Submitted to the Annals of Statistics
arXiv: 1411.1805

FAITHFUL VARIABLE SCREENING FOR HIGH-DIMENSIONAL CONVEX REGRESSION

By Min Xu*, Minhua Chen† and John Lafferty†

We study the problem of variable selection in convex nonparametric regression. Under the assumption that the true regression function is convex and sparse, we develop a screening procedure to select a subset of variables that contains the relevant variables. Our approach is a two-stage quadratic programming method that estimates a sum of one-dimensional convex functions, followed by one-dimensional concave regression fits on the residuals. In contrast to previous methods for sparse additive models, the optimization is finite dimensional and requires no tuning parameters for smoothness. Under appropriate assumptions, we prove that the procedure is faithful in the population setting, yielding no false negatives. We give a finite sample statistical analysis, and introduce algorithms for efficiently carrying out the required quadratic programs. The approach leads to computational and statistical advantages over fitting a full model, and provides an effective, practical approach to variable screening in convex regression.

1. Introduction. Shape restrictions such as monotonicity, convexity, and concavity provide a natural way of limiting the complexity of many statistical estimation problems. Shape-constrained estimation is not as well understood as more traditional nonparametric estimation involving smoothness constraints. For instance, the minimax rate of convergence for multivariate convex regression has yet to be rigorously established in full generality. Even the one-dimensional case is challenging, and has been of recent interest (Guntuboyina and Sen, 2013).

In this paper we study the problem of variable selection in multivariate convex regression. Assuming that the regression function is convex and sparse, our goal is to identify the relevant variables. We show that it suffices to estimate a sum of one-dimensional convex functions, leading to significant computational and statistical advantages. This is in contrast to general nonparametric regression, where fitting an additive model can result in false negatives. Our approach is based on a two-stage quadratic programming procedure. In the first stage, we fit an convex additive

*Machine Learning Department, Carnegie Mellon University, Pittsburgh PA 15213
†Department of Statistics, University of Chicago, Chicago IL 60637
Keywords and phrases: nonparametric regression, convex regression, variable selection, quadratic programming, additive model

model, imposing a sparsity penalty. In the second stage, we fit a concave function on the residual for each variable. As we show, this non-intuitive second stage is in general necessary. Our first result is that this procedure is faithful in the population setting, meaning that it results in no false negatives, under mild assumptions on the density of the covariates. Our second result is a finite sample statistical analysis of the procedure, where we upper bound the statistical rate of variable screening consistency. An additional contribution is to show how the required quadratic programs can be formulated to be more scalable. We give simulations to illustrate our method, showing that it performs in a manner that is consistent with our analysis.

Estimation of convex functions arises naturally in several applications. Examples include geometric programming (Boyd and Vandenberghe, 2004), computed tomography (Prince and Willsky, 1990), target reconstruction (Lele, Kulkarni and Willsky, 1992), image analysis (Goldenshluger and Zeevi, 2006) and circuit design (Hannah and Dunson, 2012). Other applications include queuing theory (Chen and Yao, 2001) and economics, where it is of interest to estimate concave utility functions (Meyer and Pratt, 1968). See Lim and Glynn (2012) for other applications. Beyond cases where the assumption of convexity is natural, the convexity assumption can be attractive as a tractable, nonparamametric relaxation of the linear model.

Recently, there has been increased research activity on shape-constrained estimation. Guntuboyina and Sen (2013) analyze univariate convex regression and show surprisingly that the risk of the MLE is adaptive to the complexity of the true function. Seijo and Sen (2011) and Lim and Glynn (2012) study maximum likelihood estimation of multivariate convex regression and independently establish its consistency. Cule, Samworth and Stewart (2010) and Kim and Samworth (2014) analyze log-concave density estimation and prove consistency of the MLE; the latter further show that log-concave density estimation has minimax risk lower bounded by $n^{-2/(d+1)}$ for $d \geq 2$, refuting a common notion that the condition of convexity is equivalent, in estimation difficulty, to the condition of having two bounded derivatives. Additive shape-constrained estimation has also been studied; Pya and Wood (2014) propose a penalized B-spline estimator while Chen and Samworth (2014) show the consistency of the MLE. To the best of our knowledge however, there has been no work on variable selection and and estimation of high-dimensional convex functions.

Variable selection in general nonparametric regression or function estimation is a notoriously difficult problem. Lafferty and Wasserman (2008) develop a greedy procedure for adjusting bandwidths in a local linear regression estimator, and show that the procedure achieves the minimax rate as if the relevant variables were isolated in advance. But the method only provably scales to dimensions p that grow logarithmically in the sample size n, i.e., $p = O(\log n)$. This is in contrast to the high dimensional scaling behavior known to hold for sparsity selection in linear models using ℓ_1 penalization, where n is logarithmic in the dimension p. Bertin and Lecué (2008) develop an optimization-based approach in the nonparametric setting, apply-

ing the lasso in a local linear model at each test point. Here again, however, the method only scales as $p = O(\log n)$, the low-dimensional regime. An approximation theory approach to the same problem is presented in DeVore, Petrova and Wojtaszczyk (2011), using techniques based on hierarchical hashing schemes, similar to those used for "junta" problems (Mossel, O'Donnell and Servedio, 2004). Here it is shown that the sample complexity scales as $n > \log p$ if one adaptively selects the points on which the high-dimensional function is evaluated.

Comminges and Dalalyan (2012) show that the exponential scaling $n = O(\log p)$ is achievable if the underlying function is assumed to be smooth with respect to a Fourier basis. They also give support for the intrinsic difficulty of variable selection in nonparametric regression, giving lower bounds showing that consistent variable selection is not possible if $n < \log p$ or if $n < \exp s$, where s is the number of relevant variables. Variable selection over kernel classes is studied by Koltchinskii and Yuan (2010).

Perhaps more closely related to the present work is the framework studied by Raskutti, Wainwright and Yu (2012) for sparse additive models, where sparse regression is considered under an additive assumption, with each component function belonging to an RKHS. An advantage of working over an RKHS is that nonparametric regression with a sparsity-inducing regularization penalty can be formulated as a finite dimensional convex cone optimization. On the other hand, smoothing parameters for the component Hilbert spaces must be chosen, leading to extra tuning parameters that are difficult to select in practice. There has also been work on estimating sparse additive models over a spline basis, for instance the work of Huang, Horowitz and Wei (2010), but these approaches too require the tuning of smoothing parameters.

While nonparametric, the convex regression problem is naturally formulated using finite dimensional convex optimization, with no additional tuning parameters. The convex additive model can be used for convenience, without assuming it to actually hold, for the purpose of variable selection. As we show, our method scales to high dimensions, with a dependence on the intrinsic dimension s that scales polynomially, rather than exponentially as in the general case analyzed in Comminges and Dalalyan (2012).

In the following section we give a high-level summary of our technical results, including additive faithfulness, variable selection consistency, and high dimensional scaling. In Section 3 we give a detailed account of our method and the conditions under which we can guarantee consistent variable selection. In Section 4 we show how the required quadratic programs can be reformulated to be more efficient and scalable. In Section 5 we give the details of our finite sample analysis, showing that a sample size growing as $n = O\big(\mathrm{poly}(s) \log p\big)$ is sufficient for variable selection. In Section 6 we report the results of simulations that illustrate our methods and theory. The full proofs are given in a technical appendix.

2. Overview of Results. In this section we provide a high-level description of our technical results. The full technical details, the precise statement of the results, and their detailed proofs are provided in following sections.

Our main contribution is an analysis of an additive approximation for identifying relevant variables in convex regression. We prove a result that shows when and how the additive approximation can be used without introducing false negatives in the population setting. In addition, we develop algorithms for the efficient implementation of the quadratic programs required by the procedure.

We first establish some notation, to be used throughout the paper. If $\boldsymbol{x}$ is a vector, we use $\boldsymbol{x}_{-k}$ to denote the vector with the k-th coordinate removed. If $\boldsymbol{v} \in \mathbb{R}^n$, then $v_{(1)}$ denotes the smallest coordinate of $\boldsymbol{v}$ in magnitude, and $v_{(j)}$ denotes the j-th smallest; $\mathbf{1}_n \in \mathbb{R}^n$ is the all ones vector. If $X \in \mathbb{R}^p$ is a random variable and $S \subset \{1, ..., p\}$, then X_S is the subvector of X restricted to the coordinates in S. Given n samples $X^{(1)}, ..., X^{(n)}$, we use $\overline{X}$ to denote the sample mean. Given a random variable X_k and a scalar x_k, we use $\mathbb{E}[\,\cdot \,|\, x_k]$ as a shorthand for $\mathbb{E}[\,\cdot \,|\, X_k = x_k]$.

2.1. *Faithful screening.* The starting point for our approach is the observation that least squares nonparametric estimation under convexity constraints is equivalent to a finite dimensional quadratic program. Specifically, the infinite dimensional optimization

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^{n} (Y_i - f(\boldsymbol{x}_i))^2 \\ &\text{subject to} && f : \mathbb{R}^p \to \mathbb{R} \text{ is convex} \end{aligned} \tag{2.1}$$

is equivalent to the finite dimensional quadratic program

$$\begin{aligned} &\text{minimize}_{f,\beta} && \sum_{i=1}^{n} (Y_i - f_i)^2 \\ &\text{subject to} && f_j \geq f_i + \beta_i^T (\boldsymbol{x}_j - \boldsymbol{x}_i), \text{ for all } i, j. \end{aligned} \tag{2.2}$$

Here f_i is the estimated function value $f(\boldsymbol{x}_i)$, and the vectors $\beta_i \in \mathbb{R}^d$ represent supporting hyperplanes to the epigraph of f. See Boyd and Vandenberghe (2004), Section 6.5.5. Importantly, this finite dimensional quadratic program does not have tuning parameters for smoothing the function.

This formulation of convex regression is subject to the curse of dimensionality. Moreover, attempting to select variables by regularizing the subgradient vectors β_i with a group sparsity penality is not effective. Intuitively, the reason is that all p components of the subgradient β_i appear in every convexity constraint $f_j \geq f_i + \beta_i^T (\boldsymbol{x}_j - \boldsymbol{x}_i)$; small changes to the subgradients may not violate the constraints. Experimentally, we find that regularization with a group sparsity penality will make the subgradients of irrelevant variables small, but may not zero them out completely.

This motivates us to consider an additive approximation. Under a convex additive model, each component of the subgradient appears only in the convexity constraint for the corresponding variable:

$$f_{ki'} \geq f_{ki} + \beta_{ki}(x_{ki'} - x_{ki}) \tag{2.3}$$

where $f_{ki} = f_k(x_{ki})$ and β_{ki} is the subgradient at point x_{ki}. As we show, this leads to an effective variable selection procedure. The shape constraints play an essential role. For general regression, using an additive approximation for variable selection may make errors. In particular, the nonlinearities in the regression function may result in an additive component being wrongly zeroed out. We show that this cannot happen for convex regression under appropriate conditions.

We say that a differentiable function f depends on variable x_k if $\partial_{x_k} f \neq 0$ with probability greater than zero. An additive approximation is given by

$$\{f_k^*\}, \mu^* := \operatorname*{arg\,min}_{f_1,\ldots,f_p,\mu} \Big\{ \mathbb{E}\Big(f(X) - \mu - \sum_{k=1}^{p} f_k(X_k) \Big)^2 \,:\, \mathbb{E} f_k(X_k) = 0 \Big\}. \tag{2.4}$$

We say that f is *additively faithful* in case $f_k^* = 0$ implies that f does not depend on coordinate k. Additive faithfulness is a desirable property since it implies that an additive approximation may allow us to screen out irrelevant variables.

Our first result shows that convex multivariate functions are additively faithful under the following assumption on the distribution of the data.

DEFINITION 2.1. Let $p(\boldsymbol{x})$ be a density supported on $[0,1]^p$. Then p satisfies the *boundary flatness condition* if for all j, and for all $\boldsymbol{x}_{-j}$,

$$\frac{\partial p(\boldsymbol{x}_{-j} \,|\, x_j)}{\partial x_j} = \frac{\partial^2 p(\boldsymbol{x}_{-j} \,|\, x_j)}{\partial x_j^2} = 0 \quad \text{at } x_j = 0 \text{ and } x_j = 1.$$

As discussed in Section 3, this is a relatively weak condition. Our first result is that this condition suffices in the population setting of convex regression.

THEOREM 1. *Let p be a positive density supported on $C = [0,1]^p$ that satisfies the boundary flatness property. If f is convex and twice differentiable, then f is additively faithful under p.*

Intuitively, an additive approximation zeroes out variable k when, fixing x_k, every "slice" of f integrates to zero. We prove this result by showing that "slices" of convex functions that integrate to zero cannot be "glued together" while still maintaining convexity.

While this shows that convex functions are additively faithful, it is difficult to estimate the optimal additive functions. The difficulty is that f_k^* need not be a convex function, as we show through a counterexample in Section 3. It may be possible to

estimate f_k^* with smoothing parameters, but, for the purpose of variable screening, it is sufficient in fact to approximate f_k^* by a *convex* additive model.

Our next result states that a convex additive fit, combined with a series of univariate concave fits, is faithful. We abuse notation in Theorem 2 and let the notation f_k^* represent convex additive components.

THEOREM 2. *Suppose $p(\boldsymbol{x})$ is a positive density on $C = [0,1]^p$ that satisfies the boundary flatness condition. Suppose that f is convex and twice-differentiable. and that $\partial_{x_k} f$, $\partial_{x_k} p(\boldsymbol{x}_{-k} \,|\, x_k)$, and $\partial_{x_k}^2 p(\boldsymbol{x}_{-k} \,|\, x_k)$ are all continuous as functions on C. Define*

$$\{f_k^*\}_{k=1}^p, \mu^* = \arg\min_{\{f_k\},\mu} \Big\{ \mathbb{E}\Big(f(X) - \mu - \sum_{k=1}^s f_k(X_k) \Big)^2 \,:\, f_k \in \mathcal{C}^1,\, \mathbb{E} f_k(X_k) = 0 \Big\} \tag{2.5}$$

where $\mathcal{C}^1$ is the set of univariate convex functions, and, with respective to f_k^*'s from above, define

$$g_k^* = \arg\min_{g_k} \Big\{ \mathbb{E}\Big(f(X) - \mu^* - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - g_k \Big)^2 \,:\, g_k \in \text{-}\mathcal{C}^1, \mathbb{E} g_k(X_k) = 0 \Big\}, \tag{2.6}$$

with -$\mathcal{C}^1$ denoting the set of univariate concave functions. Then $f_k^* = 0$ and $g_k^* = 0$ implies that f does not depend on x_k, i.e., $\partial_{x_k} f(\boldsymbol{x}) = 0$ with probability one.

This result naturally suggests a two-stage screening procedure for variable selection. In the first stage we fit a sparse convex additive model $\{\widehat{f}_k\}$. In the second stage we fit a concave function $\widehat{g}_k$ to the residual for each variable having a zero convex component $\widehat{f}_k$. If both $\widehat{f}_k = 0$ and $\widehat{g}_k = 0$, we can safely discard variable x_k. As a shorthand, we refer to this two-stage procedure as AC/DC. In the AC stage we fit an additive convex model. In the DC stage we fit decoupled concave functions on the residuals. The decoupled nature of the DC stage allows all of the fits to be carried out in parallel. The entire process involves no smoothing parameters. Our next result concerns the required optimizations, and their finite sample statistical performance.

2.2. *Optimization.* In Section 4 we present optimization algorithms for the additive convex regression stage. The convex constraints for the additive functions, analogous to the multivariate constraints (2.2), are that each component $f_k(\cdot)$ can be represented by its supporting hyperplanes, i.e.,

$$f_{ki'} \geq f_{ki} + \beta_{ki}(x_{ki'} - x_{ki}) \quad \text{for all } i, i' \tag{2.7}$$

where $f_{ki} := f_k(x_{ki})$ and β_{ki} is the subgradient at point x_{ki}. While this apparently requires $O(n^2 p)$ equations to impose the supporting hyperplane constraints, in fact, only $O(np)$ constraints suffice. This is because univariate convex functions are characterized by the condition that the subgradient, which is a scalar, must increase

monotonically. This observation leads to a reduced quadratic program with $O(np)$ variables and $O(np)$ constraints.

Directly applying a QP solver to this optimization is still computationally expensive for relatively large n and p. We thus develop a block coordinate descent method, where in each step we solve a sparse quadratic program involving $O(n)$ variables and $O(n)$ constraints. This is efficiently solved using optimization packages such as MOSEK. The details of these optimizations are given in Section 4.

2.3. *Finite sample analysis.* In Section 5 we analyze the finite sample variable selection consistency of convex additive modeling, without assuming that the true regression function f_0 is additive. Our analysis first establishes a sufficient deterministic condition for variable selection consistency, and then considers a stochastic setting. Our proof technique decomposes the KKT conditions for the optimization in a manner that is similar to the now standard *primal-dual witness* method (Wainwright, 2009).

We prove separate results that allow us to analyze false negative rates and false positive rates. To control false positives, we analyze scaling conditions on the regularization parameter λ_n for group sparsity needed to zero out irrelevant variables $k \in S^c$, where $S \subset \{1, \ldots, p\}$ is the set of variables selected by the AC/DC algorithm in the population setting. To control false negatives, we analyze the restricted regression where the variables in S^c are zeroed out, following the primal-dual strategy.

Each of our theorems uses a subset of the following assumptions:

A1: X_S, X_{S^c} are independent.

A2: f_0 is convex and twice-differentiable. $\mathbb{E} f_0(X) = 0$.

A3: $\|f_0\|_\infty \le sB$ and $\|f_k^*\| \le B$ for all k.

A4: The noise is mean-zero sub-Gaussian, independent of X.

A5: The density $p(\boldsymbol{x})$ is bounded away from 0 and satisfies the boundary flatness condition.

In Assumption A3, $f^* = \sum_k f_k^*$ denotes the optimal additive projection of f_0 in the population setting.

Our analysis involves parameters α_+ and α_-, which are measures of the signal strength of the weakest variable:

$$\alpha_+ = \inf_{f \in \mathcal{C}^p \,:\, \mathrm{supp}(f) \subsetneq \mathrm{supp}(f^*)} \Big\{ \mathbb{E}\big(f_0(X) - f(X)\big)^2 - \mathbb{E}\big(f_0(X) - f^*(X)\big)^2 \Big\}$$

$$\alpha_- = \min_{k \in S \,:\, g_k^* \neq 0} \Big\{ \mathbb{E}\big(f_0(X) - f^*(X)\big)^2 - \mathbb{E}\big(f_0(X) - f^*(X) - g_k^*(X_k)\big)^2 \Big\}.$$

Intuitively, if α_+ is small, then it is easier to make a false omission in the additive convex stage of the procedure. If α_- is small, then it is easier to make a false omission in the decoupled concave stage of the procedure.

We make strong assumptions on the covariates in A1 in order to make very weak assumptions on the true regression function f_0 in A2; in particular, we do not assume that f_0 is additive. Relaxing this condition is an important direction for future work. We also include an extra boundedness constraint to use new bracketing number results (Kim and Samworth, 2014).

Our main result is the following.

THEOREM 3. *Suppose assumptions A1-A5 hold. Let $\{\widehat{f}_i\}$ be any AC solution and let $\{\widehat{g}_k\}$ be any DC solution, both estimated with regularization parameter λ scaling as $\lambda = \Theta\Big(sB\sqrt{\frac{1}{n}\log^2 np}\Big)$. Suppose in addition that*

$$\alpha_f/\widetilde{\sigma} \geq cB^2\sqrt{\frac{s^5}{n^{4/5}}\log^2 np} \tag{2.8}$$

$$\alpha_g^2/\widetilde{\sigma} \geq cB^4\sqrt{\frac{s^5}{n^{4/5}}\log^2 2np}. \tag{2.9}$$

where $\widetilde{\sigma} \equiv \max(\sigma, B)$ and c is a constant dependent only on b, c_1.

Then, for sufficiently large n, with probability at least $1 - \frac{1}{n}$:

$$\widehat{f}_k \neq 0 \text{ or } \widehat{g}_k \neq 0 \text{ for all } k \in S$$
$$\widehat{f}_k = 0 \text{ and } \widehat{g}_k = 0 \text{ for all } k \notin S.$$

This shows that variable selection consistency is achievable under exponential scaling of the ambient dimension, $p = O(\exp(n^c))$ for some $0 < c < 1$, as for linear models. The cost of nonparametric estimation is reflected in the scaling with respect to $s = |S|$, which can grow only as $o(n^{4/25})$.

We remark that Comminges and Dalalyan (2012) show that, even with the product distribution, under traditional smoothness constraints, variable selection is achievable only if $n > O(e^s)$. Here we demonstrate that convexity yields the scaling $n = O(\text{poly}(s))$.

3. Additive Faithfulness. For a general regression function, an additive approximation may result in a relevant variable being incorrectly marked as irrelevant. Such mistakes are inherent to the approximation and may persist even in the population setting. In this section we give examples of this phenomenon, and then show how the convexity assumption changes the behavior of the additive approximation. We work with $C = [0, 1]^p$ as the support of the distribution in this section but all of our results apply to general hypercubes. We begin with a lemma that characterizes the components of the additive approximation under mild conditions.

LEMMA 3.1. *Let F be a distribution on $C = [0,1]^p$ with a positive density function p. Let $f : C \to \mathbb{R}$ be an integrable function, and define*

$$f_1^*, ..., f_p^*, \mu^* := \arg\min \left\{ \mathbb{E}\Big(f(X) - \mu - \sum_{k=1}^{p} f_k(X_k) \Big)^2 \; : \; \mathbb{E} f_k(X_k) = 0, \; \forall k = 1, \dots, p \right\}.$$

Then $\mu^* = \mathbb{E} f(X)$,

$$f_k^*(x_k) = \mathbb{E}\Big[f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) \,|\, x_k \Big] - \mathbb{E} f(X), \tag{3.1}$$

and this solution is unique.

Lemma 3.1 follows from the stationarity conditions of the optimal solution. This result is known, and criterion (3.1) is used in the backfitting algorithm for fitting additive models. We include a proof as our results build on it.

PROOF. Let $f_1^*, ..., f_p^*, \mu^*$ be the minimizers as defined. We first show that the optimal μ is $\mu^* = \mathbb{E} f(X)$ for any $f_1, ..., f_k$ such that $\mathbb{E} f_k(X_k) = 0$. This follows from the stationarity condition, which states that $\mu^* = \mathbb{E}[f(X) - \sum_k f_k(X_k)] = \mathbb{E}[f(X)]$. Uniqueness is apparent because the second derivative is strictly larger than zero and strong convexity is guaranteed.

We now turn our attention toward the f_k^*s. It must be that f_k^* minimizes

$$\mathbb{E}\Big[\big(f(X) - \mu^* - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - f_k(X_k) \big)^2 \Big] \tag{3.2}$$

subject to $\mathbb{E} f_k(X_k) = 0$. Fixing x_k, we will show that the value

$$\mathbb{E}[f(X) - \sum_{k' \neq k} f_{k'}(X_{k'}) \,|\, x_k] - \mu^* \tag{3.3}$$

uniquely minimizes

$$\min_{f_k(x_k)} \int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}) \Big(f(\boldsymbol{x}) - \sum_{k' \neq k} f_{k'}^*(x_{k'}) - f_k(x_k) - \mu^* \Big)^2 d\boldsymbol{x}_{-k}. \tag{3.4}$$

The first-order optimality condition gives us

$$\int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}) f_k(x_k) d\boldsymbol{x}_{-k} = \int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}) (f(\boldsymbol{x}) - \sum_{k' \neq k} f_{k'}^*(x_{k'}) - \mu^*) d\boldsymbol{x}_{-k} \tag{3.5}$$

$$p(x_k) f_k(x_k) = \int_{\boldsymbol{x}_{-k}} p(x_k) p(\boldsymbol{x}_{-k} \,|\, x_k) (f(\boldsymbol{x}) - \sum_{k' \neq k} f_{k'}^*(x_{k'}) - \mu^*) d\boldsymbol{x}_{-k} \tag{3.6}$$

$$f_k(x_k) = \int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}_{-k} \,|\, x_k) (f(\boldsymbol{x}) - \sum_{k' \neq k} f_{k'}(x_{k'}) - \mu^*) d\boldsymbol{x}_{-k} \tag{3.7}$$

The square error objective is strongly convex, and the second derivative with respect to $f_k(x_k)$ is $2p(x_k)$, which is always positive under the assumption that p is positive. Therefore, the solution $f_k^*(x_k) = \mathbb{E}[f(X) \,|\, x_k] - \mathbb{E}f(X)$ is unique. Noting that $\mathbb{E}[f(X) - \sum_{k' \neq k} f_{k'}(X_{k'})|x_k] - \mathbb{E}f(X)$ has mean zero as a function of x_k completes the proof. □

In the case that the distribution in Lemma 3.1 is a product distribution, the additive components take on a simple form.

COROLLARY 3.1. *Let F be a product distribution on $C = [0,1]^p$ with density function p which is positive on C. Let $\mu^*, f_k^*(x_k)$ be defined as in Lemma 3.1. Then $\mu^* = \mathbb{E}f(X)$ and $f_k^*(x_k) = \mathbb{E}[f(X) \,|\, x_k] - \mathbb{E}f(X)$ and this solution is unique.*

In particular, if F is the uniform distribution, then $f_k^*(x_k) = \int f(x_k, \boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k}$.

EXAMPLE 3.1. Using Corollary 3.1, we give two examples of *additively unfaithfulness* under the uniform distribution—where relevant variables are erroneously marked as irrelevant under an additive approximation. First, consider the following function:

$$f(x_1, x_2) = \sin(2\pi x_1)\sin(2\pi x_2) \quad \text{(egg carton)} \tag{3.8}$$

defined for $(x_1, x_2) \in [0,1]^2$. Then $\int_{x_2} f(x_1, x_2) dx_2 = 0$ and $\int_{x_1} f(x_1, x_2) dx_1 = 0$ for each x_1 and x_2. An additive approximation would set $f_1 = 0$ and $f_2 = 0$. Next, consider the function

$$f(x_1, x_2) = x_1 x_2 \quad \text{(tilting slope)} \tag{3.9}$$

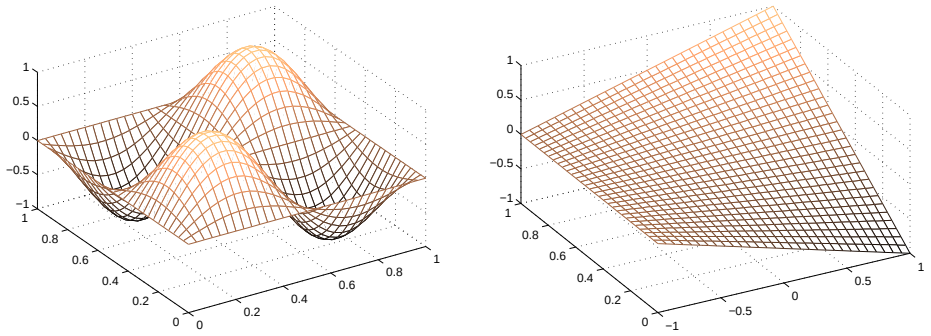
defined for $x_1 \in [-1, 1]$, $x_2 \in [0, 1]$. In this case $\int_{x_1} f(x_1, x_2) dx_1 = 0$ for each x_2; therefore, we expect $f_2 = 0$ under the additive approximation. This function, for every fixed x_2, is a zero-intercept linear function of x_1 with slope x_2.

In order to exploit additive models in variable selection, it is important to understand when the additive approximation accurately captures all of the relevant variables. We call this property *additive faithfulness*. We first formalize the concept that a multivariate function f *does not depend on* a coordinate x_k.

DEFINITION 3.1. Let $C = [0,1]^p$ and let $f : C \to \mathbb{R}$. We say that f *does not depend on coordinate* k if for all $\boldsymbol{x}_{-k}$, $f(x_k, \boldsymbol{x}_{-k})$ is a constant as a function of x_k. If f is differentiable, then f does not depend on k if $\partial_{x_k} f(x_k, \boldsymbol{x}_{-k})$ is 0 for all $\boldsymbol{x}_{-k}$.

In addition, suppose we have a distribution over C and the additive approximation

$$f_k^*, \mu^* := \underset{f_1,\ldots,f_p,\mu}{\arg\min} \Big\{ \mathbb{E}\Big[\Big(f(X) - \sum_{k=1}^{p} f_k(X_k) - \mu \Big)^2 \Big] \,:\, \mathbb{E}f_k(X_k) = 0 \Big\}. \tag{3.10}$$



(a) egg carton (b) tilting slope

FIG 1. *Two additively unfaithful functions. Relevant variables are zeroed out under an additive approximation because every "slice" of the function integrates to zero.*

We say that f is *additively faithful* under F if $f_k^* = 0$ implies that f does not depend on coordinate k.

Additive faithfulness is an attractive property because it implies that, in the population setting, the additive approximation yields a consistent variable screening.

3.1. *Additive Faithfulness of Convex Functions.* We now show that under a general class of distributions which we characterize below, convex multivariate functions are additively faithful.

DEFINITION 3.2. Let $p(\boldsymbol{x})$ be a density supported on $[0,1]^p$. $p(\boldsymbol{x})$ satisfies the *boundary flatness condition* if for all j, and for all $\boldsymbol{x}_{-j}$,

$$\frac{\partial p(\boldsymbol{x}_{-j} \,|\, x_j)}{\partial x_j} = \frac{\partial^2 p(\boldsymbol{x}_{-j} \,|\, x_j)}{\partial x_j^2} = 0 \quad \text{at } x_j = 0, x_j = 1 \tag{3.11}$$

The boundary flatness condition is a weak condition. For instance, it is satisfied when the density is flat at the boundary of support—more precisely, when the joint density satisfies the property that $\frac{\partial p(x_j, \boldsymbol{x}_{-j})}{\partial x_j} = \frac{\partial^2 p(x_j, \boldsymbol{x}_{-j})}{\partial x_j^2} = 0$ at boundary points $x_j = 0, x_j = 1$. The boundary flatness property is also trivially satisfied when p is a product density.

The following theorem is the main result of this section.

THEOREM 3.1. *Let p be a positive density supported on $C = [0,1]^p$ that satisfies the boundary flatness property. If f is convex and twice differentiable, then f is additively faithful under p.*

We pause to give some intuition before we presenting the full proof. Suppose that the underlying distribution has a product density. Then we know from Lemma 3.1 that the additive approximation zeroes out k when, fixing x_k, every "slice" of f

integrates to zero. We prove Theorem 3.1 by showing that "slices" of convex functions that integrate to zero cannot be "glued together" while still maintaining convexity.

PROOF. Fixing k and using the result of Lemma 3.1, we need only show that for all x_k, $\mathbb{E}[f(X) - \sum_{k'} f_{k'}(X_{k'}) \,|\, x_k] - \mathbb{E}f(X) = 0$ implies that f does not depend on coordinate k, i.e., $\partial_{x_k} f(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}$.

Let us use the shorthand notation that $r(\boldsymbol{x}_{-k}) = \sum_{k' \neq k} f_{k'}(x_{k'})$ and assume without loss of generality that $\mu^* = E[f(X)] = 0$. We then assume that for all x_k,

$$\mathbb{E}[f(X) - r(X_{-k}) \,|\, x_k] \equiv \int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}_{-k} \,|\, x_k)\big(f(\boldsymbol{x}) - r(\boldsymbol{x}_{-k})\big) = 0. \tag{3.12}$$

We let $p'(\boldsymbol{x}_{-k} \,|\, x_k)$ denote $\frac{\partial p(\boldsymbol{x}_{-k} \,|\, x_k)}{\partial x_k}$ and $p''(\boldsymbol{x}_{-k} \,|\, x_k)$ denote $\frac{\partial^2 p(\boldsymbol{x}_{-k} \,|\, x_k)}{\partial x_k^2}$ and likewise for $f'(x_k, \boldsymbol{x}_{-k})$ and $f''(x_k, \boldsymbol{x}_{-k})$. We then differentiate under the integral, valid because all functions are bounded, and obtain

$$\int_{\boldsymbol{x}_{-k}} p'(\boldsymbol{x}_{-k} \,|\, x_k)\big(f(\boldsymbol{x}) - r(\boldsymbol{x}_{-k})\big) + p(\boldsymbol{x}_{-k} \,|\, x_k) f'(x_k, \boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k} = 0 \tag{3.13}$$

$$\int_{\boldsymbol{x}_{-k}} p''(\boldsymbol{x}_{-k} \,|\, x_k)\big(f(\boldsymbol{x}) - r(\boldsymbol{x}_{-k})\big) + 2p'(\boldsymbol{x}_{-k} \,|\, x_k) f'(x_k, \boldsymbol{x}_{-k}) + p(\boldsymbol{x}_{-k} \,|\, x_k) f''(x_k, \boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k} = 0. \tag{3.14}$$

By the boundary flatness condition, we have that $p''(\boldsymbol{x}_{-k} \,|\, x_k)$ and $p'(\boldsymbol{x}_{-k} \,|\, x_k)$ are zero at $x_k = x_k^0 \equiv 0$. The integral equations then reduce to the following:

$$\int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}_{-k} \,|\, x_k^0) f'(x_k^0, \boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k} = 0 \tag{3.15}$$

$$\int_{\boldsymbol{x}_{-k}} p(\boldsymbol{x}_{-k} \,|\, x_k^0) f''(x_k^0, \boldsymbol{x}_{-k}) d\boldsymbol{x}_{-k} = 0. \tag{3.16}$$

Because f is convex, $f(x_k, \boldsymbol{x}_{-k})$ must be a convex function of x_k for all $\boldsymbol{x}_{-k}$. Therefore, for all $\boldsymbol{x}_{-k}$, $f''(x_k^0, \boldsymbol{x}_{-k}) \geq 0$. Since $p(\boldsymbol{x}_{-k} \,|\, x_k^0) > 0$ by the assumption that p is a positive density, we have that $\forall \boldsymbol{x}_{-k}, f''(x_k^0, \boldsymbol{x}_{-k}) = 0$ necessarily.

The Hessian of f at $(x_k^0, \boldsymbol{x}_{-k})$ then has a zero at the k-th main diagonal entry. A positive semidefinite matrix with a zero on the k-th main diagonal entry must have only zeros on the k-th row and column; see proposition 7.1.10 of Horn and Johnson (1990). Thus, at all $\boldsymbol{x}_{-k}$, the gradient of $f'(x_k^0, \boldsymbol{x}_{-k})$ with respect to $\boldsymbol{x}_{-k}$ must be zero. Therefore, $f'(x_k^0, \boldsymbol{x}_{-k})$ must be constant for all $\boldsymbol{x}_{-k}$. By equation 3.15, we conclude that $f'(x_k^0, \boldsymbol{x}_{-k}) = 0$ for all $\boldsymbol{x}_{-k}$. We can use the same reasoning for the case where $x_k = x_k^1$ and deduce that $f'(x_k^1, \boldsymbol{x}_{-k}) = 0$ for all $\boldsymbol{x}_{-k}$.

Because $f(x_k, \boldsymbol{x}_{-k})$ as a function of x_k is convex, it must be that, for all $x_k \in (0, 1)$ and for all $\boldsymbol{x}_{-k}$,

$$0 = f'(x_k^0, \boldsymbol{x}_{-k}) \leq f'(x_k, \boldsymbol{x}_{-k}) \leq f'(x_k^1, \boldsymbol{x}_{-k}) = 0 \tag{3.17}$$

Therefore f does not depend on x_k.

□

Theorem 3.1 plays an important role in our finite sample analysis, where we show that the additive approximation is variable screening consistent, even when the true function is not additive.

REMARK 3.1. We assume twice differentiability in Theorems 3.1 to simplify the proof. We expect, however, that this this smoothness condition is not necessary—every convex function can be approximated arbitrarily well by a smooth convex function.

REMARK 3.2. We have not found natural conditions under which the opposite direction of additive faithfulness holds—conditions implying that if f does not depend on coordinate k, then f_k^* will be zero in the additive approximation. Suppose, for example, that f is only a function of X_1, X_2, and that (X_1, X_2, X_3) follows a degenerate 3-dimensional distribution where $X_3 = f(X_1, X_2) - f^*(X_1) - f_2^*(X_2)$. In this case X_3 exactly captures the additive approximation error. The best additive approximation of f would have a component $f_3^*(x_3) = x_3$ even though f does not depend on x_3.

REMARK 3.3. In Theorem 3.1, we do not assume a parametric form for the additive components. The additive approximations may not be faithful if we assume a parametric form for each of the components. For example, suppose we approximate a convex function $f(X)$ by a linear form $X\beta$. The optimal linear function in the population setting is $\beta^* = \Sigma^{-1}\mathrm{Cov}(X, f(X))$. Suppose the X's are independent and follow a symmetric distribution and suppose $f(\boldsymbol{x}) = x_1^2 - \mathbb{E}[X_1^2]$, then $\beta_1^* = \mathbb{E}[X_1 f(X)] = \mathbb{E}[X_1^3 - X_1\mathbb{E}[X_1^2]] = 0$.

REMARK 3.4. It is possible to get a similar result for distributions with unbounded support, by using a limit condition $\lim_{|x_k|\to\infty} \frac{\partial p(\boldsymbol{x}_{-k}\,|\,x_k)}{\partial x_k} = 0$. Such a limit condition however is not obeyed by many common distributions such as the multivariate Gaussian distribution. The next example shows that certain convex functions are not additive faithful under certain multivariate Gaussian distributions.

EXAMPLE 3.2. Consider a two dimensional quadratic function $f(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T} H\boldsymbol{x} + c$ where $H = \begin{pmatrix} H_{11} & H_{12} \\ H_{12} & H_{22} \end{pmatrix}$ is positive definite and a Gaussian distribution $X \sim N(0, \Sigma)$ where $\Sigma = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}$. As we show in Section 9 of the Appendix, the additive approximation has the following closed form.

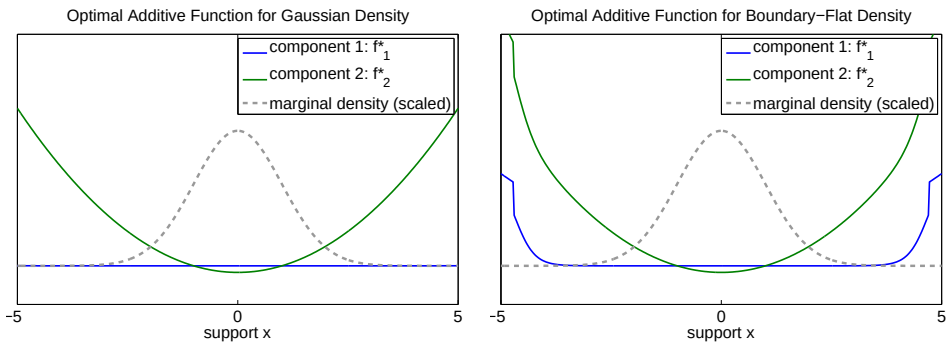
$$f_1^*(x_1) = \left(\frac{T_1 - T_2\alpha^2}{1-\alpha^4}\right) x_1^2 + c_1$$
$$f_2^*(x_2) = \left(\frac{T_2 - T_1\alpha^2}{1-\alpha^4}\right) x_2^2 + c_2$$

Where $T_1 = H_{11} + 2H_{12}\alpha + H_{22}\alpha^2$, $T_2 = H_{22} + 2H_{12}\alpha + H_{11}\alpha^2$, c_1, c_2 are constants such that f_1^* and f_2^* both have mean zero. Let $H = \begin{pmatrix} 1.6 & 2 \\ 2 & 5 \end{pmatrix}$, then it is easy to check that if $\alpha = -\frac{1}{2}$, then $f_1^* = 0$ and additive faithfulness is violated, if $\alpha > \frac{1}{2}$, then f_1^* is a concave function. We take the setting where $\alpha = -0.5$, compute the optimal additive functions via numerical simulation, and show the results in Figure 2(a)–f_1^* is zero as expected.

Although the Gaussian distribution does not satisfy the boundary flatness condition, it is possible to approximate the Gaussian distribution arbitrarily well with distributions that do satisfy the boundary flatness conditions.

EXAMPLE 3.3. Let Σ be as in Example 3.2 with $\alpha = -0.5$ so that $f_1^* = 0$. Consider a mixture $\lambda U[-(b+\epsilon), b+\epsilon]^2 + (1-\lambda)N_b(0, \Sigma)$ where $N_b(0, \Sigma)$ is the density of a *truncated* bivariate Gaussian bounded in $[-b, b]^2$ and $U[-(b+\epsilon), b+\epsilon]^2$ is the uniform distribution over a square. The uniform distribution is supported over a slightly larger square to satisfy the boundary flatness conditions.

When b is large, ϵ is small, and λ is small, the mixture closely approximates the Gaussian distribution but is still additively faithful for convex functions. Figure 2(b) shows the optimal additive components under the mixture distribution, computed by numerical integration with $b = 5, \epsilon = 0.3, \lambda = 0.0001$. True to our theory, f_1^*, which is zero under the Gaussian distribution, is nonzero under the mixture approximation to the Gaussian distribution. We note that the magnitude $\mathbb{E}f_1^*(X_1)^2$, although nonzero, is very small, consistent with the fact that the mixture distribution closely approximates the Gaussian distribution.



(a) Gaussian distribution (b) Mixture approximation

FIG 2. *Optimal additive projection of the quadratic function described in Example 3.2 under both the Gaussian distribution described in Example 3.2 and under the approximately Gaussian mixture distribution described in Example 3.3. For the mixture approximation, we used* $b = 5, \epsilon = 0.3, \lambda = 0.0001$ *where the parameters are defined in Example 3.3. This example shows the effect and the importance of the boundary flatness conditions.*

3.2. *Convex Additive Models.* Although convex functions are additively faithful—under appropriate conditions—it is difficult to estimate the optimal additive functions f_k^*s as defined in equation (3.10). The reason is that f_k^* need not be a convex function, as example 3.2 and example 3.3 show. It may be possible to estimate f_k^* via smoothing, but we prefer an approach that is free of smoothing parameters. Since the true regression function f is convex, we approximate the additive model with a *convex* additive model. We abuse notation and, for the rest of the paper, use the notation f_k^* to represent convex additive fits:

$$\{f_k^*\}_{k=1}^p = \arg\min \left\{ \mathbb{E}\Big(f(X) - \sum_{k=1}^p f_k(X_k) \Big)^2 \;:\; f_k \in \mathcal{C}^1,\, \mathbb{E} f_k(X_k) = 0 \right\} \tag{3.18}$$

where $\mathcal{C}^1$ is the set of univariate convex functions. The convex functions $\{f_k^*\}$ are not additively faithful by themselves, i.e., it could be that the true function f depends on variable k but $f_k^* = 0$. However, faithfulness can be restored by coupling the f_k^*'s with a set of univariate concave fits on the *residual* $f - f^*$:

$$g_k^* = \arg\min \left\{ \mathbb{E}\Big(f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - g_k \Big)^2 \;:\; g_k \in \text{-}\mathcal{C}^1, \mathbb{E} g_k(X_k) = 0 \right\}. \tag{3.19}$$

THEOREM 3.2. *Suppose $p(\boldsymbol{x})$ is a positive density on $C = [0,1]^p$ that satisfies the boundary flatness condition. Suppose that f is convex and twice differentiable, and that $\partial_{x_k} p(\boldsymbol{x}_{-k} \,|\, x_k)$, and $\partial_{x_k}^2 p(\boldsymbol{x}_{-k} \,|\, x_k)$ are all continuous as functions of x_k. Let f_k^* and g_k^* be as defined in equations (3.18) and (3.19), then the f_k^*'s and the g_k^*'s are unique. Furthermore, $f_k^* = 0$ and $g_k^* = 0$ implies that $\partial_{x_k} f(\boldsymbol{x}) = 0$, that is, f does not depend on x_k.*

Before we can prove the theorem, we need a lemma that generalizes Theorem 3.1.

LEMMA 3.2. *Suppose $p(x)$ is a positive density on $C = [0,1]^p$ satisfying the boundary flatness condition. Let $f(\boldsymbol{x})$ be a convex twice differentiable function on C. Let $\phi(\boldsymbol{x}_{-k})$ be any function that does not depend on x_k. Then, we have that the unconstrained univariate function*

$$h_k^* = \arg\min_{f_k} \mathbb{E}\Big[\big(f(X) - \phi(X_{-k}) - h_k(X_k) \big)^2 \Big] \tag{3.20}$$

is given by $h_k^*(x_k) = \mathbb{E}\big[f(X) - \phi(X_{-k}) \,|\, x_k \big]$, and $h_k^* = 0$ implies that $\partial_{x_k} f(\boldsymbol{x}) = 0$.

PROOF. In the proof of Theorem 3.1, the only property of $r(\boldsymbol{x}_{-k})$ we used was the fact that $\partial_{x_k} r(\boldsymbol{x}_{-k}) = 0$. Therefore, the proof here is identical to that of Theorem 3.1 except that we replace $r(\boldsymbol{x}_{-k})$ with $\phi(\boldsymbol{x}_{-k})$. □

PROOF OF THEOREM 3.2. Fix k. Let f_k^* and g_k^* be defined as in equation 3.18 and equation 3.19. Let $\phi(\boldsymbol{x}_{-k}) \equiv \sum_{k' \neq k} f_{k'}^*(x_{k'})$.

Then we have that

$$f_k^* = \arg\min_{f_k} \left\{ \mathbb{E}\big(f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - f_k\big)^2 \ : \ f_k \in \mathcal{C}^1,\, \mathbb{E} f_k(X_k) = 0 \right\} \tag{3.21}$$

$$g_k^* = \arg\min_{g_k} \left\{ \mathbb{E}\big(f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - g_k\big)^2 \ : \ g_k \in \text{-}\mathcal{C}^1,\, \mathbb{E} g_k(X_k) = 0 \right\} \tag{3.22}$$

Let us suppose that $f_k^* = g_k^* = 0$. It must be then that

$$\arg\min_{c \in \mathbb{R}} \mathbb{E}\left(f(X) - \phi(X_{-k}) - c(X_k^2 - m_k^2)\right)^2 = 0$$

where $m_k^2 \equiv \mathbb{E}X_k^2$; this is because $c(x_k^2 - m_k^2)$ is either convex or concave in x_k and it is centered, i.e. $\mathbb{E}[X_k^2 - m_k^2] = 0$. Since the optimum has a closed form $c^* = \frac{\mathbb{E}\left[(f(X) - \phi(X_{-k}))(X_k^2 - m_k^2)\right]}{\mathbb{E}X_k^2}$, we deduce that

$$\begin{aligned}\mathbb{E}\big[(f(X) - \phi(X_{-k}))(X_k^2 - m_k^2)\big] \\ = \mathbb{E}[(f(X) - \phi(X_{-k}))X_k^2] = \mathbb{E}[\mathbb{E}[f(X) - \phi(X_{-k}) \,|\, X_k]X_k^2] = 0\end{aligned}$$

We denote $h_k^*(x_k) = \mathbb{E}[f(X) - \phi(X_{-k}) \,|\, x_k]$.

Under the derivative continuity conditions in the theorem, we apply Lemma 8.3 in the appendix and know that $h_k^*(x_k)$ is twice-differentiable and has a second derivative bounded away from $-\infty$. Therefore, for a large enough positive scalar α, $h_k^*(x_k) + \alpha(x_k^2 - m_k^2)$ has a non-negative second derivative and is thus convex.

Because we assumed $f^* = g^* = 0$, it must be that

$$\arg\min_{c \in \mathbb{R}} \mathbb{E}\Big(f(X) - \phi(X_{-k}) - c\big(h_k^*(X_k) + \alpha(X_k^2 - m_k^2)\big)\Big)^2 = 0$$

This is because $c\big(h_k^*(x_k) + \alpha(x_k^2 - m_k^2)\big)$ is convex for $c \geq 0$ and concave for $c \leq 0$ and it is a centered function.

Again, $c^* = \frac{\mathbb{E}[(f(X) - \phi(X_{-k}))\big(h_k^*(X_k) + \alpha(X_k^2 - m_k^2)\big)]}{\mathbb{E}\big(h_k^*(X_k) + \alpha(X_k^2 - m_k^2)\big)^2} = 0$, so

$$\begin{aligned}\mathbb{E}[(f(X) - \phi(X_{-k}))\big(h_k^*(X_k) + \alpha(X_k^2 - m_k^2)\big)] &= \mathbb{E}[(f(X) - \phi(X_{-k}))h_k^*(X_k)] \\ &= \mathbb{E}\Big[\mathbb{E}[f(X) - \phi(X_{-k}) \,|\, X_k]h_k^*(X_k)\Big] \\ &= \mathbb{E}h_k^*(X_k)^2 = 0\end{aligned}$$

where the first equality follows because $\mathbb{E}[(f(X) - \phi(X_{-k}))(X_k^2 - m_k^2)] = 0$. Therefore, we get that $h_k^* = 0$. Now we use Lemma 3.2 with $\phi(\boldsymbol{x}_{-k}) = f(\boldsymbol{x}) - \sum_{k' \neq k} f_{k'}^*(x_{k'})$ and conclude that $f_k^* = 0$ and $g_k^* = 0$ together imply that f does not depend on x_k.

Now we turn to uniqueness. Suppose for sake of contradiction that f^* and f' are optimal solutions to (3.18) and $\mathbb{E}(f' - f^*)^2 > 0$. $f^* + \lambda(f' - f^*)$ for any $\lambda \in [0, 1]$

must then also be an optimal solution by convexity of the objective and constraint. However, the second derivative of the objective $\mathbb{E}(f - f^* - \lambda(f' - f^*))^2$ with respect to λ is $2\mathbb{E}(f' - f^*)^2 > 0$. The objective is thus strongly convex and $\mathbb{E}(f^* - f')^2 = 0$. We now apply Lemma 8.4 by letting $\phi_k = f_k^* - f_k'$. We conclude that $\mathbb{E}(f_k^* - f_k')^2 = 0$ for all k. The uniqueness of g^* is proved similarly. □

3.3. *Estimation Procedure.* Theorem 3.2 naturally suggests a two-stage screening procedure for variable selection in the population setting. In the first stage, we fit a convex additive model.

$$f_1^*, ..., f_p^* = \underset{f_1,...,f_p \in \mathcal{C}_0^1}{\arg\min}\, \mathbb{E}\Big(f(X) - \sum_{k=1}^{p} f_k(X_k)\Big)^2 \tag{3.23}$$

where we denote $\mathcal{C}_0^1$ ($\text{-}\mathcal{C}_0^1$) as the set of one-dimensional convex (resp. concave) functions with population mean zero. In the second stage, for every variable marked as irrelevant in the first stage, we fit a univariate *concave* function separately on the residual for that variable. for each k such that $f_k^* = 0$:

$$g_k^* = \underset{g_k \in \text{-}\mathcal{C}_0^1}{\arg\min}\, \mathbb{E}\Big(f(X) - \sum_{k'} f_{k'}^*(X_{k'}) - g_k(X_k)\Big)^2 \tag{3.24}$$

We screen out S^C, any variable k that is zero after the second stage, and output S.

$$S^c = \big\{k : f_k^* = 0 \text{ and } g_k^* = 0\big\}. \tag{3.25}$$

We refer to this procedure as AC/DC (additive convex/decoupled concave). Theorem 3.2 guarantees that the true set of relevant variables S_0 must be a subset of S.

It is straightforward to construct a finite sample variable screening procedure, which we describe in Figure 3. We use an ℓ_∞/ℓ_1 penalty in equation (3.26) and an ℓ_∞ penalty in equation (3.24) to encourage sparsity. Other penalties can also produce sparse estimates, such as a penalty on the derivative of each of the component functions. The $\|\cdot\|_\infty$ norm is convenient for both theoretical analysis and implementation.

The optimization in (3.26) appears to be infinite dimensional, but it is equivalent to a finite dimensional quadratic program. In the following section, we give the details of this optimization, and show how it can be reformulated to be more computationally efficient.

4. Optimization. We now describe in detail the optimization algorithm for the additive convex regression stage. The second decoupled concave regression stage follows a very similar procedure.

Let $\boldsymbol{x}_i \in \mathbb{R}^p$ be the covariate, let y_i be the response and let ϵ_i be the mean zero noise. The regression function $f(\cdot)$ we estimate is the sum of univariate functions $f_k(\cdot)$

Input: $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$, regularization parameter λ.

AC Stage: Estimate a sparse additive convex model:

$$\widehat{f}_1, ..., \widehat{f}_p, \widehat{\mu} = \underset{f_1,...,f_p \in \mathcal{C}_0^1}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \mu - \sum_{k=1}^{p} f_k(x_{ik}) \Big)^2 + \lambda \sum_{k=1}^{p} \|f_k\|_\infty \tag{3.26}$$

DC Stage: Estimate concave functions for each k such that $\|\widehat{f}_k\|_\infty = 0$:

$$\widehat{g}_k = \underset{g_k \in \text{-}\mathcal{C}_0^1}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \widehat{\mu} - \sum_{k'} \widehat{f}_{k'}(x_{ik'}) - g_k(x_{ik}) \Big)^2 + \lambda \|g_k\|_\infty \tag{3.27}$$

Output: Component functions $\{\widehat{f}_k\}$ and relevant variables $\widehat{S}$ where

$$\widehat{S}^c = \big\{ k : \|\widehat{f}_k\| = 0 \text{ and } \|\widehat{g}_k\| = 0 \big\}. \tag{3.28}$$

Fig 3. *The AC/DC algorithm for variable selection in convex regression. The AC stage fits a sparse additive convex regression model, using a quadratic program that imposes an group sparsity penalty for each component function. The DC stage fits decoupled concave functions on the residuals, for each component that is zeroed out in the AC stage.*

in each variable dimension and a scalar offset μ. We impose additional constraints that each function $f_k(\cdot)$ is convex, which can be represented by its supporting hyperplanes, i.e.,

$$f_{i'k} \geq f_{ik} + \beta_{ik}(x_{i'k} - x_{ik}) \quad \text{for all } i, i' = 1, \ldots, n, \tag{4.1}$$

where $f_{ik} := f_k(x_{ik})$ is the function value and β_{ik} is a subgradient at point x_{ik}. This ostensibly requires $O(n^2 p)$ constraints to impose the supporting hyperplane constraints. In fact, only $O(np)$ constraints suffice, since univariate convex functions are characterized by the condition that the subgradient, which is a scalar, must increase monotonically. This observation leads to the optimization

$$\begin{aligned} \min_{\{f_k, \beta_k\}, \mu} \quad & \frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \mu - \sum_{k=1}^{p} f_{ik} \Big)^2 + \lambda \sum_{k=1}^{p} \|f_k\|_\infty \\ \text{subject to} \quad & \text{for all } k = 1, \ldots, p\text{:} \\ & f_{\pi_k(i+1)k} = f_{\pi_k(i)k} + \beta_{\pi_k(i)k}(x_{\pi_k(i+1)k} - x_{\pi_k(i)k}), \text{ for } i = 1, \ldots, n-1 \\ & \sum_{i=1}^{n} f_{ik} = 0, \\ & \beta_{\pi_k(i+1)k} \geq \beta_{\pi_k(i)k} \text{ for } i = 1, \ldots, n-1. \end{aligned} \tag{4.2}$$

Here f_k denotes the vector $f_k = (f_{1k}, f_{2_k} \ldots, f_{nk})^T \in \mathbb{R}^n$ and $\{\pi_k(1), \pi_k(2), \ldots, \pi_k(n)\}$

are the indices in the sorted ordering of the values of coordinate k:

$$x_{\pi_k(1)k} \le x_{\pi_k(2)k} \le \cdots \le x_{\pi_k(n)k}. \tag{4.3}$$

We can solve for μ explicitly as $\mu = \frac{1}{n}\sum_{i=1}^n y_i = \overline{y}$. This follows from the KKT conditions and the constraints $\sum_i f_{ki} = 0$.

The sparse convex additive model optimization in (4.2) is a quadratic program with $O(np)$ variables and $O(np)$ constraints. Directly applying a QP solver for f and β is computationally expensive for relatively large n and p. However, notice that variables in different feature dimensions are only coupled in the squared error term $(y_i - \mu - \sum_{k=1}^p f_{ik})^2$. Hence, we can apply the block coordinate descent method, where in each step we solve the following QP subproblem for $\{f_k, \beta_k\}$ with the other variables fixed. In matrix notation, the optimization is

$$\begin{aligned} \min_{f_k,\beta_k,\gamma_k} \quad & \frac{1}{2n}\|r_k - f_k\|_2^2 + \lambda\gamma_k \\ \text{such that} \quad & P_k f_k = \operatorname{diag}(P_k \boldsymbol{x}_k)\beta_k \\ & D_k\beta_k \le 0 \\ & -\gamma_k \mathbf{1}_n \le f_k \le \gamma_k \mathbf{1}_n \\ & \mathbf{1}_n^\mathsf{T} f_k = 0 \end{aligned} \tag{4.4}$$

where $\beta_k \in \mathbb{R}^{n-1}$ is the vector $\beta_k = (\beta_{1k}, \ldots, \beta_{(n-1)k})^T$, and $r_k \in \mathbb{R}^n$ is the residual vector $r_k = (y_i - \widehat{\mu} - \sum_{k' \ne k} f_{ik'})^T$. In addition, $P_k \in \mathbb{R}^{(n-1)\times n}$ is a permutation matrix where the i-th row is all zeros except for the value -1 in position $\pi_k(i)$ and the value 1 in position $\pi_k(i+1)$, and $D_k \in \mathbb{R}^{(n-2)\times(n-1)}$ is another permutation matrix where the i-th row is all zeros except for a value 1 in position $\pi_k(i)$ and a value -1 in position $\pi_k(i+1)$. We denote by $\operatorname{diag}(v)$ the diagonal matrix with diagonal entries v. The extra variable γ_k is introduced to impose the regularization penalty involving the ℓ_∞ norm.

This QP subproblem involves $O(n)$ variables, $O(n)$ constraints and a sparse structure, which can be solved efficiently using optimization packages. In our experiments we use MOSEK (www.mosek.com). We cycle through all covariates k from 1 to p multiple times until convergence. Empirically, we observe that the algorithm converges in only a few cycles. We also implemented an ADMM solver for (4.2) (Boyd *et al.*, 2011), but found that it is not as efficient as this blockwise QP solver.

After optimization, the function estimate for an input vector $\boldsymbol{x}$ is, according to (4.1),

$$\widehat{f}(\boldsymbol{x}) = \sum_{k=1}^p \widehat{f}_k(x_k) + \widehat{\mu} = \sum_{k=1}^p \max_i \Big\{\widehat{f}_{ik} + \widehat{\beta}_{ik}(x_k - x_{ik})\Big\} + \widehat{\mu}. \tag{4.5}$$

The univariate concave function estimation required in the DC stage is a straightforward modification of optimization (4.4). It is only necessary to modify the linear inequality constraints so that the subgradients are non-increasing: $\beta_{\pi_k(i+1)k} \le \beta_{\pi_k(i)k}$.

4.1. *Alternative Formulation.* Optimization (4.2) can be reformulated in terms of the second derivatives. The alternative formulation replaces the order constraints $\beta_{\pi_k(i+1)k} \geq \beta_{\pi_k(i)k}$ with positivity constraints, which simplifies the analysis.

Define $d_{\pi_k(i)k}$ as the second derivative: $d_{\pi_k(1)k} = \beta_{\pi_k(1)k}$, and $d_{\pi_k(i)k} = \beta_{\pi_k(i)k} - \beta_{\pi_k(i-1)k}$ for $i > 1$. The convexity constraint is equivalent to the constraint that $d_{\pi_k(i)k} \geq 0$ for all $i > 1$.

It is easy to verify that $\beta_{\pi_k(i)k} = \sum_{j \leq i} d_{\pi_k(j)k}$ and

$$\begin{aligned}
f_k(x_{\pi_k(i)k}) =& f_k(x_{\pi_k(i-1)k}) + \beta_{\pi_k(i-1)k}(x_{\pi_k(i)k} - x_{\pi_k(i-1)k}) \\
=& f_k(x_{\pi_k(1)k}) + \sum_{j<i} \beta_{\pi_k(j)k}(x_{\pi_k(j)k} - x_{\pi_k(j-1)k}) \\
=& f_k(x_{\pi_k(1)k}) + \sum_{j<i} \sum_{j' \leq j} d_{\pi_k(j')k}(x_{\pi_k(j)k} - x_{\pi_k(j-1)k}) \\
=& f_k(x_{\pi_k(1)k}) + \sum_{j'<i} d_{\pi_k(j')k} \sum_{i>j\geq j'} (x_{\pi_k(j)k} - x_{\pi_k(j-1)k}) \\
=& f_k(x_{\pi_k(1)k}) + \sum_{j'<i} d_{\pi_k(j')k}(x_{\pi_k(i)k} - x_{\pi_k(j')k}).
\end{aligned}$$

We can write this more compactly in matrix notation as

$$\begin{bmatrix} f_k(x_{\pi_k(1)k}) \\ f_k(x_{\pi_k(2)k}) \\ \vdots \\ f_k(x_{\pi_k(n)k}) \end{bmatrix} = \begin{bmatrix} (x_{1k} - x_{\pi_k(1)k})_+ & \cdots & (x_{1k} - x_{\pi_k(n-1)k})_+ \\ \cdots & & \\ (x_{nk} - x_{\pi_k(1)k})_+ & \cdots & (x_{nk} - x_{\pi_k(n-1)k})_+ \end{bmatrix} \begin{bmatrix} d_{\pi_k(1)k} \\ \cdots \\ d_{\pi_k(n-1)k} \end{bmatrix} + \mu_k$$

$$\equiv \Delta_k d_k + \mu_k \tag{4.6}$$

where Δ_k is a $n \times n-1$ matrix such that $\Delta_k(i,j) = (x_{ik} - x_{\pi_k(j)k})_+$, $d_k = (d_{\pi_k(1)k}, \ldots, d_{\pi_k(n-1)k})$, and $\mu_k = f_k(x_{\pi_k(1)k})\mathbf{1}_n$. Because f_k has to be centered, $\mu_k = -\frac{1}{n}\mathbf{1}_n^\mathsf{T}\Delta_k d_k$, and therefore

$$\Delta_k d_k + \mu_k \mathbf{1}_n = \Delta_k d_k - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\Delta_k d_k = \overline{\Delta}_k d_k \tag{4.7}$$

where $\overline{\Delta}_k \equiv \Delta_k - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\Delta_k$ is Δ_k with the mean of each column subtracted.

The above derivations prove the following proposition, which states that (4.2) has an alternative formulation.

PROPOSITION 4.1. *Let $\{\widehat{f}_k, \widehat{\beta}_k\}_{k=1,\ldots,p}$ be an optimal solution to (4.2) and suppose $\overline{Y} = 0$. Define vectors $\widehat{d}_k \in \mathbb{R}^{n-1}$ such that $\widehat{d}_{\pi_k(1)k} = \widehat{\beta}_{\pi_k(1)k}$ and $\widehat{d}_{\pi_k(i)k} = \widehat{\beta}_{\pi_k(i)k} - \widehat{\beta}_{\pi_k(i-1)k}$ for $i > 1$. Then $\widehat{f}_k = \overline{\Delta}_k\widehat{d}_k$ and $\widehat{d}_k$ is an optimal solution to the following optimization:*

$$\min_{\{d_k \in \mathbb{R}^{n-1}\}_{k=1,\ldots p}} \frac{1}{2n}\Big\|Y - \sum_{k=1}^p \overline{\Delta}_k d_k\Big\|_2^2 + \lambda_n \sum_{k=1}^p \|\overline{\Delta}_k d_k\|_\infty \tag{4.8}$$
$$\text{such that } d_{\pi_k(2)k}, \ldots, d_{\pi_k(n-1)k} \geq 0 \qquad \textit{(convexity)}.$$

Likewise, suppose $\{\widehat{d}_k\}_{k=1,\dots p}$ *is a solution to* (4.8)*, define* $\widehat{\beta}_{\pi_k(i)k} = \sum_{j\le i} \widehat{d}_{\pi_k(j)k}$ *and* $\widehat{f}_k = \overline{\Delta}_k \widehat{d}_k$*. Then* $\{\widehat{f}_k, \widehat{\beta}_k\}_{k=1,\dots,p}$ *is an optimal solution to* (4.2)*.* $\overline{\Delta}$ *is the* n *by* $n-1$ *matrix defined by* (4.7)*.*

The decoupled concave postprocessing stage optimization is again similar. Specifically, suppose $\widehat{d}_k$ is the output of optimization (4.8), and define the residual vector

$$\widehat{r} = Y - \sum_{k=1}^{p} \overline{\Delta}_k \widehat{d}_k. \tag{4.9}$$

Then for all k such that $\widehat{d}_k = 0$, the DC stage optimization is formulated as

$$\begin{aligned} \min_{c_k} \quad & \frac{1}{2n}\left\|\widehat{r} - \Delta_k c_k\right\|_2^2 + \lambda_n \|\Delta_k c_k\|_\infty \\ \text{such that} \quad & c_{\pi_k(2)k}, \dots, c_{\pi_k(n-1)k} \le 0 \qquad \text{(concavity)}. \end{aligned} \tag{4.10}$$

We can use either the off-centered Δ_k matrix or the centered $\overline{\Delta}_k$ matrix because the concave estimations are decoupled and hence are not subject to non-identifiability under additive constants.

5. Analysis of Variable Screening Consistency. Our goal is to show that variable screening consistency. That is, as $n, p \to \infty$, $\mathbb{P}(\widehat{S} = S)$ approaches 0 where $\widehat{S}$ is the set of variables outputted by AC/DC in the finite sample setting (Algorithm 3) and S is the set of variables outputted in the population setting (3.25).

We divide our analysis into two parts. We first establish a sufficient deterministic condition for consistency of the sparsity pattern screening procedure. We then consider the stochastic setting and argue that the deterministic conditions hold with high probability. Note that in all of our results and analysis, we let c, C represent absolute constants; the actual values of c, C may change from line to line. We derived two equivalent optimizations for AC/DC: (4.2) outputs $\widehat{f}_k, \widehat{g}_k$ and (4.8) outputs the second derivatives $\widehat{d}_k$. Their equivalence is established in Proposition 4.1 and we use both $\widehat{d}_k$ and $\widehat{f}_k$ in our analysis. We will also assume in this section that the true regression function f_0 has mean-zero and therefore, we will omit the intercept term $\widehat{mu}$ from our estimation procedure.

In our analysis, we assume that an upper bound B to $\|\widehat{f}_k\|_\infty$ is imposed in the optimization procedure, where B is chosen to also upper bound $\|f_k^*\|_\infty$. This B-boundedness constraint is added so that we may use the convex function bracketing results from Kim and Samworth (2014) to establish uniform convergence between the empirical risk and the population risk. We emphasize that this constraint is not needed in practice and we do not use it for any of our simulations.

5.1. *Deterministic Setting.* We analyze Optimization 4.8 and construct an additive convex solution $\{\widehat{d}_k\}_{k=1,\dots,p}$ that is zero for $k \in S^c$, where S is the set of relevant variables, and show that it satisfies the KKT conditions for optimality of

optimization (4.8). We define $\widehat{d}_k$ for $k \in S$ to be a solution to the restricted regression (defined below). We also show that $\widehat{c}_k = 0$ satisfies the optimality condition of optimization (4.10) for all $k \in S^c$.

DEFINITION 5.1. We define the *restricted regression* problem

$$\min_{d_k} \frac{1}{n}\left\|Y - \sum_{k\in S} \overline{\Delta}_k d_k\right\|_2^2 + \lambda_n \sum_{k\in S} \|\overline{\Delta}_k d_k\|_\infty \quad \text{such that } d_{k,1}, \ldots, d_{k,n-1} \geq 0$$

where we restrict the indices k in optimization (4.8) to lie in some set S which contains the true relevant variables.

THEOREM 5.1 (Deterministic setting). *Let $\{\widehat{d}_k\}_{k\in S}$ be a minimizer of the restricted regression as defined above. Let $\widehat{r} := Y - \sum_{k\in S} \overline{\Delta}_k \widehat{d}_k$ be the restricted regression residual.*

Let $\pi_k(i)$ be a reordering of X_k in ascending order so that $X_{k\pi_k(n)}$ is the largest entry. Let $\mathbf{1}_{\pi_k(i:n)}$ be 1 on the coordinates $\pi_k(i), \pi_k(i+1), ..., \pi_k(n)$ and 0 elsewhere. Define $\mathsf{range}_k = X_{k\pi_k(n)} - X_{k\pi_k(1)}$.

Suppose for all $k \in S^c$, for all $i = 1, \ldots, n$, $\lambda_n > \mathsf{range}_k |\frac{32}{n} \widehat{r}^\mathsf{T} \mathbf{1}_{\pi_k(i:n)}|$. Suppose also that for all $k \in S^c$, $\max_{i=1,\ldots,n-1} \frac{X_{k\pi_k(i+1)} - X_{k\pi_k(i)}}{\mathsf{range}_k} \geq \frac{1}{16}$, and $\mathsf{range}_k \geq 1$.

Then the following two statements hold.

1. *Let $\widehat{d}_k = 0$ for $k \in S^c$. Then $\{\widehat{d}_k\}_{k=1,\ldots,p}$ is an optimal solution to optimization (4.8). Furthermore, any solution to the optimization program (4.8) must be zero on S^c.*
2. *For all $k \in S^c$, the solution $\widehat{c}_k$ to optimization (4.10) must be zero and unique.*

Theorem 5.1 states that the estimator produces no false positive so long as λ_n upper bounds the partial sums of the residual $\widehat{r}$ and that the maximum gap between ordered values of X_k is small.

This result holds regardless of whether or not we impose the boundedness conditions in optimization (4.8) and (4.10). The full proof of Theorem 5.1 is in Section 8.1 of the Appendix. We allow S in Theorem 5.1 to be any set containing the relevant variables; in Lasso analysis, S is taken to be the set of relevant variables; we will take S to be the set of variables chosen by the additive convex and decoupled concave procedure in the population setting, which is guaranteed to contain the relevant variables because of additive faithfulness.

Theorem 5.1 allows us to separately analyze the false negative rates and false positive rates. To control false positives, we analyze the condition on λ_n for $k \in S^c$. To control false negatives, we need only analyze the restricted regression with only $|S|$ variables.

The proof of Theorem 5.1 analyses the KKT conditions of optimization (4.8). This parallels the now standard *primal-dual witness* technique (Wainwright, 2009).

However, we cannot derive analogous *mutual incoherence* conditions because the estimation is nonparametric—even the low dimensional restricted regression has $s(n-1)$ variables. The details of the proof are given in Section 8.1 of the Appendix.

5.2. *Probabilistic Setting.* In the probabilistic setting we treat the covariates as random. We adopt the following standard setup:

1. The data $X^{(1)}, \dots, X^{(n)} \sim F$ are iid from a distribution F with a density $p(\boldsymbol{x})$ that is supported and strictly positive on $\mathcal{X} = [-1, 1]^p$.
2. The response is $Y = f_0(X) + W$ where W is independent, zero-mean noise; thus $Y^{(i)} = f_0(X^{(i)}) + W^{(i)}$.
3. The regression function f_0 satisfies $f_0(X) = f_0(X_{S_0})$, where $S_0 = \{1, \dots, s_0\}$ is the set of relevant variables.

Let $\mathcal{C}^1$ denote the set of univariate convex functions supported on $[-1, 1]$, and let $\mathcal{C}_1^p$ denote the set of convex additive functions $\mathcal{C}_1^p \equiv \{f : f = \sum_{k=1}^p f_k,\ f_k \in \mathcal{C}^1\}$. Let $f^*(\boldsymbol{x}) = \sum_{k=1}^p f_k^*(x_k)$ be the population risk minimizer in $\mathcal{C}_1^p$,

$$f^* = \arg\min_{f \in \mathcal{C}_1^p} \mathbb{E}\big(f_0(X) - f^*(X)\big)^2. \tag{5.1}$$

f^* is the unique minimizer by Theorem 3.2. Similarly, we define -$\mathcal{C}^1$ as the set of univariate concave functions supported on $[-1, 1]$ and define

$$g_k^* = \arg\min_{g_k \in \text{-}\mathcal{C}^1} \mathbb{E}\big(f_0(X) - f^*(X) - g_k(X_k)\big)^2. \tag{5.2}$$

The $\widehat{g}_k$'s are unique minimizers as well. We let $S = \{k = 1, \dots, p : f_k^* \neq 0 \text{ or } g_k^* \neq 0\}$ and let $s = |S|$. By additive faithfulness (Theorem 3.2), it must be that $S_0 \subset S$ and thus $s \geq s_0$. In some cases, such as when $X_{S_0}, X_{S_0^c}$ are independent, we have $S = S_0$. Each of our theorems will use a subset of the following assumptions:

A1: X_S, X_{S^c} are independent.

A2: f_0 is convex and twice-differentiable. $\mathbb{E} f_0(X) = 0$.

A3: $\|f_0\|_\infty \leq sB$ and $\|f_k^*\|_\infty \leq B$ for all k.

A4: W is mean-zero sub-Gaussian, independent of X, with scale σ; i.e., for all $t \in \mathbb{R}$, $\mathbb{E} e^{t\epsilon} \leq e^{\sigma^2 t^2/2}$.

A5: The density $p(\boldsymbol{x})$ is bounded away from 0 and satisfies the boundary flatness condition. (Definition 3.2)

By assumption A1, f_k^* must be zero for $k \notin S$. We define α_+, α_- as a measure of the signal strength of the weakest variable:

$$\begin{aligned} \alpha_+ &= \inf_{f \in \mathcal{C}_1^p :\, \text{supp}(f) \supsetneq \text{supp}(f^*)} \Big\{\mathbb{E}\big(f_0(X) - f(X)\big)^2 - \mathbb{E}\big(f_0(X) - f^*(X)\big)^2\Big\} \\ \alpha_- &= \min_{k \in S :\, g_k^* \neq 0} \Big\{\mathbb{E}\big(f_0(X) - f^*(X)\big)^2 - \mathbb{E}\big(f_0(X) - f^*(X) - g_k^*(X_k)\big)^2\Big\} \end{aligned} \tag{5.3}$$

α_+ is a lower bound on the excess risk incurred by any additive convex function whose support is strictly smaller than f^*; $\alpha_+ > 0$ since f^* is the unique risk minimizer. Likewise, α_- lower bounds the excess risk of any decoupled concave fit of the residual $f_0 - f^*$ that is strictly more sparse than the optimal decoupled concave fit $\{\widehat{g}_k^*\}$; $\alpha_- > 0$ by the uniqueness of $\{g_k^*\}$ as well. These quantities play the role of the absolute value of the smallest nonzero coefficient in the true linear model in lasso theory. Intuitively, if α_+ is small, then it is easier to make a false omission in the additive convex stage of the procedure. If α_- is small, then it is easier to make a false omission in the decoupled concave stage of the procedure.

REMARK 5.1. We make strong assumptions on the covariates in A1 in order to make weak assumptions on the true regression function f_0 in A2. In particular, we do not assume that f_0 is additive. In important direction for future work is to weaken assumption A1. Our simulation experiments indicate that the procedure can be effective even when the relevant and irrelevant variables are correlated.

THEOREM 5.2 (Controlling false positives). *Suppose assumptions A1-A5 hold. Define* $\widetilde{\sigma} \equiv \max(\sigma, B)$ *and define* $\mathsf{range}_k = X_{k\pi_k(n)} - X_{k\pi_k(1)}$. *Suppose* $p \leq O\big(\exp(cn)\big)$ *and* $n \geq C$ *for some positive constants* $0 < c < 1$ *and* C. *Suppose also*

$$\lambda_n \geq 512 s \widetilde{\sigma} \sqrt{\frac{\log^2 np}{n}}. \tag{5.4}$$

Then with probability at least $1 - \frac{24}{n}$, *for all* $k \in S^c$, *for all* $i = 1, \ldots, n$,

$$\lambda_n \geq \mathsf{range}_k \left| \frac{32}{n} \widehat{r}^{\mathsf{T}} \mathbf{1}_{(i':n)_k} \right|, \tag{5.5}$$

$\max_{i'} \frac{X_{k\pi_k(i'+1)} - X_{k\pi_k(i')}}{\mathsf{range}_k} \leq \frac{1}{16}$, $\mathsf{range}_k \geq 1$, *and both the AC solution* $\widehat{f}_k$ *from optimization* (4.8) *and the DC solution* $\widehat{g}_k$ *from optimization* (4.10) *are zero.*

The proof of Theorem 5.2 exploits independence of $\widehat{r}$ and X_k under assumption A1; when $\widehat{r}$ and X_k are independent, $\widehat{r}^{\mathsf{T}} \mathbf{1}_{(i':n)}$ is the sum of $n - i' + 1$ random coordinates of $\widehat{r}$. We can then use concentration of measure results for sampling without replacement to argue that $|\frac{1}{n} \widehat{r}^{\mathsf{T}} \mathbf{1}_{(i':n)}|$ is small with high probability. The result of Theorem 5.1 is then used. The full proof of Theorem 5.2 is in Section 8.2 of the Appendix.

THEOREM 5.3 (Controlling false negatives). *Suppose assumptions A1-A5 hold. Let* $\widehat{f}$ *be any AC solution to the restricted regression with* B*-boundedness constraint, and let* $\widehat{g}_k$ *be any DC solution to the restricted regression with* B*-boundedness constraint. Let* $\widetilde{\sigma}$ *denote* $\max(\sigma, B)$. *Suppose*

$$\lambda_n \leq 512 s \widetilde{\sigma} \sqrt{\frac{\log^2 np}{n}} \tag{5.6}$$

and that n is sufficiently large so that, for some constant $c' > 1$,

$$\frac{n^{4/5}}{\log np} \geq c' B^4 \widetilde{\sigma}^2 s^5. \tag{5.7}$$

Assume that the signal-to-noise ratio satisfies

$$\frac{\alpha_+}{\widetilde{\sigma}} \geq cB^2 \sqrt{\frac{s^5}{n^{4/5}} \log^2 np} \tag{5.8}$$

$$\frac{\alpha_-^2}{\widetilde{\sigma}} \geq cB^2 \sqrt{\frac{s^5}{n^{4/5}} \log^2 np} \tag{5.9}$$

where c is a constant. Then with probability at least $1 - \frac{C}{n}$ for some constant C, $\widehat{f}_k \neq 0$ or $\widehat{g}_k \neq 0$ for all $k \in S$.

This is a finite sample version of Theorem 3.1. We need stronger assumptions in Theorem 5.3 to use our additive faithfulness result, Theorem 3.1. The full proof of Theorem 5.3 is in Section 8.3 of the appendix.

Combining Theorems 5.2 and 5.3 we obtain the following result.

COROLLARY 5.1. *Suppose the assumptions of Theorem 5.2 and Theorem 5.3 hold. Then with probability at least $1 - \frac{C}{n}$*

$$\widehat{f}_k \neq 0 \text{ or } \widehat{g}_k \neq 0 \text{ for all } k \in S \tag{5.10}$$

$$\widehat{f}_k = 0 \text{ and } \widehat{g}_k = 0 \text{ for all } k \notin S \tag{5.11}$$

for some constant C.

The above corollary implies that consistent variable selection is achievable with an exponential scaling of the ambient dimension scaling, $p = O(\exp(n^c))$ for some $0 < c < 1$, just as in parametric models. The cost of nonparametric modeling through shape constraints is reflected in the scaling with respect to the number of relevant variables, which can scale as $s = o(n^{4/25})$.

REMARK 5.2. Comminges and Dalalyan (2012) have shown that under tradtional smoothness constraints, even with a product distribution, variable selection is achievable only if $n > O(e^{s_0})$. It is interesting to observe that because of additive faithfulness, the convexity assumption enables a much better scaling of $n = O(\text{poly}(s_0))$, demonstrating that geometric constraints can be quite different from the previously studied smoothness conditions.

6. Experiments. We first illustrate our methods using a simulation of the model

$$Y_i = x_{iS}^\top Q x_{iS} + \epsilon_i \quad (i = 1, 2, \ldots, n).$$

Here x_i denotes data sample i drawn from $\mathcal{N}(0, \Sigma)$, x_{iS} is a subset of x_i with dimension $|S| = 5$, where S represents the relevant variable set, and ϵ_i is additive noise drawn from $\mathcal{N}(0, 1)$. The matrix Q is a symmetric positive definite matrix of dimension $|S| \times |S|$. Note that if Q is diagonal, then the true function is convex additive; otherwise the true function is convex but not additive. For all simulations in this section, we set $\lambda = 4\sqrt{\log(np)/n}$.

In the first simulation, we vary the ambient dimension p. We set Q as one on the diagonal and 1/2 on the off-diagonal with 0.5 probability, and choose $n = 100, 200, \ldots, 1000$ and $p = 64, 128, 256$ and 512. We use independent design by setting $\Sigma = \mathrm{I}_p$. For each (n, p) combination, we generate 100 independent data sets. For each data set we use the AC/DC algorithm and mark feature j as irrelevant if both the AC estimate $\|\widehat{f}_j\|_\infty$ and the DC estimate $\|\widehat{g}_k\|_\infty$ are smaller than 10^{-6}. We plot the probability of exact support recovery over the 100 data sets in Figure 4(a). We observe that the algorithm performs consistent variable selection even if the dimensionality is large. To give the reader a sense of the running speed, for a data set with $n = 1000$ and $p = 512$, the code runs in roughly two minutes on a machine with 2.3 GHz Intel Core i5 CPU and 4 GB memory.

In the second simulation, we vary the sparsity of the Q matrix, that is, we vary the extent to which the true function is non-additive. We generate four Q matrices plotted in Figure 4(b), where the diagonal elements are all one and the off-diagonal elements are $\frac{1}{2}$ with probability α ($\alpha = 0, 0.2, 0.5, 1$ for the four cases). We fix $p = 128$ and choose $n = 100, 200, \ldots, 1000$. We use independent design by seting $\Sigma = \mathrm{I}_p$. We again run the AC/DC optimization on 100 independently generated data sets and plot the probability of recovery in Figure 4(c). The results demonstrate that AC/DC performs consistent variable selection even if the true function is not additive (but still convex).

In the third simulation, we use correlated design and vary the correlation. We let x_i be drawn from $\mathcal{N}(0, \Sigma)$ instead of $\mathcal{N}(0, I_p)$, with $\Sigma_{ij} = \nu^{|i-j|}$. We use the non-additive Q, same as in the second experiment, with $\alpha = 0.5$ and fix $p = 128$. The recovery curves for $\nu = 0.2, 0.4, 0.6, 0.8$ are shown in Figure 4(d). As can be seen, for design of moderate correlation, AC/DC can still select relevant variables well.

We next use the Boston housing data rather than simulated data. This data set contains 13 covariates, 506 samples and one response variable indicating housing values in suburbs of Boston. The data and detailed description can be found on the UCI Machine Learning Repository website[1].

We first use all $n = 506$ samples (with standardization) in the AC/DC algorithm, using a set of candidate regularization parameters $\{\lambda^{(t)}\}$ ranging from $\lambda^{(1)} = 0$ (no regularization) to 2. For each $\lambda^{(t)}$ we obtain a function value matrix $h^{(t)}$ with $p = 13$ columns. The non-zero columns in this matrix indicate the variables selected using $\lambda^{(t)}$. We plot $\|h^{(t)}\|_\infty$ and the column-wise mean of $h^{(t)}$ versus the normalized norm

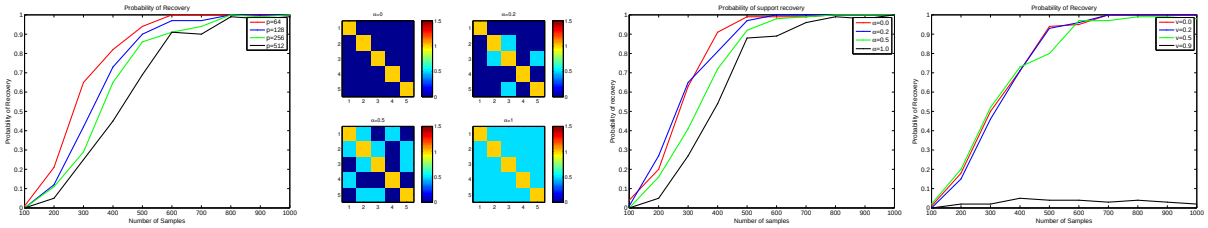
[1] http://archive.ics.uci.edu/ml/datasets/Housing

$\frac{\|h^{(t)}\|_{\infty,1}}{\|h^{(1)}\|_{\infty,1}}$ in Figures 5(a) and 5(b). For comparison, we plot the LASSO/LARS result in a similar way in Figure 5(c). From the figures we observe that the first three variables selected by AC/DC and LASSO are the same: `LSTAT`, `RM` and `PTRATIO`, consistent with previous findings (Ravikumar *et al.*, 2007). The fourth variable selected by AC/DC is `INDUS` (with $\lambda^{(t)} = 0.7$). We then refit AC/DC with only these four variables without regularization, and plot the estimated additive functions in Figure 5(e). When refitting, we constrain a component to be convex if it is non-zero in the AC stage and concave if it is non-zero in the DC stage. As can be seen, these functions contain clear nonlinear effects which cannot be captured by LASSO. The shapes of these functions, including the concave shape of the `PTRATIO` function, are in agreement with those obtained by SpAM (Ravikumar *et al.*, 2007).

Next, in order to quantitatively study the predictive performance, we run 3 times 5-fold cross validation, following the same procedure described above—training, variable selection and refitting. A plot of the mean and standard deviation of the predictive mean squared error (MSE) is shown in Figure 5(d). Since for AC/DC the same regularization level $\lambda^{(t)}$ may lead to a slightly different number of selected variables in different folds and runs, the values on the x-axis for AC/DC are not necessarily integers. The figure clearly shows that AC/DC has a lower predictive MSE than LASSO. We also compared the performance of AC/DC with that of Additive Forward Regression (AFR) presented in Liu and Chen (2009), and found that they are similar. The main advantages of AC/DC compared with AFR and SpAM are that there are no smoothing parameters required, and the optimization is formulated as a convex program, guaranteeing a global optimum.

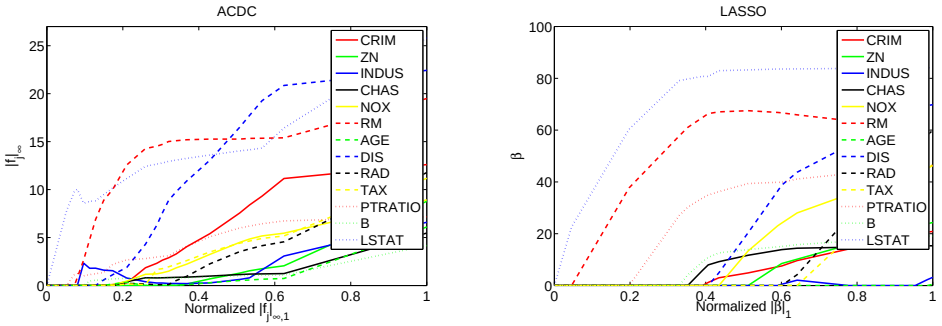
7. Discussion. We have introduced a framework for estimating high dimensional but sparse convex functions. Because of the special properties of convexity, variable selection for convex functions enjoys additive faithfulness—it suffices to carry out variable selection over an additive model, in spite of the approximation error this introduces. Sparse convex additive models can be optimized using block coordinate quadratic programming, which we have found to be effective and scalable. We established variable selection consistency results, allowing exponential scaling in the ambient dimension. We expect that the technical assumptions we have used in these analyses can be weakened; this is one direction for future work. Another interesting direction for building on this work is to allow for additive models that are a combination of convex and concave components. If the convexity/concavity of each component function is known, this again yields a convex program. The challenge is to develop a method to automatically detect the concavity or convexity pattern of the variables.

Acknowledgements. Research supported in part by NSF grants IIS-1116730, AFOSR grant FA9550-09-1-0373, ONR grant N000141210762, and an Amazon AWS in Education Machine Learning Research grant.

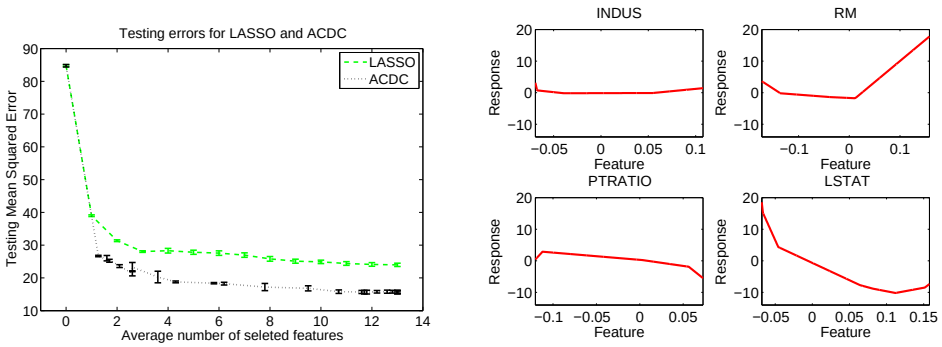


(a) additive model (b) four Q matrices (c) non-additive models (d) correlated design

FIG 4. *Support recovery results where the additive assumption is correct (a), incorrect (b), (c), and with correlated design (d).*



AC/DC $\|f_k\|_\infty$ paths LASSO $|\beta_k|$ paths



predictive MSE estimated functions from AC/DC

FIG 5. *Results on Boston housing data, showing regularization paths, MSE and fitted functions.*

References.

Bertin, K. and Lecué, G. (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic Journal of Statistics* **2** 1224–1241.

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

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011). Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning* **3** 1–122.

Chen, Y. and Samworth, R. J. (2014). Generalised additive and index models with shape constraints. *arXiv preprint arXiv:1404.2957*.

Chen, H. and Yao, D. D. (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer-Verlag.

Comminges, L. and Dalalyan, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics* **40** 2667–2696.

Cule, M., Samworth, R. and Stewart, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density (with discussion). *J. R. Stat. Soc. Ser. B, Stat. Methodol.* **72** 545–600.

DeVore, R., Petrova, G. and Wojtaszczyk, P. (2011). Approximation of functions of few variables in high dimensions. *Constructive Approximation* **33** 125–143.

Goldenshluger, A. and Zeevi, A. (2006). Recovering convex boundaries from blurred and noisy observations. *Ann. Statist* **34** 1375–1394.

Guntuboyina, A. and Sen, B. (2013). Global risk bounds and adaptation in univariate convex regression. *arXiv:1305.1648*.

Hannah, L. A. and Dunson, D. B. (2012). Ensemble Methods for Convex Regression with Applications to Geometric Programming Based Circuit Design. In *International Conference on Machine Learning (ICML)*.

Horn, R. and Johnson, C. (1990). *Matrix Analysis*. Cambridge University Press; Reprint edition.

Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of statistics* **38** 2282.

Kim, A. K. and Samworth, R. J. (2014). Global rates of convergence in log-concave density estimation. *arXiv preprint arXiv:1404.2298*.

Koltchinskii, V. and Yuan, M. (2010). Sparsity in multiple kernel learning. *The Annals of Statistics* **38** 3660–3695.

Lafferty, J. and Wasserman, L. (2008). Rodeo: sparse, greedy nonparametric regression. *The Annals of Statistics* **36** 28–63.

Lele, A. S., Kulkarni, S. R. and Willsky, A. S. (1992). Convex-polygon estimation from support-line measurements and applications to target reconstruction from laser-radar data. *Journal of the Optical Society of America, Series A* **9** 1693–1714.

Lim, E. and Glynn, P. W. (2012). Consistency of Multidimensional Convex Regression. *Operations Research* **60** 196–208.

Liu, H. and Chen, X. (2009). Nonparametric greedy algorithm for the sparse learning problems. In *Advances in Neural Information Processing Systems*.

Meyer, R. F. and Pratt, J. W. (1968). The consistent assessment and fairing of preference functions. *IEEE Trans. Systems Sci. Cybernetics* **4** 270–278.

Mossel, E., O'Donnell, R. and Servedio, R. (2004). Learning functions of k relevant variables. *Journal of Computer and System Sciences* **69** 421–434.

Prince, J. L. and Willsky, A. S. (1990). Reconstructing convex sets from support line measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12** 377–389.

Pya, N. and Wood, S. N. (2014). Shape constrained additive models. *Statistics and Computing* 1–17.

Raskutti, G., Wainwright, M. J. and Yu, B. (2012). Minimax-optimal rates for sparse additive

models over kernel classes via convex programming. *J. Mach. Learn. Res.* **13** 389–427.

Ravikumar, P., Liu, H., Lafferty, J. and Wasserman, L. (2007). Spam: Sparse additive models. In *Advances in Neural Information Processing Systems*.

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

Serfling, R. J. (1974). Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics* **2** 39–48.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. arXiv:1011.3027.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *Information Theory, IEEE Transactions on* **55** 2183–2202.

8. Supplement: Proofs of Technical Results.

8.1. *Proof of the Deterministic Condition for Sparsistency.* We restate Theorem 5.1 first for convenience. The following holds regardless of whether we impose the B-boundedness condition (see discussion at beginning of Section 5 for definition of the B-boundedness condition).

THEOREM 8.1. *Let $\{\widehat{d}_k\}_{k\in S}$ be a minimizer of the restricted regression, that is, the solution to optimization (4.8) where we restrict $k \in S$. Let $\widehat{r} := Y - \sum_{k\in S} \overline{\Delta}_k \widehat{d}_k$ be the restricted regression residual.*

Let $\pi_k(i)$ be an reordering of X_k in ascending order so that $X_{k\pi_k(n)}$ is the largest entry. Let $\mathbf{1}_{\pi_k(i:n)}$ be 1 on the coordinates $\pi_k(i), \pi_k(i+1), ..., \pi_k(n)$ and 0 elsewhere. Define $\mathsf{range}_k = X_{k\pi_k(n)} - X_{k\pi_k(1)}$.

Suppose for all $k \in S^c$ and for all $i = 1, ..., n$, $\lambda_n \geq \mathsf{range}_k |\frac{32}{n} \widehat{r}^{\mathsf{T}} \mathbf{1}_{\pi_k(i:n)}|$, and $\max_{i=1,...,n-1} \frac{X_{k\pi_k(i+1)} - X_{k\pi_k(i)}}{\mathsf{range}_k} \geq \frac{1}{16}$, and $\mathsf{range}_k \geq 1$.

Then the following are true:

1. *Let $\widehat{d}_k = 0$ for $k \in S^c$, then $\{\widehat{d}_k\}_{k=1,...,p}$ is an optimal solution to optimization 4.8. Furthermore, any solution to the optimization program 4.8 must be zero on S^c.*
2. *For all $k \in S^c$, the solution to optimization 4.10 must be zero and unique.*

PROOF. We will omit the B-boundedness constraint in our proof here. It is easy to verify that the result of the theorem still holds with the constraint added in.

We begin by considering the first item in the conclusion of the theorem. We will show that with $\{\widehat{d}_k\}_{k=1,..,p}$ as constructed, we can set the dual variables to satisfy the complementary slackness and stationary conditions: $\nabla_{d_k} \mathcal{L}(\widehat{d}) = 0$ for all k.

The Lagrangian is

$$\mathcal{L}(\{d_k\}, \nu) = \frac{1}{2n} \left\| Y - \sum_{k=1}^{p} \overline{\Delta}_k d_k \right\|_2^2 + \lambda \sum_{k=1}^{p} \|\overline{\Delta}_k d_k\|_\infty - \sum_{k=1}^{p} \sum_{i=2}^{n-1} \nu_{ki} d_{ki} \tag{8.1}$$

with the constraint that $\nu_{ki} \geq 0$ for all k, i.

Because $\{\widehat{d}_k\}_{k\in S}$ is by definition the optimal solution of the restricted regression, it is a consequence that stationarity holds for $k \in S$, that is, $\partial_{\{d_k\}_{k\in S}} \mathcal{L}(d) = 0$, and that the dual variables ν_k for $k \in S$ satisfy complementary slackness.

We now verify that stationarity holds also for $k \in S^c$. We fix one dimension $k \in S^c$ and let $\widehat{r} = Y - \sum_{k'\in S} \overline{\Delta}_{k'} \widehat{d}_{k'}$. The Lagrangian form of the optimization, in terms of just d_k, is

$$\mathcal{L}(d_k, \nu_k) = \frac{1}{2n} \left\| Y - \sum_{k'\in S} \overline{\Delta}_{k'} d_{k'} - \overline{\Delta}_k d_k \right\|_2^2 + \lambda \|\overline{\Delta}_k d_k\|_\infty - \sum_{i=2}^{n-1} \nu_{ki} d_{ki}$$

with the constraint that $v_{ki} \geq 0$ for $i = 2, ..., n-1$.

The derivative of the Lagrangian is

$$\partial_{d_k}\mathcal{L}(d_k) = -\frac{1}{n}\overline{\Delta}_k^{\mathsf{T}}(Y - \sum_{k'\in S}\overline{\Delta}_{k'}d_{k'} - \overline{\Delta}_k d_k) + \lambda\overline{\Delta}_k^{\mathsf{T}}\boldsymbol{u} - \begin{pmatrix} 0 \\ \nu_k \end{pmatrix}$$

where $\boldsymbol{u}$ is the subgradient of $\|\overline{\Delta}_k d_k\|_\infty$. If $\overline{\Delta}_k d_k = 0$, then $\boldsymbol{u}$ can be any vector whose L_1 norm is less than or equal to 1. $\nu_k \geq 0$ is a vector of Lagrangian multipliers. ν_{k1} does not exist because d_{k1} is not constrained to be non-negative.

We now substitute in $d_{k'} = \widehat{d}_{k'}$ for $k' \in S$, $d_k = 0$ for $k \in S$, and $r = \widehat{r}$ and show that the $\boldsymbol{u}, \nu_k$ dual variables can be set in a way to ensure that stationarity:

$$\partial_{d_k}\mathcal{L}(\widehat{d}_k) = -\frac{1}{n}\overline{\Delta}_k^{\mathsf{T}}\widehat{r} + \lambda\overline{\Delta}_k^{\mathsf{T}}\boldsymbol{u} - \begin{pmatrix} 0 \\ \nu_k \end{pmatrix} = 0$$

where $\|\boldsymbol{u}\| \leq 1$ and $\nu_k \geq 0$. It clear that to show stationarity, we only need to show that $[-\frac{1}{n}\overline{\Delta}_k^{\mathsf{T}}\widehat{r} + \lambda\overline{\Delta}_k^{\mathsf{T}}\boldsymbol{u}]_j = 0$ for $j = 1$ and ≥ 0 for $j = 2, ..., n-1$.

To ease notational burden, let us reorder the samples in ascending order so that the i-th sample is the i-th smallest sample. We will from here on write X_{ki} to denote $X_{k\pi_k(i)}$.

Define i^* as the largest index such that $\frac{X_{kn}-X_{ki^*}}{X_{kn}-X_{k1}} \geq 1/2$. We will construct $\boldsymbol{u} = (a - a', 0, ..., -a, 0, ..., a')$ where a, a' are positive scalars, where $-a$ lies at the i^*-th coordinate, and where the coordinates of $\boldsymbol{u}$ correspond to the new sample ordering.

We define

$$\begin{aligned}
\kappa &= \frac{1}{\lambda n}\left[\Delta_k^{\mathsf{T}}\widehat{r}\right]_1 \\
a' &= \frac{X_{kn} - X_{k1}}{X_{kn} - X_{ki^*}}\kappa + \frac{X_{ki^*} - X_{k1}}{X_{kn} - X_{ki^*}}\frac{1}{8} \\
a &= \frac{X_{kn} - X_{k1}}{X_{kn} - X_{ki^*}}\kappa + \frac{X_{kn} - X_{k1}}{X_{kn} - X_{ki^*}}\frac{1}{8}
\end{aligned}$$

and we verify two facts: first that the KKT stationarity is satisfied and second, that $\|\boldsymbol{u}\|_1 < 1$ with high probability. Our claim is proved immediately by combining these two facts.

Because $\widehat{r}$ and $\boldsymbol{u}$ are both centered vectors, $\overline{\Delta}_k^{\mathsf{T}}\widehat{r} = \Delta_k^{\mathsf{T}}\widehat{r}$ and likewise for $\boldsymbol{u}$. Therefore, we need only show that for $j = 1$, $\lambda\left[\Delta_k^{\mathsf{T}}\boldsymbol{u}\right]_j = \left[\frac{1}{n}\Delta_k^{\mathsf{T}}\widehat{r}\right]_j$ and that for $j = 2, ..., n-1$, $\lambda\left[\Delta_k^{\mathsf{T}}\boldsymbol{u}\right]_j \geq \left[\frac{1}{n}\Delta_k^{\mathsf{T}}\widehat{r}\right]_j$.

With our explicitly defined form of $\boldsymbol{u}$, we can characterize

$$\left[\Delta_k^{\mathsf{T}}\boldsymbol{u}\right]_j = \begin{cases} (-a + a')(X_{ki^*} - X_{kj}) + a'(X_{kn} - X_{ki^*}) & \text{if } j \leq i^* \\ a'(X_{kn} - X_{kj}) & \text{if } j \geq i^* \end{cases} \tag{8.2}$$

It is straightforward to check that $\left[\lambda\Delta_k^\mathsf{T}\boldsymbol{u}\right]_1 = \lambda\kappa = \frac{1}{n}\left[\Delta_k^\mathsf{T}\widehat{r}\right]_1$.

To check that other rows of stationarity condition holds, we characterize $[\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_j$:

$$\begin{aligned}
[\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_j &= \frac{1}{n}\sum_{i>j}(X_{ki} - X_{kj})\widehat{r}_i \\
&= \frac{1}{n}\sum_{i>j}\sum_{j<i'\le i}\mathsf{gap}_{i'}\widehat{r}_i \\
&= \frac{1}{n}\sum_{i'>j}\mathsf{gap}_{i'}\sum_{i\ge i'}\widehat{r}_i \\
&= \frac{1}{n}\sum_{i'>j}\mathsf{gap}_{i'}\mathbf{1}_{i':n}^\mathsf{T}\widehat{r}
\end{aligned}$$

where we denote $\mathsf{gap}_{i'} = X_{ki'} - X_{k(i'-1)}$.

We pause for a second here to give a summary of our proof strategy. We leverage two critical observations: first, any two adjacent coordinates in the vector $\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}$ cannot differ by too much. Second, we defined a, a' such that $-a + a' = -\frac{1}{8}$ so that $\lambda\Delta_k^\mathsf{T}\boldsymbol{u}$ is a sequence that strictly increases in the first half (for coordinates in $\{1, ..., i^*\}$) and strictly decreases in the second half.

We know $\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}$ and $\lambda\Delta_k^\mathsf{T}\boldsymbol{u}$ are equal in the first coordinate. We will show that the second sequence increases faster than the first sequence which will imply that the second sequence is larger than the first in the first half of the coordinates. We will then work similarly but backwards for the second half.

Following our strategy, We first compare $[\lambda\Delta_k^\mathsf{T}\boldsymbol{u}]_j$ and $[\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_j$ for $j = 1, ..., i^* - 1$.

For $j = 1, ..., i^* - 1$, we have that

$$\begin{aligned}
\lambda[\Delta_k^\mathsf{T}\boldsymbol{u}]_{j+1} - \lambda[\Delta_k^\mathsf{T}\boldsymbol{u}]_j &= \lambda(a - a')\mathsf{gap}_{j+1} \\
&\ge -\mathsf{gap}_{j+1}\frac{1}{n}\mathbf{1}_{(j+1):n}^\mathsf{T}\widehat{r} \\
&= [\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_{j+1} - [\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_j
\end{aligned}$$

The inequality follows because $a - a' = \frac{1}{8}$ and thus $\lambda(a - a') \ge \left|\frac{1}{n}\mathbf{1}_{(j+1):n}^\mathsf{T}\widehat{r}\right|$. Therefore, for all $j = 1, ..., i^*$:

$$[\lambda\Delta_k^\mathsf{T}\boldsymbol{u}]_j \ge [\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_j$$

For $j \ge i^*$, we start our comparison from $j = n - 1$. First, we claim that $a' > \frac{1}{32}$. To prove this claim, note that

$$|\kappa| = \left|\frac{1}{\lambda n}\sum_{i'>1}\mathsf{gap}_{i'}\mathbf{1}_{i':n}^\mathsf{T}\widehat{r}\right| \le \frac{1}{X_{kn} - X_{k1}}\frac{1}{32}\sum_{i'>1}\mathsf{gap}_{i'} = \frac{1}{32} \tag{8.3}$$

and that

$$\frac{X_{kn} - X_{ki^*}}{X_{kn} - X_{k1}} = \frac{X_{kn} - X_{k(i^*+1)} + X_{k(i^*+1)} - X_{ki^*}}{X_{kn} - X_{k1}} \leq \frac{1}{2} + \frac{1}{16}$$

where the inequality follows because we had assumed that $\frac{X_{k(i+1)} - X_{ki}}{X_{k(n)} - X_{k(1)}} \leq \frac{1}{16}$ for all $i = 1, ..., n-1$.

So, we have

$$\begin{aligned} a' &= \frac{X_{kn} - X_{k1}}{X_{kn} - X_{ki^*}}\kappa + \frac{X_{ki^*} - X_{k1}}{X_{kn} - X_{ki^*}}\frac{1}{8} \\ &= \frac{X_{kn} - X_{k1}}{X_{kn} - X_{ki^*}}\left(\kappa + \frac{X_{ki^*} - X_{k1}}{X_{kn} - X_{k1}}\frac{1}{8}\right) \\ &\geq \frac{X_{kn} - X_{k1}}{X_{kn} - X_{ki^*}}\left(-\frac{1}{32} + (\frac{1}{2} - \frac{1}{16})\frac{1}{8}\right) \\ &\geq \frac{1}{1/2 + 1/16}\left(-\frac{1}{32} + (\frac{1}{2} - \frac{1}{16})\frac{1}{8}\right) \\ &\geq \frac{1}{32} \end{aligned}$$

In the first inequality of the above derivation, we used the fact that $\frac{X_{ki^*} - X_{k1}}{X_{kn} - X_{k1}} \leq \frac{1}{2} - \frac{1}{16}$. In the second inequality, we used the fact that the quantity inside the parentheses is positive and $\frac{X_{kn} - X_{k1}}{X_{kn} - X_{ki^*}} \geq \frac{1}{1/2+1/16}$.

Now consider $j = n-1$.

$$[\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_{n-1} = \frac{1}{n}\mathsf{gap}_n\widehat{r}_n \leq \mathsf{gap}_n\frac{\lambda}{32} \leq \lambda\mathsf{gap}_n a' = \lambda[\Delta_k^\mathsf{T}\boldsymbol{u}]_{n-1}$$

For $j = i^*, ..., n-2$, we have that

$$\begin{aligned} \lambda[\Delta_k^\mathsf{T}\boldsymbol{u}]_j - \lambda[\Delta_k^\mathsf{T}\boldsymbol{u}]_{j+1} &= \lambda a'\mathsf{gap}_{j+1} \\ &\geq \mathsf{gap}_{j+1}\frac{1}{n}\mathbf{1}_{(j+1):n}^\mathsf{T}\widehat{r} \\ &\geq [\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_j - [\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_{j+1} \end{aligned}$$

Therefore, for $j = i^*, ..., n-2$,

$$\lambda[\Delta_k^\mathsf{T}\boldsymbol{u}]_j \geq \frac{1}{n}[\Delta_k^\mathsf{T}\widehat{r}]_j$$

We conclude then that $\lambda[\Delta_k^\mathsf{T}\boldsymbol{u}]_j \geq [\frac{1}{n}\Delta_k^\mathsf{T}\widehat{r}]_j$ for all $j = 2, ..., n-1$.

We have thus verified that the stationarity equations hold and now will bound $\|\boldsymbol{u}\|_1$.

$$\|\boldsymbol{u}\|_1 = |a - a'| + a + a' \leq \frac{1}{8} + 2a \leq \frac{1}{8} + 4|\kappa| + \frac{1}{2} \leq \frac{1}{8} + \frac{1}{8} + \frac{1}{2} < 1$$

In the third inequality, we used the fact that $|\kappa| \leq \frac{1}{32}$.

We have thus proven that there exists one solution $\{\widehat{d}_k\}_{k=1,\dots,p}$ such that $\widehat{d}_k = 0$ for all $k \in S^c$. Furthermore, we have shown that the subgradient variables $\boldsymbol{u}_k$ of the solution $\{\widehat{d}_k\}$ can be chosen such that $\|\boldsymbol{u}_k\|_1 < 1$ for all $k \in S^c$.

We now prove that if $\{\widehat{d}'_k\}_{k=1,\dots,p}$ is another solution, then it must be that $\widehat{d}'_k = 0$ for all $k \in S^c$ as well. We first claim that $\sum_{k=1}^p \overline{\Delta}_k \widehat{d}_k = \sum_{k=1}^p \overline{\Delta}_k \widehat{d}'_k$. If this were not true, then a convex combination of $\widehat{d}_k, \widehat{d}'_k$ would achieve a strictly lower objective on the quadratic term. More precisely, let $\zeta \in [0,1]$. If $\sum_{k=1}^p \overline{\Delta}_k \widehat{d}'_k \neq \sum_{k=1}^p \overline{\Delta}_k \widehat{d}_k$, then $\|Y - \sum_{k=1}^p \overline{\Delta}_k\big(\widehat{d}_k + \zeta(\widehat{d}'_k - \widehat{d}_k)\big)\|_2^2$ is strongly convex as a function of ν. Thus, it cannot be that $\widehat{d}_k$ and $\widehat{d}'_k$ both achieve optimal objective, and we have reached a contradiction.

Now, we look at the stationarity condition for both $\{\widehat{d}_k\}$ and $\{\widehat{d}'_k\}$. Let $\boldsymbol{u}_k \in \partial\|\overline{\Delta}_k \widehat{d}_k\|_\infty$ and let $\boldsymbol{u}'_k \in \partial\|\overline{\Delta}_k \widehat{d}'_k\|_\infty$ be the two sets of subgradients. Let $\{\nu_{ki}\}$ and $\{\nu'_{ki}\}$ be the two sets of positivity dual variables, for $k = 1, .., p$ and $i = 1, ...n-1$. Note that since there is no positivity constraint on d_{k1}, we let $\nu_{k1} = 0$ always.

Let us define $\overline{\Delta}$, a $n \times p(n-1)$ matrix, to denote the column-wise concatenation of $\{\overline{\Delta}_k\}_k$ and $\widehat{d}$, a $p(n-1)$ dimensional vector, to denote the concatenation of $\{\widehat{d}_k\}_k$. With this notation, we can express $\sum_{k=1}^p \overline{\Delta}_k \widehat{d}_k = \overline{\Delta}\widehat{d}$.

Since both solutions $(\widehat{d}, \boldsymbol{u}, \nu)$ and $(\widehat{d}', \boldsymbol{u}', \nu')$ must satisfy the stationarity condition, we have that

$$\overline{\Delta}^\mathsf{T}(Y - \overline{\Delta}\widehat{d}) + \lambda \sum_{k=1}^p \overline{\Delta}_k^\mathsf{T} \boldsymbol{u}_k - \nu = \overline{\Delta}^\mathsf{T}(Y - \overline{\Delta}\widehat{d}') + \lambda \sum_{k=1}^p \overline{\Delta}_k^\mathsf{T} \boldsymbol{u}'_k - \nu' = 0.$$

Multiplying both sides of the above equation by $\widehat{d}'$,

$$\widehat{d}'^\mathsf{T}\overline{\Delta}^\mathsf{T}(Y - \overline{\Delta}\widehat{d}) + \lambda \sum_{k=1}^p \widehat{d}_k'^\mathsf{T}\overline{\Delta}_k^\mathsf{T} \boldsymbol{u}_k - \widehat{d}'^\mathsf{T}\nu = \widehat{d}'^\mathsf{T}\overline{\Delta}^\mathsf{T}(Y - \overline{\Delta}\widehat{d}') + \lambda \sum_{k=1}^p \widehat{d}_k'^\mathsf{T}\overline{\Delta}_k^\mathsf{T} \boldsymbol{u}'_k - \widehat{d}'^\mathsf{T}\nu'.$$

Since $\overline{\Delta}\widehat{d}_k = \overline{\Delta}\widehat{d}$, $\widehat{d}'^\mathsf{T}\nu' = 0$ (complementary slackness), and $\widehat{d}_k'^\mathsf{T}\overline{\Delta}_k^\mathsf{T}\boldsymbol{u}'_k = \|\widehat{f}'_k\|_\infty$ (where $\widehat{f}'_k = \overline{\Delta}_k \widehat{d}'_k$), we have that

$$\lambda \sum_{k=1}^p \widehat{d}_k'^\mathsf{T}\overline{\Delta}_k^\mathsf{T} \boldsymbol{u}_k - \widehat{d}'^\mathsf{T}\nu = \lambda \sum_{k=1}^p \|\widehat{f}'_k\|_\infty.$$

On one hand, $\widehat{d}'$ is a feasible solution so $\widehat{d}'^\mathsf{T}\nu \geq 0$ and so

$$\sum_{k=1}^p \widehat{d}_k'^\mathsf{T}\overline{\Delta}_k^\mathsf{T} \boldsymbol{u}_k \geq \sum_{k=1}^p \|\widehat{f}'_k\|_\infty.$$

On the other hand, by Hölder's inequality,

$$\sum_{k=1}^p \widehat{d}_k'^\mathsf{T}\overline{\Delta}_k^\mathsf{T} \boldsymbol{u}_k \leq \sum_{k=1}^p \|\widehat{f}'_k\|_\infty \|\boldsymbol{u}_k\|_1.$$

Since $\boldsymbol{u}_k$ can be chosen so that $\|\boldsymbol{u}_k\|_1 < 1$ for all $k \in S^c$, we would get a contradiction if $\|\widehat{f}'_k\|_\infty > 0$ for some $k \in S^c$. We thus conclude that $\widehat{d}'$ must follow the same sparsity pattern.

The second item in the theorem concerning optimization 4.10 is proven in exactly the same way. The Lagrangian of optimization 4.10 is

$$\mathcal{L}_{\text{cave}}(d_k, \nu_k) = \frac{1}{2n}\left\|\widehat{r} - \overline{\Delta}_k d_k\right\|_2^2 + \lambda\|\overline{\Delta}_k d_k\|_\infty + \sum_{k=1}^{p}\sum_{i=2}^{n-1}\nu_{ki}d_{ki}.$$

with $\nu_{ki} \geq 0$. The same reasoning applies to show that $\widehat{d}_k = 0$ satisfies KKT conditions sufficient for optimality. □

8.2. *Proof of False Positive Control.* We note that in the following analysis the symbols c, C represent absolute constants. We will often abuse notation and "absorb" new absolute constants into c, C; the actual value of c, C could thus vary from line to line. We first restate the theorem for convenience.

THEOREM 8.2. *Suppose assumptions A1-A5 hold. Define* $\tilde{\sigma} \equiv \max(\sigma, B)$. *Suppose that* $p \leq O\big(\exp(cn)\big)$ *and* $n \geq C$ *for some constants* $0 < c < 1$ *and* C. *Define* $\mathsf{range}_k = X_{k\pi_k(n)} - X_{k\pi_k(1)}$.

If $\lambda_n \geq 2(8 \cdot 32)s\tilde{\sigma}\sqrt{\frac{1}{n}\log^2 np}$ *then, with probability at least* $1 - \frac{24}{n}$, *for all* $k \in S^c$, *and for all* $i' = 1, ..., n$

$$\lambda_n > \mathsf{range}_k\left|\frac{32}{n}\widehat{r}^{\mathsf{T}}\mathbf{1}_{\pi_k(i':n)}\right|$$

and $\max_{i'} \frac{X_{k\pi_k(i'+1)} - X_{k\pi_k(i')}}{\mathsf{range}_k} \leq \frac{1}{16}$ *and* $\mathsf{range}_k \geq 1$.

Therefore, for all $k \in S^c$, *both the AC solution* $\widehat{f}_k$ *from optimization 4.8, and the DC solution* $\widehat{g}_k$ *from optimization 4.10 are zero.*

PROOF. The key is to note that $\widehat{r}$ and $\Delta_{k,j}$ are independent for all $k \in S^c, j = 1, ..., n$ because $\widehat{r}$ is only dependent on X_S.

Fix j and i. Then $\widehat{r}^{\mathsf{T}}\mathbf{1}_{\pi_k(i':n)}$ is the sum of $n - i' + 1$ random coordinates of $\widehat{r}$. We will use Serfling's theorem on the concentration of measure of sampling without replacement (Corollary 8.2). We must first bound $\|\widehat{r}\|_\infty$ and $\frac{1}{n}\sum_{i=1}^n \widehat{r}_i$ before we can use Serfling's results however.

Step 1: *Bounding* $\|\widehat{r}\|_\infty$. We have $\widehat{r}_i = f_0(x_i) + w_i - \widehat{f}(x_i)$ where $\widehat{f}(x_i) = \sum_{k\in S}\overline{\Delta}_k\widehat{d}_k$ is the convex additive function outputted by the restricted regression. Note that both $f_0(x_i)$ and $\widehat{f}(x_i)$ are bounded by $2sB$. Because w_i is sub-Gaussian, $|w_i| \leq \sigma\sqrt{2\log\frac{2}{\delta}}$ with probability at least $1 - \delta$. By union bound across $i = 1, ..., n$, we have that $\|w\|_\infty \leq \sigma\sqrt{2\log\frac{2n}{\delta}}$ with probability at least $1 - \delta$.

Putting these observations together,

$$\begin{aligned}\|\widehat{r}\|_\infty &\le 2sB + \sigma\sqrt{2\log\frac{2n}{\delta}}) \\ &\le 4s\widetilde{\sigma}\sqrt{\log\frac{2n}{\delta}} \end{aligned} \tag{8.4}$$

with probability at least $1-\delta$, where we have defined $\widetilde{\sigma} = \max(\sigma, B)$, and assumed that $\sqrt{\log\frac{2np}{\delta}} \ge 1$. This assumption holds under the conditions in the theorem which state that $p \le \exp(cn)$ and $n \ge C$ for some small constant c and large constant C.

Step 2: *Bounding* $|\frac{1}{n}\widehat{r}^\mathsf{T}\mathbf{1}|$. We have that

$$\begin{aligned}\frac{1}{n}\widehat{r}^\mathsf{T}\mathbf{1} &= \frac{1}{n}\sum_{i=1}^n f_0(x_i) + w_i - \widehat{f}(x_i) \\ &= \frac{1}{n}\sum_{i=1}^n f_0(x_i) + w_i \quad (\widehat{f} \text{ is centered}).\end{aligned}$$

Since $|f_0(x_i)| \le sB$, the first term $|\frac{1}{n}\sum_{i=1}^n f_0(x_i)|$ is at most $sB\sqrt{\frac{2}{n}\log\frac{2}{\delta}}$ with probability at most $1-\delta$ by Hoeffding's inequality. Since w_i is sub-Gaussian, the second term $|\frac{1}{n}\sum_{i=1}^n w_i|$ is at most $\sigma\sqrt{\frac{2}{n}\log\frac{2}{\delta}}$ with probability at most $1-\delta$. Taking a union bound, we have that

$$\begin{aligned}|\frac{1}{n}\widehat{r}^\mathsf{T}\mathbf{1}| &\le sB\sqrt{\frac{2}{n}\log\frac{4}{\delta}} + \sigma\sqrt{\frac{2}{n}\log\frac{4}{\delta}} \\ &\le 4s\widetilde{\sigma}\sqrt{\frac{1}{n}\log\frac{4}{\delta}}\end{aligned} \tag{8.5}$$

with probability at least $1-\delta$.

Step 3: *Apply Serfling's theorem.* For any $k \in S^c$, Serfling's theorem states that with probability at least $1-\delta$

$$\left|\frac{1}{n}\widehat{r}^\mathsf{T}\mathbf{1}_{\pi_k(i':n)}\right| \le 2\|\widehat{r}\|_\infty\sqrt{\frac{1}{n}\log\frac{2}{\delta}} + \left|\frac{1}{n}\widehat{r}^\mathsf{T}\mathbf{1}\right|$$

We need Serfling's theorem to hold for all $k = 1, ..., p$ and $i' = 1, ..., n$. We also need the events that $\|\widehat{r}\|_\infty$ and $|\frac{1}{n}\widehat{r}^\mathsf{T}\mathbf{1}|$ are small to hold. Using a union bound, with probability at least $1-\delta$, for all k, i',

$$\begin{aligned}\left|\frac{1}{n}\widehat{r}^\mathsf{T}\mathbf{1}_{\pi_k(i':n)}\right| &\le 2\|\widehat{r}\|_\infty\sqrt{\frac{1}{n}\log\frac{6np}{\delta}} + \left|\frac{1}{n}\widehat{r}^\mathsf{T}\mathbf{1}\right| \\ &\le 4s\widetilde{\sigma}\sqrt{\log\frac{6n}{\delta}}\sqrt{\frac{1}{n}\log\frac{6np}{\delta}} + 4s\widetilde{\sigma}\sqrt{\frac{1}{n}\log\frac{12}{\delta}} \\ &\le 8s\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2\frac{12np}{\delta}}\end{aligned}$$

In the second inequality, we used equation (8.4) and equation (8.5) from steps 1 and 2 respectively. Setting $\delta = \frac{12}{n}$ gives the desired expression.

Finally, we note that $2 \geq (X_{k\pi_k(n)} - X_{k\pi_k(1)})$ since $X_k \subset [-1, 1]$. This concludes the proof for the first part of the theorem.

To prove the second and the third claims, let the interval $[-1, 1]$ be divided into 64 non-overlapping segments each of length $1/32$. Because X_k is drawn from a density with a lower bound $c_l > 0$, the probability that every segment contains some samples X_{ki}'s is at least $1-64\left(1 - \frac{1}{32}c_l\right)^n$. Let $\mathcal{E}_k$ denote the event that every segment contains some samples.

Define $\mathsf{gap}_i = X_{k\pi_k(i+1)} - X_{k\pi_k(i)}$ for $i = 1, ..., n-1$ and define $\mathsf{gap}_0 = X_{k\pi_k(1)} - (-1)$ and $\mathsf{gap}_n = 1 - X_{k\pi_k(n)}$.

If any $\mathsf{gap}_i \geq \frac{1}{16}$, then gap_i has to contain one of the segments. Therefore, under event $\mathcal{E}_k$, it must be that $\mathsf{gap}_i \leq \frac{1}{16}$ for all i.

Thus, we have that $\mathsf{range}_k \geq 2 - 1/8 \geq 1$ and that for all i,

$$\frac{X_{k\pi_k(i+1)} - X_{k\pi_k(i)}}{\mathsf{range}_k} \geq \frac{1/16}{2 - 1/8} \geq 1/16$$

Taking an union bound for each $k \in S^c$, the probability of that all $\mathcal{E}_k$ hold is at least $1 - p64\left(1 - \frac{1}{32}c_l\right)^n$.

$p64\left(1 - \frac{1}{32}c_l\right)^n = 64p\exp(-c'n)$ for some positive constants $0 < c' < 1$ dependent on c_l. Therefore, if $p \leq exp(cn)$ for some $0 < c < c'$ and if n is larger than some constant C, $64p\exp(-c'n) \leq 64\exp(-(c' - c'')n) \leq \frac{12}{n}$.

Taking an union bound with the event that λ_n upper bounds the partial sums of $\widehat{r}$ and we establish the claim.

□

8.3. *Proof of False Negative Control.* We begin by introducing some notation.

8.3.1. *Notation.* If $f : \mathbb{R}^s \to \mathbb{R}$, we define $\|f\|_P \equiv \mathbb{E}f(X)^2$. Given samples $X_1, ..., X_n$, we denote $\|f\|_n \equiv \frac{1}{n}\sum_{i=1}^n f(X_i)^2$ and $\langle f, g\rangle_n \equiv \frac{1}{n}\sum_{i=1}^n f(X_i)g(X_i)$.

Let $\mathcal{C}^1$ denote the set of univariate convex functions supported on $[-1, 1]$. Let $\mathcal{C}^1_B \equiv \{f \in \mathcal{C}^1 : \|f\|_\infty \leq B\}$ denote the set of B-bounded univariate convex functions. Define $\mathcal{C}^s$ as the set of convex additive functions and $\mathcal{C}^s_B$ likewise as the set of convex additive functions whose components are B-bounded:

$$\mathcal{C}^s \equiv \{f : f = \sum_{k=1}^s f_k,\ f_k \in \mathcal{C}^1\}$$

$$\mathcal{C}^s_B \equiv \{f \in \mathcal{C}^s : f = \sum_{k=1}^s f_k,\ \|f_k\|_\infty \leq B\}.$$

Let $f^*(x) = \sum_{k=1}^s f_k^*(x_k)$ be the population risk minimizer:

$$f^* = \arg\min_{f \in \mathcal{C}^s} \|f_0 - f^*\|_P^2$$

We let sB be an upper bound on $\|f_0\|_\infty$ and B be an upper bound on $\|f_k^*\|_\infty$. It follows that $\|f^*\|_\infty \le sB$.

We define $\widehat{f}$ as the empirical risk minimizer:

$$\widehat{f} = \arg\min \left\{ \|y - f\|_n^2 + \lambda \sum_{k=1}^{s} \|f_k\|_\infty \,:\, f \in \mathcal{C}_B^s,\, \mathbf{1}_n^\mathsf{T} f_k = 0 \right\}$$

For $k \in \{1, ..., s\}$, define g_k^* to be the decoupled concave population risk minimizer

$$g_k^* \equiv \arg\min_{g_k \in \text{-}\mathcal{C}^1} \|f_0 - f^* - g_k\|_P^2.$$

In our proof, we will analyze g_k^* for each k such that $f_k^* = 0$. Likewise, we define the empirical version:

$$\widehat{g}_k \equiv \arg\min \left\{ \|f_0 - \widehat{f} - g_k\|_n^2 \,:\, g_k \in \text{-}\mathcal{C}_B^1\,, \mathbf{1}_n^\mathsf{T} g_k = 0 \right\}.$$

By the definition of the AC/DC procedure, $\widehat{g}_k$ is defined only for an index k that has zero as the convex additive approximation.

8.3.2. *Proof.* By additive faithfulness of the AC/DC procedure, it is known that $f_k^* \neq 0$ or $g_k^* \neq 0$ for all $k \in S$. Our argument will be to show that the risk of the AC/DC estimators $\widehat{f}, \widehat{g}$ tends to the risk of the population optimal functions f^*, g^*:

$$\begin{aligned} \|f_0 - \widehat{f}\|_P^2 &= \|f_0 - f^*\|_P^2 + \mathrm{err}_+(n) \\ \|f_0 - f^* - \widehat{g}_k\|_P^2 &= \|f_0 - f^* - g_k^*\|_P^2 + \mathrm{err}_-(n) \quad \text{for all } k \in S \text{ where } f_k^* = 0, \end{aligned}$$

where the estimation errors $\mathrm{err}_+(n)$ and $\mathrm{err}_-(n)$ decrease with n at some rate.

Assuming this, suppose that $\widehat{f}_k = 0$ and $f_k^* \neq 0$. Then when n is large enough such that $\mathrm{err}_+(n)$ and $\mathrm{err}_-(n)$ are smaller than α_+ and α_- defined in equation (5.3), we reach a contradiction. This is because the risk $\|f_0 - f^*\|_P$ of f^* is strictly larger by α_+ than the risk of the best approximation whose k-th component is constrained to be zero. Similarly, suppose $f_k^* = 0$ and $g_k^* \neq 0$. Then when n is large enough, $\widehat{g}_k$ must not be zero.

Theorem 8.3 and Theorem 8.4 characterize $\mathrm{err}_+(n)$ and $\mathrm{err}_-(n)$ respectively.

THEOREM 8.3. *Let $\widetilde{\sigma} \equiv \max(\sigma, B)$, and let $\widehat{f}$ be the minimizer of the restricted regression with $\lambda \le 512 s\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2 np}$. Suppose $n \ge c_1 s\sqrt{sB}$. Then with probability at least $1 - \frac{C}{n}$,*

$$\|f_0 - \widehat{f}\|_P^2 - \|f_0 - f^*\|_P^2 \le cB^2\widetilde{\sigma}\sqrt{\frac{s^5}{n^{4/5}}\log^2 np}, \tag{8.6}$$

where c_1 is an absolute constant and c, C are constants possibly dependent on b.

PROOF. Our proof proceeds in three steps. First, we bound the difference of empirical risks $\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^*\|_n^2$. Second, we bound the cross-term in the bound using a bracketing entropy argument for convex function classes. Finally, we combine the previous two steps to complete the argument.

Step 1. The function $\widehat{f}$ minimizes the penalized empricial risk by definition. We would thus like to say that the penalized empirical risk of $\widehat{f}$ is no larger than that of f^*. We cannot do a direct comparison, however, because the empirical mean $\frac{1}{n}\sum_i f_k^*(x_{ik})$ is close to, but not exactly zero. We thus have to work first with the function $f^* - \overline{f}^*$. We have that

$$\|y - \widehat{f}\|_n^2 + \lambda \sum_{k=1}^s \|\widehat{f}_k\|_\infty \leq \|y - f^* + \overline{f}^*\|_n^2 + \lambda \sum_{k=1}^s \|f_k^* - \overline{f}_k^*\|_\infty$$

Plugging in $y = f_0 + w$, we obtain

$$\begin{aligned}
\|f_0 + w - \widehat{f}\|_n^2 + \lambda \sum_{k=1}^s \Big(\|\widehat{f}_k\|_\infty - \|f_k^* - \overline{f}_k^*\|_\infty \Big) &\leq \|f_0 + w - f^* + \overline{f}^*\|_n^2 \\
\|f_0 - \widehat{f}\|_n^2 + 2\langle w, f_0 - \widehat{f}\rangle_n + \lambda \sum_{k=1}^s \Big(\|\widehat{f}_k\|_\infty - \|f_k^* - \overline{f}_k^*\|_\infty \Big) & \\
\leq \|f_0 - f^* + \overline{f}^*\|_n^2 + 2\langle w, f_0 - f^* + \overline{f}^*\rangle & \\
\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^* + \overline{f}^*\|_n^2 + \lambda \sum_{k=1}^s \Big(\|\widehat{f}_k\|_\infty - \|f_k^* - \overline{f}_k^*\|_\infty \Big) &\leq 2\langle w, \widehat{f} - f^* + \overline{f}^*\rangle.
\end{aligned}$$

The middle term can be bounded under the assumption that $\|f_k^* - \overline{f}_k^*\|_\infty \leq 2B$; thus,

$$\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^* + \overline{f}^*\|_n^2 \leq 2\langle w, \widehat{f} - f^* + \overline{f}^*\rangle + \lambda 2sB.$$

Using Lemma 8.2, we can remove $\overline{f}^*$ from the lefthand side. Thus with probability at least $1 - \delta$,

$$\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^*\|_n^2 \leq 2\langle w, \widehat{f} - f^* + \overline{f}^*\rangle + \lambda 2sB + c(sB)^2 \frac{1}{n} \log \frac{2}{\delta}. \tag{8.7}$$

Step 2. We now upper bound the cross term $2\langle w, \widehat{f} - f^* + \overline{f}^*\rangle$ using bracketing entropy.

Define $\mathcal{G} = \{f - f^* + \overline{f}^* : f \in \mathcal{C}_B^s\}$ as the set of convex additive functions centered around the function $f^* - \overline{f}^*$. By Corollary 8.3, there is an ϵ-bracketing of $\mathcal{G}$ whose size is bounded by $\log N_{[]}(2\epsilon, \mathcal{G}, L_1(P)) \leq sK^{**}\left(\frac{2sB}{\epsilon}\right)^{1/2}$, for all $\epsilon \in (0, sB\epsilon_3]$. Let us suppose condition 8.11 holds. Then, by Corollary 8.4, with probability at least $1 - \delta$, each bracketing pair (h_U, h_L) is close in $L_1(P_n)$ norm, i.e., for all (h_U, h_L), $\frac{1}{n}\sum_{i=1}^n |h_U(X_i) - h_L(X_i)| \leq 2\epsilon + sB\sqrt{\frac{sK^{**}(2sB)^{1/2}\log\frac{1}{\delta}}{2\epsilon^{1/2}n}}$. We verify at the end of the proof that condition 8.11 indeed holds.

For each $h \in \mathcal{G}$, there exists a pair (h_U, h_L) such that $h_U(X_i) - h_L(X_i) \geq h(X_i) - h_L(X_i) \geq 0$. Therefore, with probability at least $1-\delta$, uniformly for all $h \in \mathcal{G}$:

$$\frac{1}{n}\sum_{i=1}^{n}|h(X_i) - h_L(X_i)| \leq \frac{1}{n}\sum_{i=1}^{n}|h_U(X_i) - h_L(X_i)| \leq 2\epsilon + (sB)\sqrt{\frac{sK^{**}(2sB)^{1/2}\log\frac{1}{\delta}}{2\epsilon^{1/2}n}}.$$

We denote $\epsilon_{n,\delta} \equiv (sB)\sqrt{\frac{sK^{**}(2sB)^{1/2}\log\frac{1}{\delta}}{2\epsilon^{1/2}n}}$. Let $\mathcal{E}_{[]}$ denote the event that for each $h \in \mathcal{G}$, there exists h_L in the ϵ-bracketing such that $\|h - h_L\|_{L_{P_n}} \leq 2\epsilon + \epsilon_{n,\delta}$. Then $\mathcal{E}_{[]}$ has probability at most $1-\delta$ as shown.

Let $\mathcal{E}_{\|w\|_\infty}$ denote the event that $\|w\|_\infty \leq \sigma\sqrt{2\log\frac{2n}{\delta}}$. Then $\mathcal{E}_{\|w\|_\infty}$ has probability at most $1-\delta$. We now take an union bound over $\mathcal{E}_{\|w\|_\infty}$ and $\mathcal{E}_{[]}$ and get that, with probability at most $1-2\delta$, for all h

$$|\langle w, h - h_L\rangle_n| \leq \|w\|_\infty \frac{1}{n}\sum_{i=1}^{n}|h(X_i) - h_L(X_i)| \leq \sigma\sqrt{2\log\frac{4n}{\delta}}\left(\epsilon + \epsilon_{n,2\delta}\right).$$

Because w is a sub-Gaussian random variable, we have that for any fixed vector $h_L = (h_L(X_1), ..., h_L(X_n))$, with probability at least $1-\delta$, $|\langle w, h_L\rangle_n| \leq \|h_L\|_n \sigma\sqrt{\frac{2}{n}\log\frac{2}{\delta}}$. Using another union bound, we have that the event $\sup_{h_L}|\langle w, h_L\rangle| \leq sB\sigma\sqrt{\frac{2}{n}\log\frac{2N_{[]}}{\delta}}$ has probability at most $1-\delta$.

Putting this together, we have that

$$\begin{aligned}
&|\langle w, h\rangle_n| \leq |\langle w, h_L\rangle_n| + |\langle w, h - h_L\rangle_n| \\
&|\sup_{h\in\mathcal{G}}\langle w, h\rangle_n| \leq |\sup_{h^L}\langle w, h_L\rangle_n| + \sigma\sqrt{2\log\frac{2n}{\delta}}(2\epsilon + \epsilon_{n,2\delta}) \\
&\leq sB\sigma\sqrt{2\frac{\log N_{[]} + \log\frac{1}{\delta}}{n}} + \sigma\sqrt{2\log\frac{2n}{\delta}}(2\epsilon + \epsilon_{n,\delta}) \\
&\leq sB\sigma\sqrt{2\frac{sK^{**}(2sB)^{1/2}\log\frac{1}{\delta}}{n\epsilon^{1/2}}} + \sigma\sqrt{2\log\frac{2n}{\delta}}(2\epsilon + \epsilon_{n,\delta}) \\
&\leq sB\sigma\sqrt{2\frac{sK^{**}(2sB)^{1/2}\log\frac{1}{\delta}}{n\epsilon^{1/2}}} + 2\sigma\sqrt{2\log\frac{2n}{\delta}}\epsilon + sB\sigma\sqrt{2\frac{sK^{**}(2sB)^{1/2}\log\frac{1}{\delta}}{n\epsilon^{1/2}}\log\frac{2n}{\delta}} \\
&\leq 2\sigma\sqrt{2\log\frac{2n}{\delta}}\epsilon + 2sB\sigma\sqrt{\frac{sK^{**}(2sB)^{1/2}\log^2\frac{2n}{\delta}}{n\epsilon^{1/2}}}.
\end{aligned}$$

To balance the two terms, we set $\epsilon = sB\sqrt{\frac{(sK^{**}(sB)^{1/2})^{4/5}}{n^{4/5}}}$. It is easy to verify that if $n \geq c_1 s\sqrt{sB}$ for some absolute constant c_1, then $\epsilon \in (0, sB\epsilon_3]$ for some absolute constant ϵ_3 as required by the bracketing number results (Corollary 8.3). Furthermore, conditions (8.11) also hold.

In summary, we have that probability at least $1-\delta$,

$$|\sup_{h\in\mathcal{G}}\langle w, h\rangle| \leq csB\sigma\sqrt{\frac{s^{6/5}B^{2/5}\log^2\frac{Cn}{\delta}}{n^{4/5}}} \leq csB\sigma\sqrt{\frac{s(sB)^{1/2}\log^2\frac{Cn}{\delta}}{n^{4/5}}}$$

where we absorbed K^{**} into the constant c and the union bound multipliers into the constant C.

Plugging this result into equation (8.7) we get that, with probability at least $1-2\delta$,

$$\begin{aligned}
\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^*\|_n^2 &\leq csB\sigma\sqrt{\frac{s(sB)^{1/2}\log^2\frac{Cn}{\delta}}{n^{4/5}}} + \lambda 2sB + c(sB)^2\frac{1}{n}\log\frac{2}{\delta} \\
\|f_0 - \widehat{f}\|_n^2 - \|f_0 - f^*\|_n^2 &\leq csB\sigma\sqrt{\frac{s(sB)^{1/2}\log^2\frac{Cn}{\delta}}{n^{4/5}}} + \lambda 2sB \\
&\leq cB\sigma\sqrt{\frac{s^4B^{1/2}}{n^{4/5}}\log^2\frac{Cn}{\delta}} + \lambda 2sB \qquad (8.8)
\end{aligned}$$

Step 3. Continuing from equation (8.8), we use Lemma 8.1 and another union bound to obtain that, with probability at least $1-3\delta$,

$$\begin{aligned}
\|f_0 - \widehat{f}\|_P^2 - \|f_0 - f^*\|_P^2 &\leq cB^2\sigma\sqrt{\frac{s^4}{n^{4/5}}\log^2\frac{Cn}{\delta}} + \lambda 2sB + cB^3\sqrt{\frac{s^5}{n^{4/5}}\log\frac{2}{\delta}} \\
&\leq cB^2\widetilde{\sigma}\sqrt{\frac{s^5}{n^{4/5}}\log^2\frac{Cn}{\delta}} + \lambda 2sB
\end{aligned}$$

Substituting in $\lambda \leq 512s\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2 np}$ and $\delta = \frac{C}{n}$ we obtain the statement of the theorem. □

THEOREM 8.4. *Let $\widehat{g}_k$ denote the minimizer of the concave postprocessing step with $\lambda_n \leq 512s\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2 np}$. Let $\widetilde{\sigma} \equiv \max(\sigma, B)$. Suppose n is sufficiently large that $\frac{n^{4/5}}{\log^2 np} \geq c'B^4\widetilde{\sigma}^2s^5$ where $c' \geq 1$ is a constant. Then with probability at least $1 - \frac{C}{n}$, for all $k = 1, ..., s$,*

$$\|f_0 - f^* - \widehat{g}_k\|_P^2 - \|f_0 - f^* - g_k^*\|_P^2 \leq cB^2\widetilde{\sigma}^{1/2}\sqrt[4]{\frac{s^5}{n^{4/5}}\log^2 np}.$$

PROOF. This proof is similar to that of Theorem 8.3; it requires a few more steps because $\widehat{g}_k$ is fitted against $f_0 - \widehat{f}$ instead of $f_0 - f^*$. We start with the following decomposition:

$$\begin{aligned}
\|f_0 - f^* - \widehat{g}_k\|_P^2 - \|f_0 - f^* - g_k^*\|_P^2 = &\underbrace{\|f_0 - \widehat{f} - \widehat{g}_k\|_P^2 - \|f_0 - \widehat{f} - g_k^*\|_P^2}_{\text{term 1}} + \\
&\underbrace{\|f_0 - f^* - \widehat{g}_k\|_P^2 - \|f_0 - \widehat{f} - \widehat{g}_k\|_P^2}_{\text{term 2}} + \\
&\underbrace{\|f_0 - \widehat{f} - g_k^*\|_P^2 - \|f_0 - f^* - g_k^*\|_P^2}_{\text{term 3}}. \qquad (8.9)
\end{aligned}$$

We now bound each of the terms. The proof proceeds almost identically to that of Theorem 8.3, because convex and concave functions have the same bracketing number.

Step 1. To bound term 1, we start from the definition of $\widehat{g}_k$ and obtain

$$\begin{aligned}\|y-\widehat{f}-\widehat{g}_k\|_n^2+\lambda_n\|\widehat{g}\|_\infty &\le \|y-\widehat{f}-g_k^*\|_n^2+\lambda_n\|g^*\|_\infty\\ \|y-\widehat{f}-\widehat{g}_k\|_n^2 &\le \|y-\widehat{f}-g_k^*\|_n^2+\lambda_n 2B\end{aligned}$$

$$\begin{aligned}\|f_0-\widehat{f}-\widehat{g}_k+w\|_n^2 &\le \|f_0-\widehat{f}-g_k^*+w\|_n^2+\lambda_n 2B\\ \|f_0-\widehat{f}-\widehat{g}_k\|_n^2-\|f_0-\widehat{f}-g_k^*\|_n^2 &\le 2\langle w,\widehat{g}_k-g_k^*\rangle_n+\lambda_n 2B.\end{aligned}$$

Using the same bracketing analysis as in Step 2 of the proof of Theorem 8.3 but setting $s=1$, we have, with probability at least $1-\delta$,

$$\|f_0-\widehat{f}-\widehat{g}_k\|_n^2-\|f_0-\widehat{f}-g_k^*\|_n^2 \le cB^2\sigma\sqrt{\frac{1}{n^{4/5}}\log\frac{C}{\delta}}+\lambda_n 2B.$$

The condition $n\ge c_1 s\sqrt{sB}$ in the proof of Theorem 8.3 is satisfied here because we assume that $n^{4/5}\ge c_1 B^4\widetilde{\sigma}^2 s^5\log^2 np$ in the statement of the theorem. Using the uniform convergence result of Lemma 8.1, with probability at least $1-\delta$,

$$\begin{aligned}\|f_0-\widehat{f}-\widehat{g}_k\|_P^2-\|f_0-\widehat{f}-g_k^*\|_P^2 &\le cB^2\sigma\sqrt{\frac{1}{n}\log\frac{Cn}{\delta}}+\lambda_n 2B+cB^3\sqrt{\frac{s^5}{n^{4/5}}\log\frac{2}{\delta}}\\ &\le cB^2\widetilde{\sigma}\sqrt{\frac{s^5}{n^{4/5}}\log\frac{C}{\delta}}+\lambda_n 2B\end{aligned}$$

Finally, plugging in $\lambda_n\le 9s\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2 np}$, we obtain

$$\begin{aligned}\|f_0-\widehat{f}-\widehat{g}_k\|_P^2-\|f_0-\widehat{f}-g_k^*\|_P^2 &\le cB^2\widetilde{\sigma}\sqrt{\frac{s^5}{n^{4/5}}\log\frac{C}{\delta}}+2sB\widetilde{\sigma}\sqrt{\frac{1}{n}\log^2 np}\\ \|f_0-\widehat{f}-\widehat{g}_k\|_P^2-\|f_0-\widehat{f}-g_k^*\|_P^2 &\le cB^2\widetilde{\sigma}\sqrt{\frac{s^5}{n^{4/5}}\log^2\frac{Cnp}{\delta}}\end{aligned}$$

with probability at least $1-\delta$.

Step 2. We now bound term 3.

$$\begin{aligned}\|f_0-\widehat{f}-g_k^*\|_P^2-\|f_0-f^*-g_k^*\|_P^2 &\le \|f_0-\widehat{f}\|_P^2-\|f_0-f^*\|_P^2-2\langle f_0-\widehat{f},g_k^*\rangle_P+2\langle f_0-f^*,g_k^*\rangle_P\\ &\le cB^2\widetilde{\sigma}\sqrt{\frac{s^5}{n^{4/5}}\log^2 np}+2|\langle\widehat{f}-f^*,g_k^*\rangle_P|\\ &\le cB^2\widetilde{\sigma}\sqrt{\frac{s^5}{n^{4/5}}\log^2 np}+2\|\widehat{f}-f^*\|_P\|g_k^*\|_P\\ &\le cB^2\widetilde{\sigma}\sqrt{\frac{s^5}{n^{4/5}}\log^2 np}+cB\sqrt{B^2\widetilde{\sigma}\sqrt{\frac{s^5}{n^{4/5}}\log^2 np}}\\ &\le cB^2\widetilde{\sigma}^{1/2}\sqrt[4]{\frac{s^5}{n^{4/5}}\log^2 np}\end{aligned}$$

with probability at least $1 - \frac{C}{n}$, by Theorem 8.3. To obtain the fourth inequality, we used the fact that $\|\widehat{f} - f^*\|^2 \leq \|f_0 - \widehat{f}\|_P^2 - \|f_0 - f^*\|_P^2$, which follows from the fact that f^* is the projection of f_0 onto the set of additive convex functions and the set of additive convex functions is convex itself. The last inequality holds because there is a condition in the theorem which states n is large enough such that $B^2\widetilde{\sigma}\sqrt{\frac{s^5}{n^{4/5}}\log^2 np} \leq 1$. The same derivation and the same bound likewise holds for term 2.

Step 3. Collecting the results and plugging them into equation (8.9), we have, with probability at least $1 - 2\delta$:

$$\|f_0 - f^* - \widehat{g}_k\|_P^2 - \|f_0 - f^* - g_k^*\|_P^2 \leq cB^2\widetilde{\sigma}^{1/2}\sqrt[4]{\frac{s^5}{n^{4/5}}\log^2\frac{4np}{\delta}}$$

Taking a union bound across the s dimensions completes the result. □

8.3.3. *Support Lemmas.*

LEMMA 8.1. *Suppose $n \geq c_1 s\sqrt{sB}$ for some absolute constant c_1. Then, with probability at least $1 - \delta$:*

$$\sup_{f \in \mathcal{C}_B^s} \left| \|f_0 - f\|_n^2 - \|f_0 - f\|_P^2 \right| \leq cB^3\sqrt{\frac{s^5}{n^{4/5}}\log\frac{2}{\delta}}$$

where c_1 is some absolute constant and c, C are constants possibly dependent on b.

PROOF. Let $\mathcal{G}$ denote the off-centered set of convex functions, that is, $\mathcal{G} \equiv \mathcal{C}^s - f_0$. Note that if $h \in \mathcal{G}$, then $\|h\|_\infty = \|f_0 - f\|_\infty \leq 4sB$. There exists an ϵ-bracketing of $\mathcal{G}$, and by Corollary 8.3, the bracketing has size at most $\log N_{[]}(2\epsilon, \mathcal{C}^s, L_1(P)) \leq sK^{**}\left(\frac{2sB}{\epsilon}\right)^{1/2}$. By Corollary 8.4, we know that with probability at least $1 - \delta$, $\|h_U - h_L\|_{L_1(P_n)} \leq \epsilon + \epsilon_{n,\delta}$ for all pairs (h_U, h_L) in the bracketing, where $\epsilon_{n,\delta} = sB\sqrt{\frac{K^{**}(2sB)^{1/2}\log\frac{2}{\delta}}{2\epsilon^{1/2}n}}$. Corollary 8.4 necessitates $\epsilon \in (0, sB\epsilon_3]$ for some absolute constant ϵ_3; we will verify that this condition holds for large enough n when we set the actual value of ϵ. For a particular function $h \in \mathcal{G}$, we can construct $\psi_L \equiv \min(|h_U|, |h_L|)$ and $\psi_U \equiv \max(|h_U|, |h_L|)$ so that

$$\psi_L^2 \leq h^2 \leq \psi_U^2.$$

We can then bound the $L_1(P)$ norm of $\psi_U^2 - \psi_L^2$ as

$$\begin{aligned}\int(\psi_U^2(x) - \psi_L^2(x))p(x)dx &\leq \int |h_U^2(x) - h_L^2(x)|p(x)dx \\ &\leq \int |h_U(x) - h_L(x)|\,|h_U(x) + h_L(x)|p(x)dx \\ &\leq 2sB\epsilon\end{aligned}$$

Now we can bound $\|h\|_n^2 - \|h\|_P^2$ as

$$\frac{1}{n}\sum_{i=1}^n \psi_L(X_i)^2 - \mathbb{E}\psi_U(X)^2 \leq \|h\|_n^2 - \|h\|_P^2 \leq \frac{1}{n}\sum_{i=1}^n \psi_U(X_i)^2 - \mathbb{E}\psi_L(X)^2 \tag{8.10}$$

Since $\psi_L(X_i)^2$ and $\psi_U(X_i)^2$ are bounded random variables with upper bound $(sB)^2$, Hoeffding's inequality and union bound give that, with probability at least $1-\delta$,, for all ψ_L (and likewise ψ_U)

$$\left|\frac{1}{n}\sum_{i=1}^n \psi_L(X_i)^2 - \mathbb{E}\psi_L(X)^2\right| \leq (sB)^2\sqrt{\frac{sK^{**}(sB)^{1/2}\log\frac{2}{\delta}}{\epsilon^{1/2}n}}$$

Plugging this into equation (8.10) above, we have that:

$$\begin{aligned}&\mathbb{E}\psi_L(X)^2 - \mathbb{E}\psi_U(X)^2 - (sB)^2\sqrt{\frac{sK^{**}(sB)^{1/2}\log\frac{2}{\delta}}{\epsilon^{1/2}n}}\\ &\leq \|h\|_n^2 - \|h\|_P^2 \leq \mathbb{E}\psi_U(X)^2 - \mathbb{E}\psi_L(X)^2 + (sB)^2\sqrt{\frac{sK^{**}(sB)^{1/2}\log\frac{2}{\delta}}{\epsilon^{1/2}n}}.\end{aligned}$$

Using the $L_1(P)$ norm of $\psi_U^2 - \psi_L^2$ result, we have

$$-sB\epsilon - (sB)^2\sqrt{\frac{sK^{**}(sB)^{1/2}\log\frac{2}{\delta}}{\epsilon^{1/2}n}} \leq \|h\|_n^2 - \|h\|_P^2 \leq sB\epsilon + (sB)^2\sqrt{\frac{sK^{**}(sB)^{1/2}\log\frac{2}{\delta}}{\epsilon^{1/2}n}}$$

We balance the terms by choosing $\epsilon = \left(\frac{(sB)^2 sK^{**}(sB)^{1/2}}{n}\right)^{2/5}$. One can easily verify that $\epsilon \leq sB\epsilon_3$ condition needed by Corollary 8.4 is satisfied when $n \geq c_1 s\sqrt{sB}$ for some absolute constant c_1. We have then that, with probability at least $1-\delta$,

$$\sup_{h\in\mathcal{G}}\left|\|h\|_n^2 - \|h\|_P^2\right| \leq cB^3\sqrt{\frac{s^5\log\frac{2}{\delta}}{n^{4/5}}}$$

The theorem follows immediately.

□

LEMMA 8.2. *Let f_0 and f^* be defined as in Section 8.3.1. Define $\overline{f}^* = \frac{1}{n}\sum_{i=1}^n f^*(X_i)$. Then, with probability at least $1-2\delta$,*

$$\left|\|f_0 - f^*\|_n^2 - \|f_0 - f^* + \overline{f}^*\|_n^2\right| \leq c(sB)^2\frac{1}{n}\log\frac{4}{\delta}$$

PROOF. We decompose the empirical norm as

$$\begin{aligned}\|f_0 - f^* + \overline{f}^*\|_n^2 &= \|f_0 - f^*\|_n^2 + 2\langle f_0 - f^*, \overline{f}^*\rangle + \overline{f}^{*2}\\ &= \|f_0 - f^*\|_n^2 + 2\overline{f}^*\langle f_0 - f^*, \mathbf{1}\rangle_n + \overline{f}^{*2}\\ &= \|f_0 - f^*\|_n^2 + 2\overline{f}^*\overline{f}_0 - \overline{f}^{*2}.\end{aligned}$$

Now $\overline{f}^* = \frac{1}{n}\sum_{i=1}^n f^*(X_i)$ is the average of n bounded mean-zero random variables and therefore, with probability at least $1-\delta$, $|\overline{f}^*| \le 4sB\sqrt{\frac{1}{n}\log\frac{2}{\delta}}$. The same reasoning likewise applies to $\overline{f}_0 = \frac{1}{n}\sum_{i=1}^n f_0(X_i)$.

Taking a union bound and we have that, with probability at least $1-2\delta$,

$$|\overline{f}^*||\overline{f}_0| \le c(sB)^2\frac{1}{n}\log\frac{2}{\delta}$$
$$\overline{f}^{*2} \le c(sB)^2\frac{1}{n}\log\frac{2}{\delta}$$

Therefore, with probability at least $1-2\delta$,

$$\|f_0 - f^*\|_n^2 - c(sB)^2\frac{1}{n}\log\frac{2}{\delta} \le \|f_0 - f^* + \overline{f}^*\|_n^2 \le \|f_0 - f^*\|_n^2 + c(sB)^2\frac{1}{n}\log\frac{2}{\delta}$$

□

8.3.4. *Supporting Lemma for Theorem 3.2.*

LEMMA 8.3. *Let $f : [0,1]^p \to \mathbb{R}$ be a twice differentiable function. Suppose $p(\boldsymbol{x})$ is a density on $[0,1]^p$ such that $\partial_{x_k} p(\boldsymbol{x}_{-k} \,|\, x_k)$ and $\partial^2_{x_k} p(\boldsymbol{x}_{-k} \,|\, x_k)$ are continuous as functions of x_k. Let $\phi(\boldsymbol{x}_{-k})$ be a continuous function not dependent on x_k.*

Then, $h_k^*(x_k) \equiv \mathbb{E}[f(X) - \phi(X_{-k}) \,|\, x_k]$ is twice differentiable and has a second derivative lower bounded away from $-\infty$.

PROOF. We can write

$$h_k^*(x_k) = \int_{\boldsymbol{x}_{-k}} \big(f(\boldsymbol{x}) - \phi(\boldsymbol{x}_{-k})\big)p(\boldsymbol{x}_{-k} \,|\, x_k)d\boldsymbol{x}_{-k}$$

The integrand is bounded because it is a sum-product of continuous functions over a compact set. Therefore, we can differentiate under the integral and derive that

$$\partial_{x_k} h_k^*(x_k) = \int_{\boldsymbol{x}_{-k}} f'(\boldsymbol{x})p(\boldsymbol{x}_{-k} \,|\, x_k) + (f(\boldsymbol{x}) - \phi(\boldsymbol{x}_{-k})p'(\boldsymbol{x}_{-k} \,|\, x_k)d\boldsymbol{x}_{-k}$$
$$\partial^2_{x_k} h_k^*(x_k) = \int_{\boldsymbol{x}_{-k}} f''(\boldsymbol{x})p(\boldsymbol{x}_{-k} \,|\, x_k) + 2f'(\boldsymbol{x})p'(\boldsymbol{x}_{-k} \,|\, x_k) + (f(\boldsymbol{x}) - \phi(\boldsymbol{x}_{-k})p''(\boldsymbol{x}_{-k} \,|\, x_k)d\boldsymbol{x}_{-k}$$

where we have used the shorthand $f'(\boldsymbol{x}), p'(\boldsymbol{x}_{-k} \,|\, x_k)$ to denote $\partial_{x_k} f(\boldsymbol{x})$, $\partial_{x_k} p(\boldsymbol{x}_{-k} \,|\, x_k)$, etc.

This proves that $h_k^*(x_k)$ is twice-differentiable. To see that the second derivative is lower bounded, we note that $f''(\boldsymbol{x})p(\boldsymbol{x}_{-k} \,|\, x_k)$ is non-negative and the other terms in the second-derivative are all continuous functions on a compact set and thus bounded. □

LEMMA 8.4. *Let $p(\boldsymbol{x})$ be a positive density over $[0,1]^p$. Let $\phi(\boldsymbol{x}) = \sum_{j=1}^p \phi_j(x_j)$ be an additive function.*

Suppose that for all j, $\phi_j(x_j)$ is bounded and $\mathbb{E}\phi_j(X_j) = 0$. Suppose for some j, $\mathbb{E}\phi_j(X_j)^2 > 0$, then it must be that $\mathbb{E}\phi(X)^2 > 0$.

PROOF. Suppose, for sake of contradiction, that

$$\mathbb{E}\phi(X)^2 = \mathbb{E}\left(\phi_j(X_j) + \phi_{-j}(X_{-j})\right)^2 = 0.$$

Let $A_+ = \{x_j : \phi_j(x_j) \geq 0\}$. Since $\mathbb{E}\phi_j(X_j) = 0$, $\mathbb{E}\phi_j(X_j)^2 > 0$, and ϕ_j is bounded, it must be that both A_+ has probability greater than 0.

Now, define $B_+ = \{x_{-j} : \phi_{-j}(x_{-j}) \geq 0\}$. B_+ then must have probability greater than 0 as well.

Since $p(\boldsymbol{x})$ is a positive density, the set $A_+ \times B_+$ must have a positive probability. However, $\phi > 0$ on $A_+ \times B_+ \subset [0,1]^p$ which implies $\mathbb{E}\phi(X)^2 > 0$.

□

8.3.5. *Concentration of Measure.* A sub-exponential random is the square of a sub-Gaussian random variable Vershynin (2010).

PROPOSITION 8.1. *(Subexponential Concentration Vershynin (2010)) Let $X_1, ..., X_n$ be zero-mean independent subexponential random variables with subexponential scale K. Then*

$$P(|\frac{1}{n}\sum_{i=1}^{n} X_i| \geq \epsilon) \leq 2\exp\left[-cn\min\left(\frac{\epsilon^2}{K^2}, \frac{\epsilon}{K}\right)\right]$$

where $c > 0$ is an absolute constant.

For uncentered subexponential random variables, we can use the following fact. If X_i subexponential with scale K, then $X_i - \mathbb{E}[X_i]$ is also subexponential with scale at most $2K$. Restating, we can set

$$c\min\left(\frac{\epsilon^2}{K^2}, \frac{\epsilon}{K}\right) = \frac{1}{n}\log\frac{1}{\delta}.$$

Thus, with probability at least $1-\delta$, the deviation is at most

$$K\max\left(\sqrt{\frac{1}{cn}\log\frac{C}{\delta}}, \frac{1}{cn}\log\frac{C}{\delta}\right).$$

COROLLARY 8.1. *Let $W_1, ..., W_n$ be n independent sub-Gaussian random variables with sub-Gaussian scale σ. Then, for all $n > n_0$, with probability at least $1 - \frac{1}{n}$,*

$$\frac{1}{n}\sum_{i=1}^{n} W_i^2 \leq c\sigma^2.$$

PROOF. Using the subexponential concentration inequality, we know that, with probability at least $1 - \frac{1}{n}$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} W_i^2 - \mathbb{E}W^2\right| \leq \sigma^2\max\left(\sqrt{\frac{1}{cn}\log\frac{C}{\delta}}, \frac{1}{cn}\log\frac{C}{\delta}\right).$$

First, let $\delta = \frac{1}{n}$. Suppose n is large enough such that $\frac{1}{cn}\log Cn < 1$. Then, we have, with probability at least $1 - \frac{1}{n}$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} W_i^2 &\leq c\sigma^2\Big(1 + \sqrt{\frac{1}{cn}\log Cn}\Big) \\ &\leq 2c\sigma^2.\end{aligned}$$

□

8.3.6. *Sampling Without Replacement.*

LEMMA 8.5. *(Serfling (1974)) Let $x_1, ..., x_N$ be a finite list, $\overline{x} = \mu$. Let $X_1, ..., X_n$ be sampled from x without replacement.*

Let $b = \max_i x_i$ and $a = \min_i x_i$. Let $r_n = 1 - \frac{n-1}{N}$. Let $S_n = \sum_i X_i$. Then we have that

$$\mathbb{P}(S_n - n\mu \geq n\epsilon) \leq \exp\left(-2n\epsilon^2 \frac{1}{r_n(b-a)^2}\right).$$

COROLLARY 8.2. *Suppose $\mu = 0$.*

$$\mathbb{P}\left(\frac{1}{N}S_n \geq \epsilon\right) \leq \exp\left(-2N\epsilon^2 \frac{1}{(b-a)^2}\right)$$

And, by union bound, we have that

$$\mathbb{P}\left(|\frac{1}{N}S_n| \geq \epsilon\right) \leq 2\exp\left(-2N\epsilon^2 \frac{1}{(b-a)^2}\right)$$

A simple restatement is that with probability at least $1 - \delta$, the deviation $|\frac{1}{N}S_n|$ is at most $(b-a)\sqrt{\frac{1}{2N}\log\frac{2}{\delta}}$.

PROOF.

$$\mathbb{P}\left(\frac{1}{N}S_n \geq \epsilon\right) = \mathbb{P}\left(S_n \geq \frac{N}{n}n\epsilon\right) \leq \exp\left(-2n\frac{N^2}{n^2}\epsilon^2 \frac{1}{r_n(b-a)^2}\right).$$

We note that $r_n \leq 1$ always, and $n \leq N$ always. Thus,

$$\exp\left(-2n\frac{N^2}{n^2}\epsilon^2 \frac{1}{r_n(b-a)^2}\right) \leq \exp\left(-2N\epsilon^2 \frac{1}{(b-a)^2}\right)$$

completing the proof. □

8.3.7. *Bracketing Numbers for Convex Functions.*

DEFINITION 8.1. Let $\mathcal{C}$ be a set of functions. For a given ϵ and metric ρ (which we take to be L_2 or $L_2(P)$), we define a *bracketing* of $\mathcal{C}$ to be a set of pairs of functions $\{(f_L, f_U)\}$ satisfying (1) $\rho(f_L, f_U) \leq \epsilon$ and (2) for any $f \in \mathcal{C}$, there exist a pair (f_L, f_U) where $f^U \geq f \geq f^L$.

We let $N_{[]}(\epsilon, \boldsymbol{C}, \rho)$ denote the size of the smallest bracketing of $\mathcal{C}$

PROPOSITION 8.2. *(Proposition 16 in Kim and Samworth (2014)) Let $\mathcal{C}$ be the set of convex functions supported on $[-1,1]^d$ and uniformly bounded by B. Then there exist constants ϵ_3 and K^{**}, dependent on d, such that*

$$\log N_{[]}(2\epsilon, \mathcal{C}, L_2) \leq K^{**} \left(\frac{2B}{\epsilon}\right)^{d/2}$$

for all $\epsilon \in (0, B\epsilon_3]$.

It is trivial to extend Kim and Samworth's result to the $L_2(P)$ norm for an absolutely continuous distribution P.

PROPOSITION 8.3. *Let P be a distribution with a density p. Let $\mathcal{C}, B, \epsilon_3, K^{**}$ be defined as in Proposition 8.2. Then,*

$$\log N_{[]}(2\epsilon, \mathcal{C}, L_1(P)) \leq K^{**} \left(\frac{2B}{\epsilon}\right)^{d/2}$$

for all $\epsilon \in (0, B\epsilon_3]$.

PROOF. Let $\mathcal{C}_\epsilon$ be the bracketing the satisfies the size bound in Proposition 8.3. Let $(f_L, f_U) \in \mathcal{C}_\epsilon$. Then we have that:

$$\begin{aligned}
\|f_L - f_U\|_{L_1(P)} &= \int |f_L(x) - f_U(x)| p(x) dx \\
&\leq \left(\int |f_L(x) - f_U(x)|^2 dx\right)^{1/2} \left(\int p(x)^2 dx\right)^{1/2} \\
&\leq \left(\int |f_L(x) - f_U(x)|^2 dx\right)^{1/2} \\
&\leq \|f_L - f_U\|_{L_2} \leq \epsilon.
\end{aligned}$$

On the third line, we used the fact that $\int p(x)^2 dx \leq \left(\int p(x) dx\right)^2 \leq 1$. □

It is also simple to extend the bracketing number result to additive convex functions. As before, let $\mathcal{C}^s$ be the set of additive convex functions with s components.

COROLLARY 8.3. *Let P be a distribution with a density p. Let B, ϵ_3, K^{**} be defined as in Proposition 8.2. Then,*

$$\log N_{[]}(2\epsilon, \mathcal{C}^s, L_1(P)) \leq sK^{**}\left(\frac{2sB}{\epsilon}\right)^{1/2}$$

for all $\epsilon \in (0, sB\epsilon_3]$.

PROOF. Let $f \in \mathcal{C}^s$. We can construct an ϵ-bracketing for f through ϵ/s-bracketings for each of the components $\{f_k\}_{k=1,...,s}$:

$$f_U = \sum_{k=1}^{s} f_{Uk} \qquad f_L = \sum_{k=1}^{s} f_{Lk}$$

It is clear that $f_U \geq f \geq f_L$. It is also clear that $\|f_U - f_L\|_{L_1(P)} \leq \sum_{k=1}^{s} \|f_{Uk} - f_{Lk}\|_{L_1(P)} \leq \epsilon$. □

The following result follows from Corollary 8.3 directly by a union bound.

COROLLARY 8.4. *Let $X_1, ..., X_n$ be random samples from a distribution P. Let $1 > \delta > 0$. Let $\mathcal{C}^s_\epsilon$ be an ϵ-bracketing of $\mathcal{C}^s$ with respect to the $L_1(P)$-norm whose size is at most $N_{[]}(2\epsilon, \mathcal{C}^s, L_1(P))$. Let $\epsilon \in (0, sB\epsilon_3]$.*

Then, with probability at least $1 - \delta$, for all pairs $(f_L, f_U) \in \mathcal{C}^s_\epsilon$, we have that

$$\frac{1}{n}\sum_{i=1}^{n} |f_L(X_i) - f_U(X_i)| \leq \epsilon + \epsilon_{n,\delta}$$

where

$$\epsilon_{n,\delta} \equiv sB\sqrt{\frac{\log N_{[]}(2\epsilon, \mathcal{C}^s, L_2(P)) + \log\frac{1}{\delta}}{2n}} = \sqrt{\frac{sK^{**}(sB)^{1/2}}{2\epsilon^{1/2}n} + \frac{1}{2n}\log\frac{1}{\delta}}.$$

PROOF. Noting that $|f_L(X_i) - f_U(X_i)|$ is at most sB and there are $N_{[]}(2\epsilon, \mathcal{C}^s, L_1(P))$ pairs (f_L, f_U), the inequality follows from a direct application of a union bound and Hoeffding's Inequality. □

To make the expression in this corollary easier to work with, we derive an upper bound for $\epsilon_{n,\delta}$. Suppose

$$\epsilon^{1/2} \leq 2sK^{**}(sB)^{1/2} \quad \text{and} \quad \log\frac{1}{\delta} \geq 2. \tag{8.11}$$

Then we have that

$$\epsilon_n \leq sB\sqrt{\frac{sK^{**}(sB)^{1/2}\log\frac{1}{\delta}}{\epsilon^{1/2}n}}.$$

9. Gaussian Example. Let H be a positive definite matrix and let $f(x_1, x_2) = H_{11}x_1^2 + 2H_{12}x_1x_2 + H_{22}x_2^2 + c$ be a quadratic form where c is a constant such that $\mathbb{E}[f(X)] = 0$. Let $X \sim N(0, \Sigma)$ be a random bivariate Gaussian vector with covariance $\Sigma = [1, \alpha; \alpha, 1]$

PROPOSITION 9.1. *Let $f_1^*(x_1) + f_2^*(x_2)$ be the additive projection of f under the bivariate Gaussian distribution. That is,*

$$f_1^*, f_2^* \equiv \operatorname*{arg\,min}_{f_1, f_2} \left\{ \mathbb{E}\left(f(X) - f_1(X_1) - f_2(X_2)\right)^2 \,:\, \mathbb{E}[f_1(X_1)] = \mathbb{E}[f_2(X_2)] = 0 \right\}$$

Then, we have that

$$f_1^*(x_1) = \left(\frac{T_1 - T_2\alpha^2}{1 - \alpha^4}\right) x_1^2 + c_1$$
$$f_2^*(x_2) = \left(\frac{T_2 - T_1\alpha^2}{1 - \alpha^4}\right) x_2^2 + c_2$$

where $T_1 = H_{11} + 2H_{12}\alpha + H_{22}\alpha^2$ and $T_2 = H_{22} + 2H_{12}\alpha + H_{11}\alpha^2$ and c_1, c_2 are constants such that $\mathbb{E}[f_1^*(X_1)] = \mathbb{E}[f_2^*(X_2)] = 0$.

PROOF. By Lemma 3.1, we need only verify that f_1^*, f_2^* satisfy

$$f_1^*(x_1) = \mathbb{E}[f(X) - f_2^*(X_2) \,|\, x_1]$$
$$f_2^*(x_2) = \mathbb{E}[f(X) - f_1^*(X_1) \,|\, x_2].$$

Let us guess that f_1^*, f_2^* are quadratic forms $f_1^*(x_1) = a_1x_1^2 + c_1$, $f_2^*(x_2) = a_2x_2^2 + c_2$ and verify that there exist a_1, a_2 to satisfy the above equations. Since we are not interested in constants, we define $\simeq$ to be equality up to a constant. Then,

$$\begin{aligned}
&\mathbb{E}[f(X) - f_2^*(X_2) \,|\, x_1] \\
&\simeq \mathbb{E}[H_{11}X_1^2 + 2H_{12}X_1X_2 + H_{22}X_2^2 - a_2X_2^2 \,|\, x_1] \\
&\simeq H_{11}x_1^2 + 2H_{12}x_1\mathbb{E}[X_2 \,|\, x_1] + H_{22}\mathbb{E}[X_2^2 \,|\, x_1] - a_2\mathbb{E}[X_2^2 \,|\, x_1] \\
&\simeq H_{11}x_1^2 + 2H_{12}\alpha x_1^2 + H_{22}\alpha^2 x_1^2 - a_2\alpha^2 x_1^2 \\
&\simeq (H_{11} + 2H_{12}\alpha + H_{22}\alpha^2 - a_2\alpha^2)x_1^2.
\end{aligned}$$

Likewise, we have that

$$\mathbb{E}[f(X) - f_1^2(X_1) \,|\, x_2] \simeq (H_{22} + 2H_{12}\alpha + H_{22}\alpha^2 - a_1\alpha^2)x_2^2.$$

Thus, a_1, a_2 need only satisfy the linear system

$$T_1 - a_2\alpha^2 = a_1$$
$$T_2 - a_1\alpha^2 = a_2$$

where $T_1 = H_{11} + 2H_{12}\alpha + H_{22}\alpha^2$ and $T_2 = H_{22} + 2H_{12}\alpha + H_{11}\alpha^2$. It is then simple to solve the system and verify that a_1, a_2 are as specified. □