we obtain

$$\begin{aligned}
\int_{\tilde{\Omega}_{k(n)}} (\tilde{f} - f^*)^2 d\mathbb{P} &\leq C_1(d) \sum_{\tilde{\triangle} \in \bigcup_{i=1}^{k(n)} \mathcal{S}_X^i} \left(\log(n)^{2d-1} \inf_{w \text{ affine}} \int_{\tilde{\triangle}} (f^* - w)^2 d\mathbb{P} + \frac{\log(n)^{2d}}{n} + \mathbb{P}(\tilde{\triangle})n^{-\frac{4}{d}}\right) \\
&\leq C_d\left(\log(n)^{2d-1}\int_{\tilde{\Omega}_{k(n)}} (f_{k(n)} - f^*)^2 d\mathbb{P} + \log(n)^{3d} n^{-\frac{4}{d+4}}\right) \\
&\leq O_d(n^{-\frac{4}{d+4}}\log(n)^{3d}),
\end{aligned} \tag{19}$$

where we used the fact that the cardinality of $\bigcup_{i=1}^{k(n)} \mathcal{S}_X^i$ is bounded by $O_d(n^{\frac{d}{d+4}} \log(n)^d)$, the disjointness of all the simplices comprising $\tilde{\Omega}_{k(n)}$, and finally the definition of $f_{k(n)}$. Next, it is not hard to show that $\|\tilde{f}\|_\infty \leq C$ (simply because $\tilde{f}$ is 1-Lipschitz, Ω is contained in the unit ball, and $\int_{\Omega_{k(n)}} (\tilde{f} - f^*)^2 \leq 4$). Thus, we obtain that

$$\int_{\Omega_{k(n)} \setminus \tilde{\Omega}_{k(n)}} (\tilde{f} - f^*)^2 d\mathbb{P} \leq \|\tilde{f}\|_\infty \mathbb{P}(\Omega_{k(n)} \setminus \tilde{\Omega}_{k(n)}) \leq C_1 \sum_{i=1}^{k(n)} \mathbb{P}(\triangle_i \setminus \bigcup \mathcal{S}_X^i)$$
$$\leq O_d(n^{-\frac{4}{d+4}} \log(n)^d).$$

where we used part (2) of Lemma 11. Combining the last two equations, we obtain that

$$\int_{\Omega_{k(n)}} (\tilde{f} - f^*)^2 d\mathbb{P} \leq O_d(n^{-\frac{4}{d+4}} \log(n)^{3d}). \tag{20}$$

Finally, since $\mathbb{P}(\Omega \setminus \Omega_{k(n)}) \leq C(d) n^{-\frac{d+2}{d+4}} \log(n)^{d-1}$, we can estimate f^* simply by 1 on $\Omega \backslash \Omega_{k(n)}$ and the error of doing so will be asymptotically negligible ($n^{-\frac{d+2}{d+4}} \ll n^{-\frac{4}{d+4}}$), so $\min\{\tilde{f}, 1\}$ is a minimax optimal estimator on all of Ω. (Recall that this is an improper estimator, and we can apply the procedure MP described in Appendix E to obtain a proper estimator.) It is not hard to see that the runtime of the above algorithm is $O_d(n^{O(d)})$. The proof of Theorem 1 is complete.

Acknowledgments

The first author is supported by NSF through award DMS-2031883 and from the Simons Foundation through Award 814639 for the Collaboration on the Theoretical Foundations of Deep Learning, as well as support from NSF through grant DMS-1953181. The second author was supported by the ERC under the European Union's Horizon 2020 research and innovation programme (grant agreement No 770127). The first author would like to acknowledge Prof. Alexander Rakhlin, Prof. Mattias Reitzner and Dr. Nikita Zhivotevsky for the useful discussions. The authors deeply acknowledge Prof. Shiri Artstein-Avidan for introducing the two authors to each other.

References

Shiri Artstein-Avidan, Apostolos Giannopoulos, and Vitali D Milman. *Asymptotic geometric analysis, Part I*, volume 202 of *Mathematical Surveys and Monographs; v. 202*. American Mathematical Society, 2015.

Gábor Balázs. Convex regression: theory, practice, and applications. 2016.

Gábor Balázs. Adaptively partitioning max-affine estimators for convex regression. In *International Conference on Artificial Intelligence and Statistics*, pages 860–874. PMLR, 2022.

Gábor Balázs, András György, and Csaba Szepesvári. Near-optimal max-affine estimators for convex regression. In *Artificial Intelligence and Statistics*, pages 56–64. PMLR, 2015.

Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.

Dimitris Bertsimas and Nishanth Mundru. Sparse convex regression. *INFORMS Journal on Computing*, 33(1):262–279, 2021.

Lucien Birgé and Pascal Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1-2):113–150, 1993.

Jose Blanchet, Peter W Glynn, Jun Yan, and Zhengqing Zhou. Multivariate distributionally robust convex regression under absolute error loss. *Advances in Neural Information Processing Systems*, 32:11817–11826, 2019.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Olivier Bousquet. Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms. 2002.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

EM Bronshtein. ε-entropy of convex sets and functions. *Siberian Mathematical Journal*, 17(3): 393–398, 1976.

Victor-Emmanuel Brunel. Adaptive estimation of convex polytopes and convex sets from noisy data. *Electronic Journal of Statistics*, 7:1301–1327, 2013.

Victor-Emmanuel Brunel. Adaptive estimation of convex and polytopal density support. *Probability Theory and Related Fields*, 164(1-2):1–16, 2016.

I. Bárány and D. G. Larman. Convex bodies, economic cap coverings, random polytopes. *Mathematika*, 35(2):274–291, 1988. doi: 10.1112/S0025579300015266.

Imre Bárány. Intrinsic volumes and f-vectors of random polytopes. *Mathematische Annalen*, 285 (4):671–699, 1989.

Timothy Carpenter, Ilias Diakonikolas, Anastasios Sidiropoulos, and Alistair Stewart. Near-optimal sample complexity bounds for maximum likelihood estimation of multivariate log-concave densities. In *Conference On Learning Theory*, pages 1234–1262, 2018.

Sourav Chatterjee. A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42(6):2340–2381, 2014.

Wenyu Chen and Rahul Mazumder. Multivariate convex regression at scale. *arXiv preprint arXiv:2005.11588*, 2020.

Wenyu Chen, Rahul Mazumder, and Richard J Samworth. A new computational framework for log-concave density estimation. *arXiv preprint arXiv:2105.11387*, 2021.

Madeleine Cule, Richard Samworth, and Michael Stewart. Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(5):545–607, 2010.

Jules Depersin and Guillaume Lecué. Robust subgaussian estimation of a mean vector in nearly linear time. *arXiv preprint arXiv:1906.03058*, 2019.

Luc Devroye. *Non-uniform random variate generation*. Springer-Verlag, 1986.

Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I Oliveira. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Efficient robust proper learning of log-concave distributions. *arXiv preprint arXiv:1606.03077*, 2016.

Ilias Diakonikolas, Anastasios Sidiropoulos, and Alistair Stewart. A polynomial time algorithm for maximum likelihood estimation of multivariate log-concave densities. *arXiv preprint arXiv:1812.05524*, 2018.

Rex A Dwyer. On the convex hull of random points in a polytope. *Journal of Applied Probability*, 25(4):688–699, 1988.

Fuchang Gao and Jon A Wellner. Entropy of convex functions on $\mathbb{R}^d$. *Constructive approximation*, 46(3):565–592, 2017.

Richard J Gardner. *Geometric tomography*, volume 6. Cambridge University Press Cambridge, 1995.

Richard J Gardner, Markus Kiderlen, and Peyman Milanfar. Convergence of algorithms for reconstructing convex bodies and directional measures. *The Annals of Statistics*, 34(3):1331–1374, 2006.

Avishek Ghosh, Ashwin Pananjady, Adityanand Guntuboyina, and Kannan Ramchandran. Max-affine regression: Parameter estimation for gaussian designs. *IEEE Transactions on Information Theory*, 2021.

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

Adityanand Guntuboyina and Bodhisattva Sen. Covering numbers for convex functions. *IEEE Transactions on Information Theory*, 59(4):1957–1965, 2013.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

Qiyang Han. Global empirical risk minimizers with “shape constraints” are rate optimal in general dimensions. *arXiv preprint arXiv:1905.12823*, 2019.

Qiyang Han and Jon A Wellner. Multivariate convex regression: global risk bounds and adaptation. *arXiv preprint arXiv:1601.06844*, 2016.

Qiyang Han, Tengyao Wang, Sabyasachi Chatterjee, and Richard J Samworth. Isotonic regression in general dimensions. *The Annals of Statistics*, 47(5):2440–2471, 2019.

Lauren Hannah and David Dunson. Ensemble methods for convex regression with applications to geometric programming based circuit design. *arXiv preprint arXiv:1206.4645*, 2012.

Lauren A Hannah and David B Dunson. Multivariate convex regression with adaptive partitioning. *The Journal of Machine Learning Research*, 14(1):3261–3294, 2013.

Samuel B Hopkins. Sub-gaussian mean estimation in polynomial time. *arXiv preprint arXiv:1809.07425*, page 120, 2018.

Samuel B Hopkins, Jerry Li, and Fred Zhang. Robust and heavy-tailed mean estimation made simple, via regret minimization. *arXiv preprint arXiv:2007.15839*, 2020.

Arlene KH Kim and Richard J Samworth. Global rates of convergence in log-concave density estimation. *The Annals of Statistics*, 44(6):2756–2779, 2016.

Gil Kur and Alexander Rakhlin. On the minimal error of empirical risk minimization. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 2849–2852. PMLR, 2021.

Gil Kur, Alexander Rakhlin, and Adityanand Guntuboyina. On suboptimality of least squares with application to estimation of convex bodies. pages 1–19.

Gil Kur, Yuval Dagan, and Alexander Rakhlin. Optimality of maximum likelihood for log-concave density estimation and bounded convex regression. *arXiv preprint arXiv:1903.05315*, 2019.

Gil Kur, Fuchang Gao, Adityanand Guntuboyina, and Bodhisattva Sen. Multivariate convex regression: Adaptation and sub-optimality of the least squares estimators. *in preparation*, 2020.

Meixia Lin, Defeng Sun, and Kim-Chuan Toh. Efficient algorithms for multivariate shape-constrained convex regression problems. *arXiv preprint arXiv:2002.11410*, 2020.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

Monika Ludwig, Carsten Schütt, and Elisabeth Werner. *Approximation of the Euclidean ball by polytopes*. Citeseer, 2006.

Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *The Annals of Statistics*, 47(2):783–794, 2019.

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

P. McMullen. The maximum numbers of faces of a convex polytope. *Mathematika*, 17(2):179–184, 1970.

Jaouad Mourtada, Tomas Vaškevičius, and Nikita Zhivotovskiy. Distribution-free robust linear regression. *arXiv preprint arXiv:2102.12919*, 2021.

Eliza O'Reilly and Venkat Chandrasekaran. Spectrahedral regression. *arXiv preprint arXiv:2110.14779*, 2021.

Ashwin Pananjady and Richard J Samworth. Isotonic regression with unknown permutations: Statistics, computation and adaptation. *The Annals of Statistics*, 50(1):324–350, 2022.

Warren B Powell and Huseyin Topaloglu. Stochastic programming in transportation and logistics. *Handbooks in operations research and management science*, 10:555–635, 2003.

Alexander Rakhlin, Karthik Sridharan, and Alexandre B Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

Matthias Reitzner, Carsten Schuett, and EM Werner. The convex hull of random points on the boundary of a simple polytope. *arXiv preprint arXiv:1911.05917*, 2019.

Richard J Samworth. Recent progress in log-concave density estimation. *Statistical Science*, 33(4): 493–509, 2018.

Rolf Schneider. *Convex bodies: the Brunn–Minkowski theory*. Number 151. Cambridge university press, 2014.

Carsten Schütt and Elisabeth Werner. The convex floating body. *Mathematica Scandinavica*, pages 275–290, 1990.

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

Ali Siahkamari, Durmus Alp Emre Acar, Christopher Liao, Kelly Geyer, Venkatesh Saligrama, and Brian Kulis. Faster convex lipschitz regression via 2-block admm. *arXiv preprint arXiv:2111.01348*, 2021.

Max Simchowitz, Kevin Jamieson, Jordan W Suchow, and Thomas L Griffiths. Adaptive sampling for convex regression. *arXiv preprint arXiv:1808.04523*, 2018.

Yong Sheng Soh and Venkat Chandrasekaran. Fitting tractable convex sets to support function evaluations. *Discrete & Computational Geometry*, pages 1–42, 2021.

Alexandre B Tsybakov. Optimal rates of aggregation. In *Learning theory and kernel machines*, pages 303–313. Springer, 2003a.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2003b.

Sara A van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Hal R Varian. The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, pages 945–973, 1982.

Van H Vu. Sharp concentration of random polytopes. *Geometric & Functional Analysis*, 15(6): 1284–1318, 2005.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

Appendix A. Definitions and Preliminaries

Here we collect definitions needed for the appendices below.

Definition 14 *For a fixed $\epsilon \in (0,1)$, and a function class $\mathcal{F}$ equipped with a probability measure $\mathbb{Q}$, an ϵ-net is a set that has the following property: For each $f \in \mathcal{F}$ there exists an element in this set, denoted by $\Pi(f)$, such that $\|f - \Pi(f)\|_{\mathbb{Q}} \le \epsilon$.*

Definition 15 *We denote by $\mathcal{N}(\epsilon, \mathcal{F}, \mathbb{Q})$ the cardinality of the minimal ϵ-net of $\mathcal{F}$ (w.r.t to $L_2(\mathbb{Q})$).*

Also denote by $\mathcal{N}_{[]}(\epsilon, \mathcal{F}, \mathbb{Q})$ the cardinality of the minimal ϵ-net with bracketing, which is defined as a set that has the following property: For each $f \in \mathcal{F}$ there exists two elements $f_- \le f \le f_+$ such that $\|f_+ - f_-\|_{\mathbb{Q}} \le \epsilon$.

Next, we recall the definition of $\mathbb{P}$-Donsker and non $\mathbb{P}$-Donsker classes for uniformly bounded $\mathcal{F}$.

Definition 16 *$(\mathcal{F}, \mathbb{P})$ is said to be $\mathbb{P}$-Donsker if there exists $\alpha \in (0,2)$ such that for all $\epsilon \in (0,1)$, we have $\log \mathcal{N}_{[]}(\epsilon, \mathcal{F}, \mathbb{P}) = \Theta_{\mathbb{P},\mathcal{F}}(\epsilon^{-\alpha})$, and non $\mathbb{P}$-Donsker if the same holds with $\alpha \in (2,\infty)$.*

Remark 17 *It is shown in (Bronshtein, 1976; Gao and Wellner, 2017) that*

$$\log \mathcal{N}_{[]}(\epsilon, \mathcal{F}_L(\Omega), \mathbb{P}) = \Theta_d((L/\epsilon)^{d/2})$$

and

$$\log \mathcal{N}_{[]}(\epsilon, \mathcal{F}^{\Gamma}(P), \mathbb{P}) = \Theta_d(C(P)(\Gamma/\epsilon)^{d/2}).$$

Therefore, when the dimension $d \ge 5$, both $\mathcal{F}_L(\Omega)$ and $\mathcal{F}^{\Gamma}(P)$ are non-Donsker classes.

Basic notions regarding polytopes A quick but thorough treatment of the basic theory is given in, e.g. (Schneider, 2014, §2.4). A set $P \subset \mathbb{R}^d$ is called a polyhedral set if it is the intersection of a finite set of half-spaces, i.e., sets of the form $\{x \in \mathbb{R}^d : x \cdot a \le c\}$ for some $a \in \mathbb{R}^d$, $c \in \mathbb{R}$. A polyhedral set P is called a polytope if it is bounded and has nonempty interior; equivalently, a set P is a polytope if it is the convex hull of a finite set of points and has nonempty interior.

The affine hull of a set $S \subset \mathbb{R}^d$ is defined as

$$\text{aff } S = \bigcup_{k=1}^{\infty} \{ \sum_{i=1}^{k} a_i x_i : x_i \in K, a_i \in \mathbb{R} \mid \sum_{i=1}^{k} a_i = 1 \},$$

which is the minimal affine subspace of $\mathbb{R}^d$ containing S. For a convex set K, we define its dimension to be the linear dimension of its affine hull.

For any unit vector u and any convex set K, the support set $F(K, u)$ is defined as

$$F(K, u) = \{ x \in K : x \cdot u = \max_{y \in K} y \cdot u \}.$$

(If $\max_{y \in K} y \cdot u = \infty$ then $F(K, u)$ is defined to be the empty set.)

Suppose P is a polyhedral set. For any $u \in \mathbb{S}^{m-1}$, $F(P, u)$ is a polyhedral set of smaller dimension than K. Any such $F(P, u)$ is called a face of P, and if $F(P, u)$ has dimension $m - 1$, it is called a facet of P. A polyhedral set P which is neither empty nor the whole space $\mathbb{R}^d$ has a finite and nonempty set of facets, and every face of P is the intersection of some subset of the set of facets of P. If P is a polytope, all of its faces, and in particular all of its facets, are bounded. A polytope is called simplicial if all of its facets are $(m - 1)$-dimensional simplices, which is to say, each facet F of P is the convex hull of precisely m points in aff F.

Appendix B. Proofs of Missing Parts

B.1. Proof of Theorem 10

Since the squared L^2-error scales quadratically with the function to be estimated, it suffices to prove the theorem for the class of 1-Lipschitz functions. Since the range of a 1-Lipschitz function on a domain of diameter at most 1 is contained in an interval of length 1, it is no loss to assume that the range of f^* is contained in $[0, 1]$.

The construction of a k-affine approximation to any convex 1-Lipschitz function $f^* : \Omega \to [0, 1]$, uses a combination of two tools: the theory of random polytopes in convex sets, and empirical processes.

Fix a convex body $K \subset \mathbb{R}^d$ and $n \geq d + 1$. The random polytope K_n is defined to be the convex hull of n random points $X_1, \ldots, X_n \sim U(K)$, where $U(\cdot)$ denotes the uniform distribution. It is well-known and easy to justify that K_n is a simplicial polytope with probability 1: Indeed, if $X_1, \ldots, X_n$ form a facet of K_n then in particular they lie in the same affine hyperplane, and if $k \geq d + 1$, the probability that X_k lies in the affine hull H of $X_1, \ldots, X_n$ is 0, since $K \cap H$ has volume 0. For future use we note that with probability 1, the projection of every facet of K_n on the first $d - 1$ coordinates is a $(d - 1)$-dimensional simplex, by similar reasoning.

For $s \in \{0, 1, \ldots, d - 1\}$ and P a polytope, we let $f_s(P)$ denote the number of s-dimensional faces of P. The first result regarding random polytopes that we need appears in (Bárány, 1989, Corollary 3):

Theorem 18 *Let $d \geq 1$, $1 \leq s \leq d-1$ and a convex body $K \subset \mathbb{R}^d$. Then, there exists $C(d,s) \leq C_1(d)$ such that*

$$\mathbb{E}[f_s(K_n)] \leq C(d,s)n^{\frac{d-1}{d+1}}.$$

We will also use the following result that was derived in Dwyer (1988):

Theorem 19 *Let $P \subset B_d$ be a polytope, and let $Y_1, \ldots, Y_m \underset{i.i.d.}{\sim} \mathbb{P}_P$. Then, $P_m = \operatorname{conv}(Y_1, \ldots, Y_m)$ is a simplicial polytope with probability 1, and the following holds:*

$$\mathbb{E}\mathbb{P}_P(P \setminus P_m) = O_d(C(P)m^{-1}\log(m)^{d-1}),$$

The other result that we need from empirical processes appears as Lemma 8 in the main text.

We now describe our construction. Given a 1-Lipschitz function $f^* : \Omega \to [0,1]$, define the convex body

$$K = \{(x,y) : x \in \Omega, y \in [0,2] \,|\, f^*(x) \leq y\}.$$

In other words, K is the epigraph of the function f^*, intersected with the slab $\mathbb{R}^d \times [0,2]$. Note that $\operatorname{vol}_{d-1}(\Omega) \leq \operatorname{vol}_d(K) \leq 2\operatorname{vol}_{d-1}(\Omega)$, since $\operatorname{Im} f^* \subset [0,1]$.

Let $n = \lfloor k^{\frac{d+2}{d}} \rfloor$, and consider the random polytope $K_n \subset K$. Let Ω_k be the projection of K_n to $\mathbb{R}^d$, and define the function $f_k : \Omega_k \to [0,2]$ by

$$f_k(x) = \min\{y \in \mathbb{R} : (x,y) \in K_n\},$$

i.e., f_k is the lower envelope of K_n. In particular, since $K_n \subset K$, f_k lies above the graph of f^*. We would like to show that with positive probability, f_k satisfies the properties in the statement of the theorem. We treat each property in turn.

f_k is k-simplicial with probability at least $9/10$: Using Theorem 18, and Markov's inequality K_n has at most $10C(d)n^{\frac{d}{d+2}} = C'(d)k$ facets (recall that all facets of K_n are simplices with probability 1). Letting $\triangle_1, \ldots, \triangle_F$ be the bottom facets of K_n, and letting $\pi : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ be the projection onto the first factor, $\pi(\triangle_1), \ldots, \pi(\triangle_F)$ is a triangulation of Ω_k and for each $i = 1, \ldots, F$, $f_k|_{\triangle_i}$ is affine, as its graph is simply $\triangle_i$.

Bounding $\mathbb{P}(\Omega \backslash \Omega_k)$ with probability at least $9/10$: Since Ω_k is the projection of K_n to $\mathbb{R}^d$, it is equivalently defined as $\operatorname{conv}(\pi(X_1), \ldots, \pi(X_n))$ where $X_1, \ldots, X_n$ are independently chosen from the uniform distribution on K, and π is the projection onto the first d coordinates as above. $\pi(X_i)$ is not uniformly distributed on Ω, so we cannot apply Theorem 19 (and Markov's inequality directly). Instead, we re-express $\pi(X_i)$ as a mixture of a uniform distribution and another distribution, and apply Theorem 19 to the points which come from the uniform distribution.

In more detail, note that we may write $K = K_1 \cup K_2$ where $K_1 = (\Omega \times [0,1]) \cap \operatorname{epi} f$ and $K_2 = \Omega \times [1,2]$, since $f \leq 1$. Let $p = \frac{\operatorname{vol}(K_2)}{\operatorname{vol}(\Omega)} \geq \frac{1}{2}$. The uniform distribution from K can be sampled from as follows: with probability p, sample uniformly from K_2, and with probability $1-p$

sample uniformly from K_1. Clearly, if X is uniformly distributed from K_2 then $\pi(X)$ is uniformly distributed on Ω. Hence, Ω_k can be constructed as follows: draw M from the binomial distribution $B(n,p)$ with n trials and success probability p, then sample M points $X_1,\dots,X_M$ uniformly from Ω and sample $k-M$ points $X'_1,\dots,X'_{k-m}$ from some other distribution on Ω, which doesn't interest us; then set $\Omega_k = \operatorname{conv}(X_1,\dots,X_M,X'_1,\dots,X'_{k-M})$. In particular, $\mathbb{P}(\Omega\backslash\Omega_k) \geq \mathbb{P}(\Omega\backslash\Omega_M)$, so it is sufficient to bound the RHS with high probability.

By the usual tail bounds on the binomial distribution, $M \geq \frac{np}{2} \geq \frac{n}{4}$ with probability $1-e^{-\Omega(n)}$. Hence, by Theorem 19 we obtain

$$\begin{aligned}\mathbb{E}\mathbb{P}(\Omega\backslash\Omega_M) &\leq \mathbb{E}\mathbb{P}(\Omega\backslash\operatorname{conv}\{X_1,\dots,X_{n/4}\}) + C(d)e^{-c(d)n} \leq O_d(C(\Omega)n^{-1}\log(n)^{d-1})\\ &\leq O_d(C(\Omega)k^{-\frac{d+2}{d}}\log(k)^{d-1}),\end{aligned}$$

and we obtain $\mathbb{P}(\Omega\backslash\Omega_M) \leq 10C(\Omega)k^{-\frac{d+2}{d}}\log(k)^{d-1}$ with probability at least $\frac{9}{10}$ by Markov's inequality.

Bounding $\int(f-f_k)^2 d\mathbb{P}$ with probability at least $9/10$: Finally, we wish to bound the $\mathbb{L}^2(\mathbb{P})$-norm of f^*-f_k. To do this, we use the same strategy, arguing that on average, k of the points of K_n can be thought of as drawn from the uniform distribution on a thin shell of width $k^{-\frac{2}{d}}$ lying above the graph of f^*, which automatically bounds the empirical L^2-norm $\int(f^*-f_k)^2\, d\mathbb{P}_n$ and hence the L^2-norm by Lemma 8.

Now for the details. Set $\epsilon = k^{-\frac{2}{d}}$, and define

$$K_\epsilon = \{(x,y) : x\in\mathbb{R}^d, y\in[0,2] \,|\, f^*(x) \leq y \leq f^*(x)+\epsilon\},$$

i.e., $K_\epsilon \subset K$ is just the strip of width ϵ lying above the graph of f^*. By Fubini, K_ϵ has volume $\epsilon\operatorname{vol}(\Omega) \geq \frac{\epsilon}{2\operatorname{vol}(K)}$, and if X is uniformly distributed on K_ϵ, $\pi(X)$ is uniformly distributed on Ω. Hence, we can argue precisely as in the preceding: with probability $1-e^{-\Omega(n)}$,

$$L := |\{X_i : X_i \in K_\epsilon\}| \geq \frac{\epsilon n}{4} = \frac{k}{4}.$$

Conditioning on L for some $L \geq \frac{k}{4}$ and letting $X_1,\dots,X_L$ be the points drawn from K which lie in K_ϵ we have that $\pi(X_1),\dots,\pi(X_L)$ are uniformly distributed on Ω. Moreover, for any $i\in\{1,\dots,n\}$, $X_i\in K_n$ and so it lies above the graph of f_k, but also $X_i \in K_\epsilon$ and so it lies below the graph of $f^*+\epsilon$. Combining these two facts yields

$$\forall 1\leq i\leq L: \quad f_k(\pi(X_i)) \leq (X_i)_{d+1} \leq f^*(\pi(X_i))+\epsilon,$$

where $(\cdot)_{d+1}$ denotes the $d+1$ coordinate. Hence,

$$\forall 1\leq i\leq L: \quad f^*(\pi(X_i)) \leq f_k(\pi(X_i)) \leq f^*(\pi(X_i)) + Ck^{-2/d}. \tag{21}$$

Thus, letting $\mathbb{P}_L = \frac{1}{L}\sum_{i=1}^L \delta_{\pi(X_i)}$ denote the empirical measure on $\pi(X_1),\dots,\pi(X_L)$, we obtain

$$\int_\Omega (f^*-f_k)^2\, d\mathbb{P}_L \leq \frac{1}{L}\sum_{i=1}^L \epsilon^2 = \epsilon^2 = k^{-\frac{4}{d}}.$$

Since the $\pi(X_i)$ are drawn uniformly from Ω, if we knew that f_k were 1-Lipschitz it would follow from Lemma 8 that

$$\int_\Omega (f^* - f_k)^2 \, d\mathbb{P} \leq k^{-\frac{4}{d}} + CL^{-\frac{4}{d}} = C' k^{-\frac{4}{d}},$$

with high probability.

We do not know, however, that f_k is 1-Lipschitz. To get around this, define the function $\hat{f}_k$ as the function on Ω_k whose graph is $\operatorname{conv}\{(\Pi(X_i), f^*(\Pi(X_i)))\}_{i=1}^L$. Unlike f_k, $\hat{f}_k$ is necessarily 1-Lipschitz since f^* is (see, e.g., the argument in the paragraph below equations (13)-(15)), so by Lemma 8, it follows that

$$\int_\Omega (f^* - \hat{f}_k)^2 \, d\mathbb{P} \leq C_1 k^{-\frac{4}{d}}$$

with probability at least $1 - C(d)\exp(-c(d)k)$. Also, by (21),

$$\forall 1 \leq i \leq L: \quad \hat{f}_k(\pi(X_i)) \leq f_k(\pi(X_i)) \leq \hat{f}_k(\pi(X_i)) + C_1 k^{-2/d}.$$

It easily follows by the definitions of f_k and $\hat{f}_k$ as convex hulls that on the domain $\Omega_{\Pi(X)} := \operatorname{conv}\{(\Pi(X_i))\}_{i=1}^L$, we have

$$f^* \leq \hat{f}_k \leq f_k \leq \hat{f}_k + Ck^{-2/d}.$$

Hence, we conclude that

$$\begin{aligned}\int_{\Omega_{\Pi(X)}} (f_k - f^*)^2 \, d\mathbb{P} &\leq 2\int_{\Omega_{\Pi(X)}} (\hat{f}_k - f^*)^2 \, d\mathbb{P} + 2\int_{\Omega_{\Pi(X)}} (f_k - \hat{f}_k)^2 \, d\mathbb{P} \\ &\leq 2\int_{\Omega_{\Pi(X)}} (\hat{f}_k - f^*)^2 \, d\mathbb{P} + 2\|f_k - \hat{f}_k\|^2_{L^\infty(\Pi(X))} \leq C_2 k^{-4/d}\end{aligned}$$

with high probability.

Now, using Theorem 19 and Markov's inequality, we also know that

$$\mathbb{P}(\Omega \setminus \Omega_{\Pi(X)}) \leq 20C(\Omega)k^{-1}\log(k)^{d-1}.$$

with probability at least $\frac{19}{20}$. Conditioned on this event, and using the fact that f_k is uniformly bounded by 1, we obtain

$$\int_{\Omega\setminus\Omega_{\Pi(X)}} (f_k - f^*)^2 \leq C(\Omega)k^{-1}\log(k)^{d-1}.$$

On the intersection of the two events defined above, which has probability at least $\frac{9}{10}$, we have $\int_\Omega (f_k - f^*)^2 \leq C_3 k^{-\frac{4}{d}} + C(\Omega)k^{-1}\log(k)^{d-1}$.

Deriving the theorem Since we have three events each of which hold with probability at least $9/10$, then the intersection of these events is not empty. Therefore, an f_k satisfying all the desired properties exists, and the theorem follows.

B.2. Proof of Lemma 11

We start with the following easy lemma:

Lemma 20 *The following event holds with probability at least* $1 - n^{-3d}$*:*

$$\forall 1 \le i \le k(n) \;\; s.t. \; \mathbb{P}(\triangle_i) \ge C_3 d \log(n)/n : \;\; 2^{-1}\mathbb{P}(\triangle_i) \le \mathbb{P}_n(\triangle_i) \le 2\mathbb{P}(\triangle_i). \tag{22}$$

Proof The lemma follows for the fact that $n \cdot \mathbb{P}_n(S) \sim Bin(n, \mathbb{P}(S))$, along with the concentration inequality (cf. (Boucheron et al., 2013)) for binomial random variables: for all $\epsilon \in (0,1)$,

$$\Pr\left(\left|\frac{\mathbb{P}_n(S)}{\mathbb{P}(S)} - 1\right| \le \epsilon\right) \le 2\exp(-c\min\{\mathbb{P}(S), 1 - \mathbb{P}(S)\}n\epsilon^2).$$

By taking $\epsilon = 1/2$, and choosing C to be large enough, we conclude that for any particular $\triangle_i$,

$$\mathbb{P}(\triangle_i) \ge C_3 d \log(n)/n : \;\; 2^{-1}\mathbb{P}(\triangle_i) \le \mathbb{P}_n(\triangle_i) \le 2\mathbb{P}(\triangle_i)$$

with probability at least $1 - n^{-(3d+1)}$. Taking the union bound over all $k(n)$ simplices, the claim follows. ■

The main step is the following lemma, which shows that for any given simplex $\triangle_i$, if we draw $Cd \log n$ points from the uniform distribution on $\triangle_i$ for sufficiently large C, then there exists some subset S of these points whose convex hull P covers almost all of the simplex and can also be triangulated by a polylogarithmic number of simplices whose vertices lie in S.

Lemma 21 *Let* $S \subset \mathbb{R}^d$ *be a simplex, and* $m \ge C_3 d \log(n)$*, for some large enough* $C_3 \ge 0$*. Let* $Y_1, \ldots, Y_m \sim \mathbb{P}_S$*. Then, with probability at least* $1 - n^{-3d}$ *there exists a set* $\mathcal{A}$ *of simplices contained in* S *with disjoint interiors of cardinality* $|\mathcal{A}| \le C_d \log(m)^{d-1}$ *such that*

$$\mathbb{P}_S(S \setminus \bigcup \mathcal{A}) = O_d(m^{-1}\log(n)\log(m)^{d-1}).$$

Proof For each $s \in \{0, 1, \ldots, d-1\}$ and P a polytope, we let $f_s(P)$ denote the number of s-dimensional faces of P. We need the following result, which was first proven in (Dwyer, 1988); for more details see the recent paper Reitzner et al. (2019).

Theorem 22 *Let* $S \subset \mathbb{R}^d$ *be a simplex, and let* $Y_1, \ldots, Y_m \sim \mathbb{P}_S$*. Then,* $S_m = \mathrm{conv}(Y_1, \ldots, Y_m)$ *is a simplicial polytope with probability* 1*, and the following holds:*

$$\mathbb{E}\mathbb{P}_S(S \setminus S_m) = O_d(m^{-1}\log(m)^{d-1}),$$

and

$$\mathbb{E}f_{d-1}(S_m) = O_d(\log(m)^{d-1}).$$

This theorem does not give us what we need directly, since it treats only expectation while we require high-probability bounds. (To the best of our knowledge, sub-Gaussian concentration bounds are not known for the random variables $f_{d-1}(P_m), \mathbb{P}(S \setminus S_m)$ when S is a simplex, cf. (Vu, 2005).) This necessitates using a partitioning strategy. We divide our $Y_1, \ldots, Y_m$ into $C_1 d \log(n)$ blocks, for C_1 to be chosen later, each with $m(n) := \frac{m}{C_1 d \log(n)}$ samples drawn uniformly from $\triangle$. Let $P_1, \ldots, P_B$ be the convex hulls of the points in each block, each of which are independent realizations of the random polytope $S_{m(n)}$. For each P_i, Markov's inequality and a union bound yield that with probability at least $\frac{1}{3}$,

$$\begin{aligned} \mathbb{P}_S(S \setminus P_i) &\leq 3 \cdot \mathbb{E}\mathbb{P}_S(S \setminus P_i) \leq C_1(d) m(n)^{-1} \log(m(n))^{d-1} \\ &= \frac{C_2(d) \log(m)^{d-1} \log(n)}{n}, \end{aligned} \tag{23}$$

and

$$f_{d-1}(P_i) \leq 3\mathbb{E} f_{d-1}(S_{m(n)}) \leq O_d(\log(m(n))^{d-1}) \leq O_d(\log(m)^{d-1}). \tag{24}$$

Since there are $C_1 d \log(n)$ independent P_i, at least one of them will satisfy these conditions with probability $1 - \left(\frac{2}{3}\right)^{C_1 d \log n}$, and we may choose C_1 so that this is at least $1 - n^{-3d}$.

Conditioned on the existence of P_i satisfying (23) and (24), we take one such P_i and triangulate it by picking any point among the original $Y_1, \ldots, Y_m$ lying in the interior of P_i and connecting it to each of the $(d-1)$-simplices making up the boundary of P_i. The set $\mathcal{A}$ is simply the set of d-simplices in this triangulation. ■

Now, to obtain Lemma 11, we condition on the event of Lemma 20 and apply Lemma 21 to each $\triangle_i$ such that $\mathbb{P}(\triangle_i) \geq Cd \log(n)/n$, with the $Y_1, \ldots, Y_m$ taken to be the points of $X_{n+1}, \ldots, X_{2n}$ drawn from $\mathbb{P}$ which fall inside of $\triangle_i$. Using the fact that $\mathbb{P}_n(\triangle_i) \geq 0.5\mathbb{P}(\triangle_i)$, we see that $m \geq Cd \log n$ for each $\triangle_i$, so Lemma 21 is in fact applicable. In addition, the bounds on the cardinality of $\mathcal{S}_X^i$ and on the volume of $\triangle_i$ left uncovered by the simplices in $\mathcal{S}_X^i$ follow immediately by substituting $c\mathbb{P}(\triangle_i)$ for m in the conclusions of Lemma 21. For i such that $\mathbb{P}(\triangle_i) \leq Cd \log(n)/n$, we take $\mathcal{S}_X^i$ to be the empty set.

B.3. Proofs of Lemmas and 12 and 13

In several places, we will use a high-probability estimator for the mean of a random variable presented in (Devroye et al., 2016):

Lemma 23 *Let $\delta \in (0,1)$ and let $Z_1, \ldots, Z_k$ be i.i.d. samples from a distribution on $\mathbb{R}$ with finite variance σ_Z^2. There exists an estimator $\hat{f}_\delta : \mathbb{R}^k \to \mathbb{R}$ with a runtime of $O(k)$, such that with probability at least $1 - \delta$,*

$$(\hat{f}_\delta(Z_1, \ldots, Z_k) - \mathbb{E}Z)^2 \leq \frac{8\sigma_Z^2 \cdot \log(2/\delta)}{k}.$$

In the first sub-subsection, we construct the estimator for the L_1 norm of a convex function g defined on a convex body K, which is the content of Lemma 12. In the second sub-subsection, we show how to “upgrade” this estimator to an estimator of the L_2-norm in the special case that K is a simplex.

The final step will be to estimate the L_2 norm of g under the assumptions of Lemma 13, by using Lemma 12 and the claim of Lemma 13 will follow.

B.3.1. Proof of Lemma 12

We only prove this Lemma for K a simplex, which is what we require for our algorithm. The proof for a general K can be done in similar fashion, by placing K in John position (besides the computational aspects).

Let S be the regular simplex inscribed in the unit ball B_d, and for each $t \in [0,1]$, denote by $S^t := (1-t)S$. We will use the following geometric facts, which can be extracted from the statements and proofs of (Gao and Wellner, 2017, Lemmas 2.6-2.7).

Lemma 24 *Let $g : S \to \mathbb{R}$ be a convex function, and let $\|g\|_1 = \int_S |g|$ be its L^1-norm. We have:*

- *$g \geq -C_d \|g\|_1$ on all of S.*
- *For each $\delta \in (0,1)$, $g|_{S^\delta}$ is is $C_d \delta^{-(d+1)} \|g\|_1$-Lipschitz and satisfies $g|_{S^\delta} \leq C_d \delta^{-(d+1)} \|g\|_1$.*

We immediately obtain the following corollary:

Corollary 25 *For any $a \in (0,1)$, there exists a constant δ such that g restricted to S^δ is uniformly bounded by $C_d s^{-(d+1)} \|g\|_1$; moreover, letting $g_- = \min(g, 0)$, we have*

$$\int_{S \setminus S^\delta} |g_-| \, dU(x) \leq a \|g\|_1.$$

Define a probability density p_S on S by the formula

$$p_S(x) := \int_S \frac{1_{B_y}(x)}{U(B_y)} dU(y), \tag{25}$$

where we define B_y to be the largest ball centered on y which is contained in S. For any simplex $\triangle$, we define the density $p_\triangle$ as the pushforward of p_S under the affine transformation T sending S to T.

Lemma 26 *Let M, M' be positive constants, let $\triangle$ be a simplex contained in the unit ball B_d, and let $g : \triangle \to [-M' \|g\|_{L_1(U(\triangle))}, M' \|g\|_{L_1(U(\triangle))}]$ be a convex $M \|g\|_{L_1(U(\triangle))}$-Lipschitz function which is orthogonal to affine functions, i.e.,*

$$\int_\triangle g w \, dU = 0$$

for any affine function w. Then there exist positive constants $c_1 = c_1(M, M', d)$, $C_1 = C_1(M, M', d)$ such that:

$$c_1 \|g\|_{L_1(U(\triangle))} \leq \int g(x) p_\triangle(x)\, dx \leq C_1 \|g\|_{L_1(U(\triangle))}, \tag{26}$$

In addition, for every affine function w,

$$\int w p_\triangle \, dx = \int_\triangle w \, dU(x).$$

Moreover,

$$\max_{x \in \triangle} p_\triangle(x) \leq \alpha_d := 2^d \frac{d+1}{d-1} \frac{v_{d-1}}{v_d}$$

where v_d is the volume of the unit ball in dimension d, and there exists an efficient algorithm to compute $p_\triangle(x)$ for any $x \in \triangle$.

The idea behind the proof of this lemma is that for any point $x \in \triangle$ and a ball $B_x \subset \triangle$, the average $\bar{g}(x)$ of g over B_x is at least $g(x)$, with equality iff g is affine on B_x. If g is nonzero and orthogonal to affine functions, the averaged function $\bar{g}$ must have positive integral, and a compactness argument then yields a lower bound on $\frac{1}{\|g\|_1} \int \bar{g}$. The full proof is given at the end of the sub-subsection.

Using the above results we can estimate the L_1 norm $g : \triangle \to [-1, 1]$.

Letting T be the unique affine transformation such that $T\triangle = S$, we define the shrunken simplex $\triangle_\delta$ by the $\triangle_\delta := T^{-1}(T\triangle)^{\delta(l,d)}$, where $a = 1/10$ and $\delta(a, d)$ is defined in Corollary 25. The proof involves analysis of several cases.

Case 1: $\|g \mathbb{1}_{\triangle \setminus \triangle_\delta}\|_1 \geq \frac{1}{2}\|g\|_1$, i.e. most of the L_1-norm of g comes from the shell $\triangle \setminus \triangle_\delta$. Using Corollary 25, we know that

$$\|g\|_1 \geq \int_{\triangle \setminus \triangle_\delta} g dU(x) = \int_{\triangle \setminus \triangle_\delta} g^+ dU(x) + \int_{\triangle \setminus \triangle_\delta} g^- dU(x) \geq (3/20) \cdot \|g\|_1.$$

Therefore it is enough to estimate the mean (scaled by $U(\triangle \setminus \triangle_\delta)$) of the r.v. $g(X)$ where $X \sim U(\triangle \setminus \triangle_\delta)$, which can be done using the samples that fall in $\triangle \setminus \triangle_\delta$. By Lemma 23 above, we conclude that using these samples we have an estimator $\hat{f}_{(1)}$ such that with probability at least $1 - \delta$,

$$\begin{aligned} \left| \hat{f}_{(1)} - \int_{\triangle \setminus \triangle_\delta} g \, dU \right|^2 &\leq U(\triangle \setminus \triangle_\delta)^2 \frac{Cd(\sigma^2 + \|g \mathbb{1}_{\triangle \setminus \triangle_\delta}\|_2^2) \log(2/\delta)}{U(\triangle \setminus \triangle_\delta) m} \\ &\leq C_d \cdot \frac{(\sigma^2 + \|g\|_2^2) \log(2/\delta)}{m}, \end{aligned} \tag{27}$$

where we used that fact that $U(\triangle \setminus \triangle_\delta) \geq c_d$.

Case 2: If we are not in Case 1, we must have $\|g\mathbb{1}_{\triangle_\delta}\|_1 \geq \frac{1}{2}\|g\|_1$, i.e., most of the L^1-norm of g comes from the inner simplex $\triangle_\delta$. Decompose $g = w_g + (g - w_g)$, where $w_g = \operatorname{argmin}_{w \text{ affine}} \|g - w\|_{L_2(U(\triangle_\delta))}$ is the $L^2(\triangle_\delta)$-projection of g onto the space of affine functions. Note that by orthogonality, we have

$$\max(\|w_g\|_{L_2(U(\triangle_\delta))}, \|g - w_g\|_{L_2(U(\triangle_\delta))}) \leq \|g\|_{L_2(U(\triangle_\delta))} \tag{28}$$

by orthogonality, while by Lemma 24, we have

$$\|g\|_{L_1(U(\triangle_\delta))} \leq \|g\|_{L_2(U(\triangle_\delta))}\| \leq C_d\|g\|_{L_1(U(\triangle_\delta))}. \tag{29}$$

By the triangle inequality, we must have either $\|w_g\mathbb{1}_{\triangle_\delta}\|_1 \geq \frac{1}{4}\|g\|_1$ or $\|(g - w_g)\mathbb{1}_{\triangle_\delta}\|_1 \geq \frac{1}{4}\|g\|_1$; we analyze each case below.

Case 2a: First, suppose $\|w_g\mathbb{1}_{\triangle_\delta}\|_1 \geq \frac{1}{4}\|g\|_1$. Using half of the samples that fall into $\triangle_\delta$, we may apply Lemma 9, giving us an affine $\hat{w}_g$ such that with probability at least $1 - \delta$,

$$\|\hat{w}_g - w_g\|_2^2 \leq C_d(\sigma^2 + \|g\|_2^2)\frac{\log(2/\delta)}{m}.$$

Writing $\bar{f}_{(2a)} := \|\hat{w}_g\|_1$, we conclude that

$$\frac{1}{4}\|g\|_1 - C_d(\sigma + \|g\|_2)\sqrt{\frac{\log(2/\delta)}{m}} \leq \bar{f}_{(2a)} \leq \|g\|_1 + C_d(\sigma + \|g\|_2)\sqrt{\frac{\log(2/\delta)}{m}}. \tag{30}$$

Note that the right-hand inequality does not require the assumption $\|w_g\mathbb{1}_{\triangle_\delta}\|_1 \geq \frac{1}{4}\|g\|_1$.

Case 2b: Now suppose $\|(g - w_g)\mathbb{1}_{\triangle_\delta}\|_1 \geq \frac{1}{4}\|g\|_1$. To estimate $\|(g - w_g)\mathbb{1}_{\triangle_\delta}\|_1$, we will use our Lemma 26. Note that by the definition of $\triangle_\delta$, we may assume by Lemma 25 that $\max\{M', M\} \leq C(d)$ (in Lemma 26). Therefore, we conclude that

$$c_1(d)\|g - w_g\|_1 \leq \int_{\triangle_\delta}(g - w_g)\cdot p_{\triangle_\delta} \leq C_1(d)\|g - w_g\|_1.$$

By our assumption $\|(g - w_g)\mathbb{1}_{\triangle_\delta}\|_1 \geq \frac{1}{4}\|g\|_1$; we also have

$$\|(g - w_g)\mathbb{1}_{\triangle_\delta}\|_1 \leq \|g\|_1 + \|\hat{w}_g\|_1 \leq \|g\|_1 + \|\hat{w}_g - w$$

so it follows that

$$c_2(d)\|g\|_1 \leq \int_{\triangle_\delta}(g - \hat{w}_g)\cdot p_{\triangle_\delta} \leq C_2(d)\|g\|_1.$$

Therefore, it is enough to estimate $\int(g - \hat{w}_g)\cdot p_{\triangle_\delta}$, using the the second half of the samples that fall into $\triangle_\delta$.

To do this, we simulate sampling from $p_{\triangle_\delta}$ given samples from $U(\triangle_\delta)$ and their corresponding noisy samples of $g - \hat{w}_g$. The idea is simply to use rejection sampling (Devroye, 1986): given a single sample $X \sim U(\triangle_\delta)$, we keep it with probability $p_{\triangle_\delta}(x)\cdot\frac{1}{\alpha_d}$. Conditioned on keeping the sample, X is distributed according to $p_{\triangle_\delta}$. If we are given $m/2$ i.i.d. samples from $U(\triangle_\delta)$, then with probability $1 - e^{-c'_d m}$ the random number of samples N we obtain from $p_\triangle$ by this method is at least $c(\alpha_d) \cdot m \geq c_1(d)m$, and conditioned on N, these samples are i.i.d. $p_{\triangle_\delta}$. We condition on this event going forward. Now, using these N samples, Lemma 23 gives an estimator $\bar{f}_{(2b)}$ such that

$$\left| \bar{f}_{(2b)} - \int (g - \hat{w}_g) \cdot p_{\triangle_\delta} \right| \leq C_d(\sigma + \|g\|_2)\sqrt{\frac{\log(2/\delta)}{m}}. \tag{31}$$

with probability at least $1 - \max\{\delta, e^{-cm}\}$.

Note that each of $\bar{f}_{(1)}, \bar{f}_{(2a)}, \bar{f}_{(2b)}$ is bounded from above by $C_d\|g\|_1 + C_d(\sigma + \|g\|_2)\sqrt{\frac{\log(2/\delta)}{m}}$, irrespective of whether we are in the case for which the estimator was designed; this follows from (27), (30), (31), respectively.

Finally, by using $\bar{f}_{(1)}, \bar{f}_{(2a)}, \bar{f}_{(2b)}$, we conclude that with probability $1 - 3\max\{\delta, e^{-cm}\}$ at least

$$c_1(d)\|g\|_1 - C_d(\sigma + \|g\|_2)\sqrt{\frac{\log(2/\delta)}{m}} \leq \bar{f}_{(1)} + \bar{f}_{(2a)} + \bar{f}_{(2b)} \leq C_1(d)\|g\|_1 + C_d|\sigma + \|g\|_2|\sqrt{\frac{\log(2/\delta)}{m}},$$

and the claim follows.

Proof [Proof of Lemma 26] Recall that it suffices to show that

$$c_1(M, M') \leq \int_S g(x)p_S(x)dx \leq C_1(M, M').$$

Note that the function g is convex and in particular subharmonic, i.e., for any ball B_x with center x contained in S we have

$$\frac{1}{U(B_x)} \int_{B_x} g dU(x) \geq g(x),$$

where U denotes the uniform measure on the regular simplex S. g is non-affine and hence strictly subharmonic (as convex harmonic functions are affine), so there exists some x such that for any ball $B_x \subset S$ centered on x, the above inequality is strict, since subharmonicity is a local property. As g is convex and in particular continuous, the inequality is strict on some open set of positive measure. We obtain that for a *non*-affine convex function that

$$\begin{aligned} \int_S g(x)p_S(x)\, dU(x) &= \int_S \int_S g(x)1_{B_y}(x)\, dU(y)\, dU(x) \\ &= \int_S \left(\frac{1}{U(B_y)} \int_{B_y} g(x)\, dx \right) dU(y) > \int_S g(y) dU(y) = 0, \end{aligned} \tag{32}$$

i.e. we showed that for a non-harmonic g that $\int_S g(x)p_S(x)\, dx > 0$.

Now, we show why Eq. (32) actually implies the lower bound of Eq. (26), which is certainly not obvious a priori. However, it follows from a standard compactness argument. The set $\mathcal{C}$ of convex M-bounded, M'-Lipschitz functions with norm 1 that is orthogonal to the affine functions is closed in $L^\infty(S)$, and also equicontinuous due to the Lipschitz condition. Hence, by the Arzela-Ascoli theorem it is compact in $L^\infty(S)$, and we conclude that

$$A = \left\{ \int_S g(x)p_S(x)dx : g \in \mathcal{C} \right\}$$

is compact; but (32) implies that $A \subset (0, \infty)$, which finally implies the existence of $c(M, M', d) > 0$ such that $S \subset [c(M, M', d), \infty)$. As for the upper bound in (26), it follows immediately from the boundedness of p_S, which we prove below.

We claim that in this case (25) can be evaluated analytically as a function of x, though the formulas are sufficiently complicated that this is best left to a computer algebra system. Indeed, we note that $y \in S$ contributes to the integral at x if and only if x is closer to y than y is to the boundary of S. The regular simplex can be divided into $d+1$ congruent cells $C_1, \ldots, C_{d+1}$ such that the points in C_i are closer to the i-th facet of the simplex than to any other facet (in fact, C_i is simply the convex hull of the barycenter of S and the ith facet); for any $y \in C_i$, $x \in B_y$ if and only if x is closer to y than y is to the hyperplane H_i containing C_i. But the locus of points equidistant from a fixed point x and a hyperplane is the higher-dimensional analog of an elliptic paraboloid, for which it's easy to write down an explicit equation. Letting $P_{i,x}$ be the set of points on x's side of the paraboloid (namely, those closer to x than to H_i), we obtain

$$p_S(x) = \sum_{i=1}^{d+1} \int_{C_i \cap P_{i,x}} \frac{dy}{v_d \cdot d(y, H_i)^d}.$$

Each region of integration $C_i \cap P_{i,x}$ is defined by several linear inequalities and a single quadratic inequality, and the integrand can be written simply as $\frac{1}{y_i^d}$ in an appropriate coordinate system. It is thus clear that the integral can be evaluated analytically, as claimed.

Finally, we need to show that $p_S(x)$ is bounded above by α_d. By symmetry,

$$p_S(x) \leq \sum_{i=1}^{d+1} \frac{dy}{\Omega_d \cdot d(y, H_i)^d} \leq \frac{d+1}{\Omega_d} \cdot \sup_{x \in \mathbb{R}^n} \int_{P_{1,x}} \frac{dy}{d(y, H_1)^d}. \tag{33}$$

Fix x, and choose coordinates such that $H_1 = \{x_1 = 0\}$ and $x = (x_0, 0, \ldots, 0)$ with $x_0 > 0$. Then for any $y = (t, z)$ with $t \in \mathbb{R}$, $z \in \mathbb{R}^{d-1}$, y lies in $P_{1,x}$ if $t^2 \geq (x_0 - w)^2 + |z|^2$, or $2tx_0 - x_0^2 \geq |z|^2$. Hence,

$$\begin{aligned}
\int_{P_{1,x}} \frac{dy}{d(y, H_1)^d} &= \int_{\frac{x_0}{2}}^{\infty} \frac{dt}{t^d} \int_{\mathbb{R}^{d-1}} 1_{|z|^2 \leq 2tx_0 - x_0^2}(z)\, dz \\
&\leq \int_{\frac{x_0}{2}}^{\infty} \frac{dt}{t^d} \cdot \Omega_{d-1} (2tx_0 - x_0^2)^{\frac{d-1}{2}} \leq \Omega_{d-1} \int_{\frac{x_0}{2}}^{\infty} \frac{dt}{t^d} (2tx_0)^{\frac{d-1}{2}} \\
&= \Omega_{d-1} 2^{\frac{d-1}{2}} x_0^{\frac{d-1}{2}} \int_{\frac{x_0}{2}}^{\infty} \frac{dt}{t^{\frac{d+1}{2}}} \\
&= \Omega_{d-1} 2^{\frac{d-1}{2}} x_0^{\frac{d-1}{2}} \cdot \left(\frac{d-1}{2}\right)^{-1} \left(\frac{x_0}{2}\right)^{-\frac{d-1}{2}} = \Omega_{d-1} \cdot \frac{2^d}{d-1},
\end{aligned}$$

and substituting in (33) gives the desired bound. ■

B.3.2. PROOF OF LEMMA 13

Recall that we are given a convex L-Lipschitz function g satisfying $\|g\|_\infty \leq L$; by homogeneity, we may assume $L = 1$. Our goal in this subsection is to estimate $\|g\|_2$ up to polylogarithmic factors given an estimate of $\|g\|_1$, where $g : \triangle \to [-1, 1]$. This part requires the additional assumption that $\|g\|_2 \geq \frac{Cd^{1/2} \log n}{n^{1/2}}$.

For this section, we will need the following classical result about the floating body of a simplex (Bárány and Larman, 1988; Schütt and Werner, 1990).

Lemma 27 *For a simplex S and $\epsilon \in (0, 1)$. let S_ϵ be its ϵ-convex floating body, defined as*

$$S_\epsilon := \bigcap \{K : K \subset S \textit{ convex }, \mathrm{vol}(S \setminus K) \leq \epsilon \, \mathrm{vol}(S)\},$$

and let $S(\epsilon) = S \backslash S_\epsilon$ be the so-called wet part of S. Then $\mathrm{vol}(S(\epsilon)) \leq C_d \epsilon \log(\epsilon^{-1})^{d-1} \mathrm{vol}(S)$.

We also note that for any particular ϵ and $x \in S$ one can check in polynomial time whether $x \in S(\epsilon)$: indeed, letting

$$\begin{aligned} H^+_{x,u} &= \{y \in \mathbb{R}^n : \langle y, u\rangle \geq \langle x, u\rangle\} \\ H_{x,u} &= \partial H^+_{x,u} = \{y \in \mathbb{R}^n : \langle y, u\rangle = \langle x, u\rangle\}, \end{aligned}$$

the function $u \mapsto U(S \cap H^+_{x,u})$ is smooth on S^{d-1} outside of the closed, lower-dimensional subset A where $H_{x,u}$ is not in general position with respect to some face of u, and, moreover, is given by an analytic expression in each of the connected components C_i of $S^{d-1} \backslash A$. It can thus be determined algorithmically whether $\min_i \inf_{x \in C_i} U(S \cap H^+_{x,u}) \leq \epsilon$, i.e., whether $x \in S(\epsilon)$.

Let $v = \mathbb{P}(S)$, and let $i_{min} = \min(\lfloor \log_2(C_d \log(n)^{d+1} v^{-1})\rfloor, 0)$; note that since $\|g\|_1 \geq v \geq \frac{\log(n)^d}{n}$, $|i_{min}| \leq C \log n$. Set $V = g^{-1}((-\infty, 2^{i_{min}}])$, and for $i = i_{min}, i_{min} + 1, \ldots, 0$, set $U_i = g^{-1}((2^i, 1])$. Note that V is convex, while each U_i, $i \geq 0$, is the complement of a convex subset of S.

We will use the following lemma:

Lemma 28 *For g and V, U_i as defined above, at least one of the following alternatives holds:*

1. $c_d \log(n)^{-d+1/2} \|g\|_2 \leq v^{-\frac{1}{2}} \|g\|_1 \leq \|g\|_2$.
2. *There exists $i_0 \in [i_{min}, 0]$ such that $2^{-i_0} \geq C_d (\log n)^{d-1} \frac{\|g\|_1}{\mathbb{P}(S)}$ and*

$$c \log(n)^{-1/2} \|g\|_2 \leq \mathbb{P}(U_{i_0})^{-1/2} \int_{U_{i_0}} g \, d\mathbb{P}. \tag{34}$$

The proof of this lemma appears at the end of this subsection.

If alternative (1) of the lemma holds, the L_1-norm of g is only a polylogarithmic factor away from the L_2-norm (up to normalizing by the measure of S, which is known to us). Therefore, we may use the L_1-estimator of the previous subsection and estimate the L_2 norm of g, up to a larger polylogarithmic factor, as we will see below.

We must therefore consider what happens when alternative (2) of Lemma 28 holds. If we could estimate the integral of g over U_{i_0}, we'd be done, but neither the index i_0 nor the set U_{i_0} are given to us. So we make use of the fact that each such U_{i_0}, being the complement of a convex subset of $\triangle$, is contained in the wet part $S(\mathbb{P}(U_i))$, which has volume at most $C_d \log(n)^{d-1}\mathbb{P}(U_i)$ by Lemma 27. We will show in the next lemma that this replacement costs us a $C_d \log(n)^{d/2}$ factor in the worst case.

More precisely, let $\epsilon_j = 2^{-2j}$, and let $S(\epsilon_j)$ be the corresponding wet part of S, as defined in Lemma 27. Then we have the following:

Lemma 29 *With g as above, we have*

$$\max_{j\in[i_{min},0]} \mathbb{P}(S(\epsilon_j))^{-1/2} \int_{S(\epsilon_j)} g d\mathbb{P} \le \|g\|_2,$$

and moreover, if alternative (2) of Lemma 27 holds, then there exists j such that $\mathbb{P}(S(\epsilon_j)) \ge \frac{Cd\log n}{n}$ and

$$c_d \log(n)^{-d/2}\|g\|_2 \le \mathbb{P}(S(\epsilon_j))^{-1/2} \int_{S(\epsilon_j)} g d\mathbb{P}.$$

Using the last lemma, we can construct an estimator for $\|g\|_2$ that is at most a polylogarithmic factor away from the true value, whether we are in case (1) or case (2) of Lemma 28.

Indeed, note that we if alternative (2) holds, we have $\mathbb{P}(S_{\epsilon_j}) \ge \frac{Cd\log n}{n}$ and hence, as in Section 2, we can assume by a union bound that the number of sample points falling in $S(\epsilon_j)$ is proportional to $\mathbb{P}(S(\epsilon_j))$. Hence, for each j the estimation of $\int_{S(\epsilon_j)} g\, d\mathbb{P}_{S(\epsilon_j)} = \mathbb{P}(S(\epsilon_j))^{-1}\int_{S(\epsilon_j)} g\, d\mathbb{P}$ can be done in a similar fashion as in §B.3.1, with an additive deviation that is proportional to $\sqrt{\frac{\log(2/\delta)}{n\mathbb{P}(S(\epsilon_j))}}$. However, since we need to estimate $\mathbb{P}(S(\epsilon_j))^{-\frac{1}{2}}\int_{S(\epsilon_j)} g\, d\mathbb{P}$, we can multiply it by $\sqrt{\mathbb{P}(S(\epsilon_j))}$, and obtain the correct deviation of $O(\sqrt{\log(2/\delta)/(\mathbb{P}(S)n)})$ (as usual, we work on the event $\mathbb{P}(S(\epsilon_j))/2 \le \mathbb{P}_n(S(\epsilon_j))$, which holds with probability at least $1 - n^{-2d}$).

We conclude that the maximum over $j \in [-c\log(n) \le i_{min}, 0]$, estimators of the means of the random variables $\mathbb{P}(S(\epsilon_j))^{\frac{1}{2}}\int_{S(\epsilon_j)} g\, d\mathbb{P}_{S(\epsilon\triangle_i)}$ and the L_1 estimator of the above sub-sub section give the claim.

Proof [Proof of Lemma 28] Let $g_- = \min(g,0)$, $g_+ = \max(g,0)$, so that $\|g\|_2^2 = \|g_-\|^2 + \|g_+\|^2$.

We claim that alternative (1) holds if $\|g_-\|_2^2 \ge \frac{1}{2}\|g\|_2^2$. Indeed, by Lemma 24, $g_- \ge -C_d v^{-1}\|g\|_1$, which immediately yields

$$\int_S g_-^2\, d\mathbb{P} \le -C_d v^{-1}\|g\|_1 \cdot \int g_-\, d\mathbb{P} \le -C_d v^{-1}\|g\|_1^2$$

i.e.,

$$v^{-1/2}\|g\|_1 \ge c_d\|g_-\|_2 \ge c_d'\|g\|_2.$$

Note that by Jensen's inequality we have that $v^{-1/2}\|g\|_1 \leq \|g\|_2$. Otherwise, we have $\|g_+\|^2 \geq \frac{1}{2}\|g\|_2^2$. Let $T_i = U_i \backslash U_{i+1} = g^{-1}((2^i, 2^{i+1}])$. We have

$$\frac{1}{2}\|g\|_2^2 \leq \|g_+\|_2^2 \leq \sum_{i=-\infty}^{0} 2^{2(i+2))}\mathbb{P}(T_i).$$

By our assumption $\|g\|_2^2 \geq C\frac{\log n}{n}$ and the fact that $v \leq 1$, the terms in the sum with $i \leq i_{min} = \log\left(C\frac{\log n}{n}\right) + 2$ cannot contribute more than half of the sum, so we have .

$$\frac{1}{4}\|g\|_2^2 \leq \|g_+\|_2^2 \leq \sum_{i=i_{min}}^{0} 2^{2(i+2))}\mathbb{P}(T_i).$$

Hence there exists $i_0 \in [i_{min}, 0]$ such that

$$\frac{\|g\|_2^2}{4\log n} \leq 2^{2(i_0+2))}\mathbb{P}(T_{i_0}) \leq 4\min_{x\in U_i} g(x)^2 \cdot \mathbb{P}(U_i) \leq 4\mathbb{P}(U_{i_0})^{-1}\left(\int_{U_i} g\, d\mathbb{P}\right)^2,$$

or

$$c\log(n)^{-\frac{1}{2}}\|g\|_2 \leq \mathbb{P}(U_{i_0})^{-\frac{1}{2}}\int_{U_{i_0}} g\, d\mathbb{P}.$$

We consider two cases: either $2^{-i} \geq C(\log n)^{d-1}v^{-1}\|g\|_1$, or $2^{-i} \leq C(\log n)^{d-1}v^{-1}\|g\|_1$. The first case leads immediately to alternative (2), while in the second case we have

$$c\log(n)^{-\frac{1}{2}}\|g\|_2 \leq \mathbb{P}(U_{i_0})^{-\frac{1}{2}}\int_{U_{i_0}} g\, d\mathbb{P} \leq 4C(\log n)^{d-1}v^{-1}\|g\|_1 \cdot \mathbb{P}(U_{i_0})^{\frac{1}{2}} \leq C(\log n)^{d-1}\|g\|_1 v^{-\frac{1}{2}},$$

which is another instance of alternative (1).

As for the right-hand inequality in alternative (1), this is simply Cauchy-Schwarz: $\left(\int |g|\, d\mathbb{P}\right)^2 \leq \int g^2\, d\mathbb{P} \cdot v$. ■

Proof [Proof of Lemma 29] The first inequality is again Cauchy-Schwarz: for any subset A of S, we have

$$\int_A g\, d\mathbb{P} \leq \left(\int_A g^2\, d\mathbb{P}\right)^{\frac{1}{2}}\left(\int_A 1\, d\mathbb{P}\right)^{\frac{1}{2}} \leq \|g\|_2 \cdot \mathbb{P}(A)^{\frac{1}{2}}.$$

As for the second statement, first note that since $\|g\|_\infty \leq 1$ and we have

$$c\log(n)^{-1/2}\|g\|_2 \leq \mathbb{P}(U_{i_0})^{-1/2}\int_{U_{i_0}} g\, d\mathbb{P} \leq \mathbb{P}(U_{i_0})^{\frac{1}{2}}.$$

Let $j = \lceil \log \mathbb{P}(U_{i_0})\rceil \geq i_{min}$, $\epsilon_j = 2^j$, so that $U_{i_0} \subset S(\epsilon_j)$ and

$$\mathbb{P}(S_{\epsilon_j}) \leq C\mathbb{P}(U_{i_0})\log(\mathbb{P}(U_{i_0})^{-1})^{d-1} \leq C\mathbb{P}(U_{i_0})(\log n)^{d-1}.$$

Recalling again that by (24), $g \geq -C\|g\|_1\mathbb{P}(S)^{-1}$, we have

$$\begin{aligned}\int_{S(\epsilon_j)} g\, d\mathbb{P} - 2^{-1}\int_{U_{i_0}} g\, d\mathbb{P} &\geq 2^{-1}\int_{U_{i_0}} g\, d\mathbb{P} + \int_{S(\epsilon_j)\setminus U_{i_0}} g\, d\mathbb{P} \\ &\geq \mathbb{P}(U_{i_0})2^{i_0} - \mathbb{P}(S(\epsilon_j)) \cdot C\|g\|_1\mathbb{P}(S)^{-1} \\ &\geq \mathbb{P}(U_{i_0})\left(2^{i_0} - C_d(\log n)^{d-1}\|g\|_1\mathbb{P}(S)^{-1}\right) \\ &\geq c \cdot \mathbb{P}(U_{i_0}) \cdot 2^{i_0} > 0,\end{aligned}$$

where we used our assumption on i_0 in the last line. Therefore, by the last two inequalities

$$\mathbb{P}(S(\epsilon_j))^{-\frac{1}{2}}\int_{S(\epsilon_j)} g\, d\mathbb{P} \geq c(\log n)^{-(d-1)/2}\mathbb{P}(U_{i_0})^{-\frac{1}{2}}\int_{U_{i_0}} g\, d\mathbb{P} \geq c(\log n)^{-d/2}\|g\|_2,$$

as claimed. Finally, note that by the assumptions of $\|g\|_2^2 \geq \frac{Cd(\log n)^2}{n}$ and $\|g\|_\infty \leq 1$, we obtain that

$$\mathbb{P}(S(\epsilon_j)) \geq \mathbb{P}(U_{i_0}) \geq (c\log(n)^{-\frac{1}{2}}\|g\|_2/\sqrt{\|g\|_\infty})^2 \geq C_1 d\frac{\log n}{n}.$$

■

Appendix C. Sketch of the Proof of Theorem 3

The modifications of Algorithm 1 to work in this setting are minimal: we simply need to replace L by Γ, and replace (14) with

$$\forall (i,j) \in [|\mathcal{S}|] \times [n]: \qquad |y_{i,j}| \leq \Gamma \tag{35}$$

$$\begin{aligned}&\forall\, 1 \leq i \leq |\mathcal{S}| \;\; \frac{1}{n}\sum_{j=1}^{n}(f(Z_{i,j}) - \hat{w}_i^\top(Z_{i,j},1))^2 \leq \hat{l}_i^2 + \Gamma\sqrt{\frac{Cd\log(n)}{n}} \\ &\forall (i,j) \in [|\mathcal{S}|] \times [n] \quad |f(Z_{i,j})| \leq \Gamma \\ &\forall (i_1,j_1),(i_2,j_2) \in [|\mathcal{S}|] \times [n] \quad f(Z_{i_2,j_2}) \geq \nabla f(Z_{i_1,j_1})^\top(Z_{i_2,j_2} - Z_{i_1,j_1}).\end{aligned}$$

For the correctness proof, we need some additional modifications. First, we replace Lemma 8 with a similar bound in the Γ-bounded setting. The following lemma is based on the L_4 entropy bound of (Gao and Wellner, 2017, Thm 1.1) and the peeling device (van de Geer, 2000, Ch. 5); it appears explicitly in (Han and Wellner, 2016)):

Lemma 30 *Let $d \geq 5$, $m \geq C^d$ and $\mathbb{Q}$ be a uniform measure on a convex polytope $P' \subset B_d$ and $Z_1, \dots, Z_m \sim \mathbb{Q}$. Then, the following holds uniformly for all $f, g \in \mathcal{F}^\Gamma(P')$*

$$2^{-1}\int_{P'}(f-g)^2 d\mathbb{Q} - C(P')\Gamma^2 m^{-\frac{4}{d}} \leq \int (f-g)^2 d\mathbb{Q}_m \leq 2\int_{P'}(f-g)^2 d\mathbb{Q} + C(P')\Gamma^2 m^{-\frac{4}{d}},$$

with probability at least $1 - C_1(P')\exp(-c_1(P')\sqrt{m})$.

Note that differently from Lemma 8, the constant before $m^{-4/d}$ depends on the domain P', and this dependence cannot be removed.

Since $\mathcal{F}^{\Gamma}(P')$ has finite L^2-entropy for every ϵ, it is in particular compact in $L^2(P')$, which means that the proof of Lemma 13 in sub-Section B.3 works for this class of functions as well.

The proof of Theorem 10 also goes through for this case, by replacing Lemma 8 by Lemma 30. The precise statement we obtain is the following:

Theorem 31 *Let $P \subset B_d$ be a convex polytope, $f \in \mathcal{F}^{\Gamma}(P)$, and some integer $k \geq (Cd)^{d/2}$, for some large enough $C \geq 0$, there exists a convex set $P_k \subset P$ and a k-simplicial convex function $f_k : P_k \to \mathbb{R}$ such that*

$$\mathbb{P}(P \setminus P_k) \leq C(P)k^{-\frac{d+2}{d}} \log(k)^{d-1}.$$

and

$$\int_{P_k} (f_k - f)^2 d\mathbb{P} \leq \Gamma^2 \cdot C(P)k^{-\frac{4}{d}}$$

The remaining lemmas and arguments in the proof of Theorem 1, can easily be seen to apply in the setting of Γ-bounded regression under polytopal support P.

Appendix D. Simplified version of our estimator

Like the estimator for our original problem, the simplified version of our estimator is based on the existence of a simplicial approximation $\hat{f}_{k(n)} : \Omega_{k(n)} \to [0,1]$ to the unknown convex function f^* (Theorem 1). Here we demonstrate how to recover f^* to within the desired accuracy if we are given the simplicial structure of $f_{k(n)}$, i.e., the set $\Omega_{k(n)}$ and the decomposition $\bigcup_{i=1}^{k(n)} \triangle_i$ of $\Omega_{k(n)}$ into simplices such that $f_{k(n)}|_{\triangle_i}$ is affine for each i. In this case the performance of our algorithm is rather better: it runs in time $O_d(n^{O(1)})$ rather than $O_d(n^{O(d)})$, and is minimax optimal up to a constant that depends on d. We can also slightly weaken the assumptions: it is no longer required that the variance σ^2 of the noise be given.

We will use the following classical estimator (Györfi et al., 2002, Thm 11.3); it is quoted here with an improved bound which is proven in (Mourtada et al., 2021, Theorem A):

Lemma 32 *Let $m \geq d+1$, $d \geq 1$ and $\mathbb{Q}$ be a probability measure that is supported on some $\Omega' \subset \mathbb{R}^d$. Consider the regression model $W = f^*(Z) + \xi$, where f^* is L-Lipschitz and $\|f^*\|_\infty \leq L$, and $Z_1, \ldots, Z_m \underset{i.i.d.}{\sim} \mathbb{Q}$. Then, the exists an estimator $\hat{f}_R$ that has an input of $\{(Z_i, W_i)\}_{i=1}^m$ and runtime of $O_d(n)$ and outputs a function such that*

$$\mathbb{E} \int (\hat{f}_R(x) - f^*(x))^2 d\mathbb{Q}(x) \leq \frac{Cd(\sigma + L)^2}{m} + \inf_{w \in \mathbb{R}^{d+1}} \int (w^\top(x,1) - f^*(x))^2 d\mathbb{Q}(x).$$

Note that this estimator is *distribution-free*: it works irrespective of the structure of $\mathbb{Q}$, nor does it require that $\mathbb{Q}$ be known.

The first step of the simplified algorithm is estimating $f^*|_{\triangle_i}$ on each $\triangle_i \subset \Omega_{k(n)}$ $(1 \le i \le k(n))$ with the estimator $\hat{f}_R$ defined in Lemma 32 (with respect to the probability measure $\mathbb{P}(\cdot|\triangle_i)$) with the input of the data points in $\mathcal{D}$ that lie in $\triangle_i$. We obtain independent regressors $\hat{f}_1, \dots, \hat{f}_{k(n)}$ such that

$$\mathbb{E}\int_{\triangle_i} (\hat{f}_i(x) - f^*(x))^2 \frac{d\mathbb{P}}{\mathbb{P}(\triangle_i)} \le \inf_{w\in\mathbb{R}^{d+1}} \int_{\triangle_i} (w^\top(x,1) - f^*(x))^2 \frac{d\mathbb{P}}{\mathbb{P}(\triangle_i)} + \mathbb{E}\min\{\frac{Cd}{\mathbb{P}_n(\triangle_i)n}, 1\}, \tag{36}$$

where the $\min\{\cdot, 1\}$ part follows from the fact that when we have less than Cd points, we can always set $\hat{f}_i$ to be the zero function.

Now, we define the function $f'(x) := \sum_{i=1}^{k(n)} \hat{f}_i(x)\mathbb{1}_{x\in\triangle_i}$, and by multiplying the last equation by $\mathbb{P}(\triangle_i)$ for each $1 \le i \le k(n)$ and taking a sum over i, we obtain that

$$\begin{aligned}\mathbb{E}\int_{\Omega_{k(n)}} (f' - f^*)^2 d\mathbb{P} &\le \sum_{i=1}^{k(n)} \inf_{w_i\in\mathbb{R}^{d+1}} \int_{\triangle_i} (w_i^\top(x,1) - f^*)^2 d\mathbb{P} + \mathbb{E}\sum_{i=1}^{k(n)} \min\{\frac{Cd\cdot\mathbb{P}(\triangle_i)}{n\cdot\mathbb{P}_n(\triangle_i)}, \mathbb{P}(\triangle_i)\}\\ &\le \int_{\Omega_{k(n)}} (f_{k(n)} - f^*)^2 d\mathbb{P} + C_1 d k(n)\cdot n^{-1} = O_d(n^{-\frac{4}{d+4}}),\end{aligned} \tag{37}$$

where in the first equation, we used the the fact that $n\cdot\mathbb{P}_n(\triangle_i) \sim Bin(n, \mathbb{P}(\triangle_i))$ (for completeness, see Lemma 20), and in the last inequality we used Eq. (6). Next, recall that Theorem 10 implies that

$$\mathbb{P}(\Omega \setminus \Omega_{k(n)}) \le C(d)k(n)^{-\frac{d+2}{d}} \le O_d(n^{-\frac{d+2}{d+4}}).$$

Therefore, if we consider the (not necessarily convex) function $\tilde{f} = f'\mathbb{1}_{\Omega_{k(n)}} + \mathbb{1}_{\Omega\setminus\Omega_{k(n)}}$, we obtain that

$$\begin{aligned}\mathbb{E}\int_{\Omega} (f' - f^*)^2 d\mathbb{P} &= \mathbb{E}\int_{\Omega\setminus\Omega_{k(n)}} (f' - f^*)^2 d\mathbb{P} + \mathbb{E}\int_{\Omega_{k(n)}} (f' - f^*)^2 d\mathbb{P} \le O_d(n^{-\frac{d+2}{d+4}} + n^{-\frac{4}{d+4}})\\ &\le O_d(n^{-\frac{4}{d+4}}).\end{aligned}$$

Thus, $\tilde{f}$ is a minimax optimal *improper* estimator. To obtain a proper estimator, we simply need to replace $\tilde{f}$ by $MP(\tilde{f})$, where MP is the procedure defined in Appendix E.

It remains only to point out that the runtime of this estimator is of order $O_d(n^{O(1)})$. Indeed, the procedure MP is essentially a convex LSE on n points, which can be formulated as a quadratic programming problem with $O(n^2)$ constraints, and hence can be computed in $O_d(n^{O(1)})$ time (Seijo and Sen, 2011). In addition, the runtime of the other estimator we use, namely the estimator of Lemma 32, is linear in the number of inputs.

Appendix E. From an Improper to a Proper Estimator

The following procedure, which we named MP, is classical and we give its description and prove its correctness here for completeness. However, note that we only give a proof for optimality in expectation; high-probability bounds can be obtained using standard concentration inequalities.

The procedure MP is defined as follows: given an improper estimator $\tilde{f}$, draw $X_1', \dots, X_{k(n)}' \underset{i.i.d.}{\sim} \mathbb{P}$, and apply the convex LSE with the input $\{(X_i', \tilde{f}(X_i'))\}_{i=1}^{k(n)}$, yielding a function $\hat{f}_1$. We remark that the convex LSE is only unique on the convex hull of the data-points $X_1', \dots, X_{k(n)}'$, and not on the entire domain Ω (Seijo and Sen, 2011), so we will show that any solution $\hat{f}_1$ of the convex LSE is optimal.

First off, we have

$$\mathbb{E}\int_\Omega (\tilde{f} - f^*)^2 d\mathbb{P}_{k(n)}' = \mathbb{E}\int_\Omega (\tilde{f} - f^*)^2 d\mathbb{P} \tag{38}$$

Also recall the classical observation that for $\hat{f}_1$ that is defined above, we know that $(\hat{f}_1(X_1'), \dots, \hat{f}_1(X_{k(n)}'))$ is precisely the projection of $(\tilde{f}(X_1'), \dots, \tilde{f}(X_{k(n)}'))$ on the *convex set*

$$\mathcal{F}_{k(n)} := \{(f(X_1'), \dots, f(X_{k(n)}')) : f \in \mathcal{F}_1(\Omega)\} \subset \mathbb{R}^{k(n)},$$

cf. (Chatterjee, 2014). Now, the function $\Pi_{\mathcal{F}_{k(n)}}$ sending a point to its projection onto $\mathcal{F}_{k(n)}$, like any projection to a convex set, is a 1-Lipschitz function, i.e.,

$$\|\Pi_{\mathcal{F}_{k(n)}}(x) - \Pi_{\mathcal{F}_{k(n)}}(y)\| \leq \|x - y\| \quad \forall x, y \in \mathbb{R}^{k(n)}.$$

We also know that $(\hat{f}_1(X_i'))_{i=1}^{k(n)} = \Pi_{\mathcal{F}_{k(n)}}(\tilde{f})$ and $\Pi_{\mathcal{F}_{k(n)}}((f^*(X_i'))_{i=1}^{k(n)}) = (f^*(X_i'))_{i=1}^{k(n)}$; substituting in the preceding equation, we therefore obtain

$$\mathbb{E}\int_\Omega (\hat{f}_1 - f^*)^2 d\mathbb{P}_{k(n)}' \leq \mathbb{E}\int_\Omega (\tilde{f} - f^*)^2 d\mathbb{P}_{k(n)}' = \mathbb{E}\int_\Omega (\tilde{f} - f^*)^2 d\mathbb{P},$$

since $\int (\cdot)^2 d\mathbb{P}_{k(n)}'$ is just $\|\cdot\|^2/k(n)$. In order to conclude the minimax optimality of $\hat{f}_1$, we know by Lemma 8 that for any function in

$$\mathcal{O} := \left\{ f \in \mathcal{F}_1 : \int_\Omega (f - \tilde{f}')^2 d\mathbb{P}_{k(n)}' = 0 \right\},$$

it holds that

$$\mathbb{E}\int (f - f^*)^2 d\mathbb{P} \leq 2\mathbb{E}\int (\tilde{f} - f^*)^2 d\mathbb{P}_{k(n)}' + Ck(n)^{-\frac{4}{d}} \leq 2\mathbb{E}\int (\tilde{f} - f^*)^2 d\mathbb{P} + C_1 n^{-\frac{4}{d+4}},$$

where we used Eq. (38) and the fact that $k(n) = n^{\frac{d}{d+4}}$. Since we showed that $\hat{f}_1$ must lie in $\mathcal{O}$, the minimax optimality of this proper estimator follows.