Thesis Proposal

Shape-Constrained Estimation in High Dimensions

Min Xu

January 2014

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
John Lafferty, Chair
Larry Wasserman
Aarti Singh
Ming Yuan

Abstract

Shape-constrained estimation techniques such as convex regression or log-concave density estimation offer attractive alternatives to traditional nonparametric methods. Shape-constrained estimation often has an easy-to-optimize likelihood, no tuning parameter, and an adaptivity property where the sample complexity adapts to the complexity of the underlying functions. In this dissertation proposal, our thesis is that some shape-constraints have an additional advantage in that they are naturally suited to high-dimensional problems, where the number of variables is large relative to the number of samples.

In the first part of this proposal, we show that convex functions have an additive faithfulness property, where the additive approximation is guaranteed to capture all relevant variables even if the true function is not additive. We design computationally efficient sparse convex additive models and prove that it achieves variable selection consistency with good sample complexity. The overall work provides a practical bandwidth-free semi-parametric generalization of the Lasso.

We then propose three directions of development for this dissertation. First, we propose to loosen the convexity assumption by learning convex-plus-concave functions, which is a vastly more general function class than convex functions. Second, we consider variable selection on general smooth functions by first decomposing the function into a convex part and a concave part and then exploiting additive faithfulness. Finally, we study graph structure learning on a shape-constrained multivariate distribution.

1 Introduction

Nonparametric estimation methods, such as kernel regression or random forest, are flexible and powerful because of they impose weak assumptions on the underlying function. The downside is that they require more time for computation and more samples for estimation. Nonparametric methods are particularly vulnerable to the curse of dimensionality. Their drawbacks are dramatically exacerbated when the data is high-dimensional, i.e. when the dataset has a large number of variables relative to the number of samples.

In parametric regression, stunning recent advances have shown that under a sparsity assumption, in which most variables are assumed to be uninformative, it is tractable to identify the relevant variables and estimate the function as if the data is low-dimensional. Some analogous results have followed for high-dimensional nonparametric regression but there is still a large gap; there currently exist no method for high-dimensional nonparametric regression that is as practical and theoretically justifiable as parametric methods like the Lasso.

This thesis tackles the problem of high-dimensional nonparametric estimation through shape constraints. Shape-constrained estimation has a rich history and extensive research on topics such as convex or monotone regression and log-concave density estimation. Shape-constraints differ from the usual smoothness assumptions in several ways:

1. It is often possible to directly optimize the likelihood, making the estimation simpler.
2. It is often free of tuning parameters, such as the bandwidth parameter in kernel regression.

3. It exhibits adaptivity; the sample complexity can adapt to the complexity of the underlying function to be learned. [3, 7]

In this thesis, we posit an additional advantage: that shape constraints are naturally suited toward high-dimensional estimation.

Our work thus far has shown that convex functions have an unique property which we call additive faithfulness. This property says that, for the purpose of selecting relevant variables, it is sufficient to assume that the convex function is additive, which significantly eases statistical and computational burden. We design efficient sparse convex additive model and show that it has good finite sample complexity. A striking feature of our model is the lack of a smoothing bandwidth parameter, which makes the model an easy-to-use semi-parametric generalization of the Lasso.

The main disadvantage of shape-constrained estimation is that shape-constraint assumptions can be too strong. For example, rarely can we be sure that the underlying regression function is convex. For our proposed work, we consider ways of generalizing shape constraints. We can for instance learn function that is convex-plus-concave, i.e., a sum of a convex and a concave part; this is computationally a straightforward extension of the sparse convex additive model. We propose also to leverage additive faithfulness to study variable selection on a general smooth function and to study graph structure learning of a shape-constrained multivariate distribution.

Prior work on high-dimensional nonparametric approaches include greedy bandwidth adjustment [10], local-linear lasso [1], and fourier coefficient thresholding [4]. Only the last method can consistently identify the relevant variables if p is allowed to increase polynomially with the sample size n, but the method requires strong assumptions and heavy computation. Much prior work has also been done in shape-constrained estimation, with statistical analysis [3, 5, 15] and computational consideration [8, 12]. This thesis aims to bridge these two directions and derive from them statistical methods that are both practical and theoretically justifable.

2 Sparse Convex Regression (completed)

We consider the problem of estimating the best convex function that explains the response variable based on the observed input variables. [1] We suppose that $y = f(x) + w$ where $x \in \mathbb{R}^p$ are drawn from some distribution and w is an independent zero-mean noise term. In the population version, the regression problem can be formulated as

$$\min_{f \in \mathcal{C}} \mathbb{E}(Y - f(X))^2 \qquad Y \in \mathbb{R}, X \in \mathbb{R}^p \tag{2.1}$$

where $\mathcal{C}$ is the set of convex functions. Although expression 2.1 is an infinite dimensional optimization, the finite sample least-square regression is a finite dimensional quadratic program:

$$\min_{f_1,\ldots,f_n,\beta_1,\ldots,\beta_n} \sum_{i=1}^{n} (Y_i - f_i)^2 \tag{2.2}$$

$$\text{subject to } h_j \geq h_i + \beta_i^{\mathsf{T}}(X_j - X_i) \quad \text{for all } i, j \tag{2.3}$$

[1] All discussions pertinent to concave regression as well by symmetry

where $f_1, ..., f_n$ are scalar variables and $\beta_1, ..., \beta_n$ are p-dimensional subgradient vector variables. Existing work has shown that this finite sample procedure is consistent [11, 15], although the rate is yet unknown.

If the dimensionality of the data is too high, that is, if p is larger than n, then optimization 2.2 will overfit–it would always be possible to achieve zero training error. Linear model overcomes this problem by assuming that the model is sparse and adding an ℓ_1-penalty, i.e. lasso. One approach for convex regression is to follow suit and impose a similar penalty on the subgradient: $\lambda \sum_{k=1,...,p} \|\beta_k\|_\infty$ where each β_k is an n-dimensional vector. Although this approach reduces overfitting, it is computationally intensive and it does not cleanly identify the relevant variables. The β matrix learned has small but non-zero weights on the irrelevant variables.

Our approoach is instead to use an additive model. Though additivity is an approximation, we show that, curiously, we can still achieve variable selection consistency for convex functions.

2.1 Additive Faithfulness

An additive model approximates a multivariable function $f(x)$, with $x \in \mathbb{R}^p$, as a sum of p univariate functions $\sum_{k=1}^p f_k(x_k)$. For general regression, additive approximation may result in a relevant variable being incorrectly marked as irrelevant. Such mistakes are inherent to the approximation and may persist even with infinite samples. In this section we give examples of this phenomenon, and then show how the convexity assumption changes the behavior of the additive approximation. We begin with a lemma that characterizes the components of the additive approximation under mild conditions.

Lemma 2.1. *Let F be a distribution on $C = [0,1]^s$ with a positive density function p. Let $f : C \to \mathbb{R}$ be an integrable function.*

Let $f_1^*, ..., f_s^*, \mu^* \equiv \arg\min\{\mathbb{E}\big(f(X) - \sum_{k=1}^s f_k(X_k) - \mu\big)^2 \,:\, \forall k, \mathbb{E}f_k(X_k) = 0\}$

Then

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

and $\mu^* = \mathbb{E}f(X)$ and this solution is unique.

Lemma 2.1 follows from the stationarity conditions of the optimal solution.

Proof. Let $f_1^*, ..., f_s^*, \mu^*$ be the minimizers as defined.

We first show that the optimal $\mu^* = \mathbb{E}f(X)$ 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 0 and strong convexity is guaranteed.

We now turn our attention toward the f_k^*'s.

It must be that f_k^* minimizes $\Big\{\mathbb{E}\big(f(X) - \mu^* - \sum_{k' \neq k} f_{k'}^*(X_{k'}) - f_k(X_k)\big)^2 \,:\, \mathbb{E}f_k(X_k) = 0\Big\}$.

Fix x_k, we will show that the value $\mathbb{E}[f(X) - \sum_{k' \neq k} f_{k'}(X_{k'}) \,|\, x_k] - \mu^*$, for all x_k, uniquely

minimizes

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

It easily follows then that the function $x_k \mapsto \mathbb{E}[f(X) - \sum_{k'} f_{k'}(X_{k'}) \,|\, x_k] - \mu^*$ is the unique f^*_k that minimizes the expected square error.We focus our attention on f^*_K, and fix x_k.

The first-order optimality condition gives us:

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

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

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

The square error objective is strongly convex. 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.

Now, we note that as a function of x_k, $\mathbb{E}[f(X) - \sum_{k' \neq k} f_{k'}(X_{k'}) | x_k] - \mathbb{E}f(X)$ has mean zero and we thus finish the proof. □

In the case that the distribution in Lemma 2.1 is a product distribution, we get particularly clean expressions for the additive components.

Corollary 2.1. *Let F be a product distribution on $\mathbf{C} = [0,1]^s$ with density function p which is positive on $\mathbf{C}$. Let $\mu^*, f^*_k(x_k)$ be defined as in Lemma 2.1. 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.*

If F is the uniform distribution, then $f^*_k(x_k) = \int f(x_k, \mathbf{x}_{-k}) d\mathbf{x}_{-k}$.

Example 2.1. Using Corollary 2.1, we give two examples of *additive unfaithfulness* under the uniform distribution, that is, examples where relevant variables are erroneously marked as irrelevant under an additive approximation. First, consider the following function:

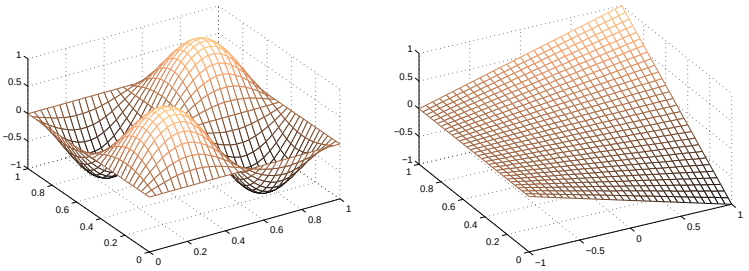
$$\text{(egg carton)} \quad f(x_1, x_2) = \sin(2\pi x_1) \sin(2\pi x_2)$$

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

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

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, 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 intuitive notion that a multivariate function f *depends on* a coordinate x_k.



(a) egg carton (b) tilting slope

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

Definition 2.1. Let F be a distribution on $\mathbf{C} = [0,1]^s$, and $f : \mathbf{C} \to \mathbb{R}$.

We say that f **depends on** coordinate k if, for all $x_k \in [0,1]$, the set $\{x'_k \in [0,1] : f(x_k, \mathbf{x}_{-k}) = f(x'_k, \mathbf{x}_{-k})$ for almost all $\mathbf{x}_{-k}\}$ has probability strictly less than 1.

Suppose we have the additive approximation:

$$f_k^*, \mu^* \equiv \underset{f_1,\ldots,f_s,\mu}{\operatorname{argmin}} \Big\{ \mathbb{E}(f(X) - \sum_{k=1}^{s} f_k(X_k) - \mu)^2 \; : \; \mathbb{E} f_k(X_k) = 0 \Big\}. \tag{2.4}$$

We say that f is **additively faithful** under F in case $f_k^* = 0 \Rightarrow f$ does not depend on coordinate k.

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

2.1.1 Additive Faithfulness of Convex Functions

Remarkably, under a general class of distributions which we characterize below, convex multivariate functions are additively faithful.

Definition 2.2. Let $p(\mathbf{x})$ be a density supported on $[0,1]^s$, p satisfies the *boundary-points condition* if, for all j, and for all $\mathbf{x}_{-j}$:

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

The boundary-points 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 properties that $\frac{\partial p(x_j, \mathbf{x}_{-j})}{\partial x_j} = \frac{\partial^2 p(x_j, \mathbf{x}_{-j})}{\partial x_j^2} = 0$ at points $x_j = 0, x_j = 1$. The boundary-points property is also trivially satisfied when p is the density of any product distributions.

The following theorem is the main result of this section.

Theorem 2.1. *Let p be a positive density supported on $C = [0,1]^s$ that satisfies the boundary-points property (definition 2.2). If f is convex and twice differentiable, then f is* additively faithful *under p.*

We pause to give some intuition before we present the full proof: suppose the underlying distribution is a product distribution for a second, then we know from lemma 2.1 that the additive

approximation zeroes out k when, fixing x_k, every "slice" of f integrates to zero. We prove Theorem 2.1 by showing that "slices" of convex functions that integrate to zero cannot be "glued" together while still maintaining convexity.

Proof. (of Theorem 2.1)
Fix k. Using the result of Lemma 2.1, 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.

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

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

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

$$\int_{\mathbf{x}_{-k}} p'(\mathbf{x}_{-k} \,|\, x_k)\big(f(\mathbf{x}) - r(\mathbf{x}_{-k})\big) + p(\mathbf{x}_{-k} \,|\, x_k) f'(x_k, \mathbf{x}_{-k}) d\mathbf{x}_{-k} = 0 \tag{2.5}$$

$$\int_{\mathbf{x}_{-k}} p''(\mathbf{x}_{-k} \,|\, x_k)\big(f(\mathbf{x}) - r(\mathbf{x}_{-k})\big) + 2p'(\mathbf{x}_{-k} \,|\, x_k) f'(x_k, \mathbf{x}_{-k}) + p(\mathbf{x}_{-k} \,|\, x_k) f''(x_k, \mathbf{x}_{-k}) d\mathbf{x}_{-k} = 0 \tag{2.6}$$

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

$$\int_{\mathbf{x}_{-k}} p(\mathbf{x}_{-k} \,|\, x_k^0) f'(x_k^0, \mathbf{x}_{-k}) d\mathbf{x}_{-k} = 0 \tag{2.7}$$

$$\int_{\mathbf{x}_{-k}} p(\mathbf{x}_{-k} \,|\, x_k^0) f''(x_k^0, \mathbf{x}_{-k}) d\mathbf{x}_{-k} = 0 \tag{2.8}$$

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

The Hessian of f at $(x_k^0, \mathbf{x}_{-k})$ 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 [2], which means that *at all $\mathbf{x}_{-k}$, the gradient of $f'(x_k^0, \mathbf{x}_{-k})$ with respect to $\mathbf{x}_{-k}$ must be zero*.

Therefore, $f'(x_k^0, \mathbf{x}_{-k})$ must be constant for all $\mathbf{x}_{-k}$. By equation 2.7, we conclude then that $f'(x_k^0, \mathbf{x}_{-k}) = 0$ for all $\mathbf{x}_{-k}$. We can use the same reasoning for the case where $x_k = x_k^1$ and

[2] See proposition 7.1.10 of Horn and Johnson [9]

deduce that $f'(x_k^1, \mathbf{x}_{-k}) = 0$ for all $\mathbf{x}_{-k}$.

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

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

Now we apply the first-order condition of convex functions to any two points $(x_k, \mathbf{x}_{-k})$ and $(x'_k, \mathbf{x}_{-k})$ where $x_k, x'_k \in [0,1]$:

$$\begin{aligned}
\forall \mathbf{x}_{-k}, f(x'_k, \mathbf{x}_{-k}) &\leq f(x_k, \mathbf{x}_{-k}) + f'(x_k, \mathbf{x}_{-k})(x'_k - x_k) \\
f(x'_k, \mathbf{x}_{-k}) &\leq f(x_k, \mathbf{x}_{-k}) \\
\forall \mathbf{x}_{-k}, f(x_k, \mathbf{x}_{-k}) &\leq f(x'_k, \mathbf{x}_{-k}) + f'(x'_k, \mathbf{x}_{-k})(x_k - x'_k) \\
f(x_k, \mathbf{x}_{-k}) &\leq f(x'_k, \mathbf{x}_{-k})
\end{aligned}$$

We thus have that $f(x_k, \mathbf{x}_{-k}) = f(x'_k, \mathbf{x}_{-k})$ for all $\mathbf{x}_{-k}$ and all pairs $x_k, x'_k \in (0,1)$. This proves that f does not depend on x_k. □

Theorem 2.1 plays an important role in our sparsistency analysis, where we show that the additive approximation is variable selection consistent (or "sparsistent"), even when the true function is not additive.

Remark 2.1. We assume twice differentiability in Theorems 2.1 to simplify the proof. We believe this smoothness condition is not necessary because every non-smooth convex function can be approximated arbitrarily well by a smooth one.

Remark 2.2. The additive component $f_k^*(x_k)$, which is equal to $\mathbb{E}[f(X) - \sum_{k' \neq k} f_{k'}^*(X_{k'}) \mid x_k] - \mathbb{E}f(X)$ as shown in Lemma 2.1, is not necessarily convex. It is easy to see that $f_k^*(x_k)$ is convex if the underlying distribution p is a product measure, and we propose to identify more general properties of p under which we can guarantee the convexity of the additive components.

2.1.2 Precision of Additive Models

Additive faithfulness guarantees no false negative errors. The opposite direction, which says that if f does not depend on coordinate k, then f_k^* will be zero in the additive approximation, is more complex to analyze. Consider as a conceptual example a 3D distribution over (X_1, X_2, X_3); suppose X_1, X_2 are independent, and f is only a function of X_1, X_2. We can then let $X_3 = f(X_1, X_2) - f_1^*(X_1) - f_2^*(X_2)$, that is, we let X_3 exactly capture the additive approximation error, then 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. Additive precision can be guaranteed if the underlying distribution p is a product measure, and we propose to study more general conditions under which we can ensure precision.

2.2 Optimization Algorithm

Univariate convex functions are characterized by the condition that the subgradient, which is a scalar, must increase monotonically. This observation leads to our optimization program, which

we call SCAM (sparse convex additive model):

$$\begin{aligned} \min_{\boldsymbol{h},\boldsymbol{\beta},\mu} \ & \frac{1}{2n}\sum_{i=1}^{n}\Big(Y_i - \sum_{k=1}^{p} h_{ki} - \mu\Big)^2 + \lambda \sum_{k=1}^{p} \|\boldsymbol{\beta}_{k\cdot}\|_\infty \\ \text{subject to } \ & h_{k(i+1)} = h_{k(i)} + \beta_{k(i)}(x_{k(i+1)} - x_{k(i)}), \\ & \sum_{i=1}^{n} h_{ki} = 0, \\ & \beta_{k(i+1)} \geq \beta_{k(i)} \ (\forall k, i) \end{aligned} \tag{2.9}$$

Here $\{(1), (2), \ldots, (n)\}$ is a reordering of $\{1, 2, \ldots, n\}$ such that $x_{k(1)} \leq x_{k(2)} \leq \cdots \leq x_{k(n)}$. We can solve for μ explicitly, as $\mu = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$ which follows from the KKT conditions and the constraints $\sum_i h_{ki} = 0$.

The ℓ_∞/ℓ_1 penalty $\sum_{k=1}^{p} \|\boldsymbol{\beta}_{k\cdot}\|_\infty$ encourages group sparsity of the vectors $\boldsymbol{\beta}_{k\cdot}$, and thus performs variable selection. We refer to this framework as the sparse convex additive model (SCAM). SCAM uses the *inner piece-wise linear function* that approximates the graph with secant lines. Notice that if we replace $\beta_{k(i+1)} \geq \beta_{k(i)}$ with $\beta_{k(i+1)} = \beta_{k(i)}$, the optimization reduces to the lasso.

The SCAM optimization in (2.9) is a quadratic program (QP) with $O(np)$ variables and $O(np)$ constraints. Directly applying a QP solver for $\boldsymbol{h}, \boldsymbol{\beta}$ would be computationally expensive for relatively large n and p. However, notice that variables in different feature dimensions are only coupled in the term $(Y_i - \sum_{k=1}^{p} h_{ki})^2$. Hence, we can apply the block coordinate descent method, where in each step we solve the following QP subproblem for $\{\boldsymbol{h}_{k\cdot}, \boldsymbol{\beta}_{k\cdot}\}$ with the other variables fixed:

$$\begin{aligned} \min_{\boldsymbol{h}_{k\cdot},\boldsymbol{\beta}_{k\cdot},\gamma_k} \ & \frac{1}{2n}\sum_{i=1}^{n}\Big((Y_i - \bar{Y} - \sum_{r\neq k} h_{ri}) - h_{ki}\Big)^2 + \lambda\gamma_k \\ \text{such that } & h_{k(i+1)} = h_{k(i)} + \beta_{k(i)}(x_{k(i+1)} - x_{k(i)}), \\ & \beta_{k(i+1)} \geq \beta_{k(i)}, \ -\gamma_k \leq \beta_{k(i)} \leq \gamma_k \\ & \sum_{i=1}^{n} h_{ki} = 0, \ (\forall i). \end{aligned}$$

The extra variable γ_k is introduced to deal with 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 (e.g., MOSEK: `http://www.mosek.com/`). We cycle through all feature dimensions (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 (2.9), but found that it is not as efficient as this QP solver.

After optimization, the function estimator for any input data $\boldsymbol{x}_j$ is

$$f(\boldsymbol{x}_j) = \sum_{k=1}^{p} f_k(x_{kj}) + \mu = \sum_{k=1}^{p} \max_i\{h_{ki} + \beta_{ki}(x_{kj} - x_{ki})\} + \mu.$$

2.3 Alternative Formulation

Optimization (2.9) can be reformulated in terms of the 2nd derivatives, a form which we analyze in our theoretical analysis. The alternative formulation replaces the ordering constraints $\beta_{k(i+1)} \geq \beta_{k(i)}$ with positivity constraints, which simplifies theoretical analysis. Define $d_{k(i)}$ as the second derivative: $d_{k(1)} = \beta_{k(1)}$, and $d_{k(2)} = \beta_{k(2)} - \beta_{k(1)}$. The convexity constraint is equivalent to the constraint that $d_{k(i)} \geq 0$ for all $i > 1$.

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

$$f_k(x_{k(i)}) = f_k(x_{k(1)}) + d_{k(1)}(x_{k(i)} - x_{k(1)}) + d_{k(2)}(x_{k(i)} - x_{k(2)}) + \cdots d_{k(i-1)}(x_{k(i)} - x_{k(i-1)}) \quad (2.10)$$

We can write this more compactly in matrix notations. First define $\Delta_{k(j)}(x_{ki}) = \max(x_{ki} - x_{k(j)}, 0)$.

$$(f_k(x_{k1}), ..., f_k(x_{kn}))^\mathsf{T} = \Delta_k d_k \equiv \begin{bmatrix} \Delta_{k(1)}(x_{k1}) & ... & \Delta_{k(n-1)}(x_{k1}) \\ ... & & \\ \Delta_{k(1)}(x_{kn}) & ... & \Delta_{k(n-1)}(x_{kn}) \end{bmatrix} \begin{bmatrix} d_{k(1)} \\ ... \\ d_{k(n-1)} \end{bmatrix}$$

Where Δ_k is a $n \times n-1$ matrix such that $\Delta_k(i,j) = \Delta_{k(j)}(x_{ki})$ and $d_k = (d_{k(1)}, ..., d_{k(n-1)})$. We can now reformulate (2.9) as an equivalent optimization program with only centering and positivity constraints:

$$\begin{aligned} &\min_{d_k, c_k} \frac{1}{2n} \left\| Y - \bar{Y}\mathbf{1}_n - \sum_{k=1}^{p} (\Delta_k d_k - c_k \mathbf{1}_n) \right\|_2^2 + \lambda_n \sum_{k=1}^{p} \|d_k\|_1 \qquad (2.11) \\ &\text{s.t. } d_{k(2)}, \ldots, d_{k(n-1)} \geq 0 \qquad \text{(convexity)} \\ &\quad c_k = \frac{1}{n} \mathbf{1}_n^\mathsf{T} \Delta_k d_k \qquad \text{(centering)} \end{aligned}$$

$\|d_k\|_1$ is not identical to $\|\boldsymbol{\beta}_{k\cdot}\|_\infty$, but it is easy to verify that $\|\boldsymbol{\beta}_{k\cdot}\|_\infty \leq \|d_k\|_1 \leq 4\|\boldsymbol{\beta}_{k\cdot}\|_\infty$.

Remark 2.3. For parts of our theoretical analysis, we will also impose onto (2.11) a boundedness constraint $-B\mathbf{1}_n \leq \Delta_k d_k + c_k \mathbf{1}_n \leq B\mathbf{1}_n$ which constrains that $\|f_k\|_\infty \leq B$, or a Lipschitz constraint $\|d_k\|_1 \leq L$ which constrains that f_k must be L-Lipschitz. We use these constraints only in the proof for technical reasons; we never need nor use these constraints in our experiments.

2.4 Variable Selection Consistency

We show in this section that sparse convex additive model is variable selection consistent, i.e., as $n \to \infty$, we have that $P(\text{supp}(\widehat{f}) = \text{supp}(f^*)) \to 0$. Because we are especially interested in the high dimensional setting, we derive our rate of consistency in terms of both sample size n and dimensionality p (and certain other quantities pertinent to the problem) and show that our procedure is variable selection consistent even if $p \to \infty$ as well at some rate p_n.

We divide our analysis into two parts. We first establish a sufficient *deterministic* condition for sparsistency. We then consider the stochastic setting and argue that the deterministic conditions hold with high probability.

2.5 Deterministic Setting

We follow [18] and define the *restricted regression* purely for theoretical purposes.

Definition 2.3. In *restricted regression*, we restrict the indices k in optimization (2.11) to lie in the support S instead of ranging from $1, ..., p$.

Our analysis then differs from the now-standard "primal-dual witness technique" [18]. Primal-dual witness explicitly solves all the dual variables, but because our optimization is more complex, we do not solve the dual variables on S; we instead write the dual variables on S^c as a function of the restricted regression *residual*, which is implicitly a function of the dual variables on S.

Theorem 2.2. *(Deterministic setting) Let $\{\widehat{d}_k, \widehat{c}_k\}_{k \in S}$ be the minimizer of the restricted regression, that is, the solution to optimization (2.11) where we restrict $k \in S$. Let $\widehat{d}_k = 0$ and $\widehat{c}_k = 0$ for $k \in S^c$. Let $\widehat{r} \equiv Y - \bar{Y}\mathbf{1}_n - \sum_{k \in S}(\Delta_k \widehat{d}_k - \widehat{c}_k \mathbf{1}_n)$ be the restricted regression residual. For $k \in \{1, ..., p\}$, Let $\Delta_{k,j} \in R^n$ be the j-th column of Δ_k, i.e. $\max(X_k - X_{k(j)}\mathbf{1}_n, 0)$.*

Suppose for all j and all $k \in S^c$, $\lambda_n > |\frac{1}{n}\widehat{r}^\mathsf{T}\Delta_{k,j}|$. Then $\widehat{\mu}$ and $\widehat{d}_k, \widehat{c}_k$ for $k = 1, ..., p$ is an optimal solution to the full regression 2.11. Furthermore, any solution to the optimization program 2.11 must be zero on S^c.

This result holds regardless of whether we impose the boundedness and Lipschitz conditions in optimization 2.11. The full proof of Theorem 2.2 is in Section 5.1 of the Appendix.

Remark 2.4. The incoherence condition of [18] is implicitly encoded in our condition on $\lambda_n, \widehat{r}, \Delta_{k,j}$. We can reconstruct the incoherence condition if we assume that the true function f_0 is linear and that our fitted functions $\widehat{f}_k$ are linear as well.

Theorem 2.2 allows us to analyze false negative rates and false positive rates separately. To control false positives, we study when the condition $\lambda_n > |\frac{1}{n}\widehat{r}^\mathsf{T}\Delta_{k,j}|$ is fulfilled for all j and all $k \in S^c$. To control false negatives, we study the restricted regression.

2.6 Probabilistic Setting

We use the following statistical setting:

1. Let F be a distribution supported and positive on $\mathcal{X} = [-b, b]^p$. Let $X^{(1)}, ..., X^{(n)} \sim F$
2. Let $Y = f_0(X) + \epsilon$ where ϵ is zero-mean noise. Let $Y^{(1)}, ..., Y^{(n)}$ be iid.
3. Let $S = \{1, ..., s\}$ denote the relevant variables where $s \leq p$, i.e., $f_0(X) = f_0(X_S)$.
4. Let $f_1^*, ..., f_s^* \equiv \arg\min_{f_1, ..., f_s}\{\mathbb{E}\Big(f_0(X) - \mathbb{E}f_0(X) - \sum_{k=1}^s f_k(X_k)\Big)^2 \mid \mathbb{E}[f_k(X_k)] = 0\}$.

Each of our theorems will use a subset of the following assumptions:

A1: X_S, X_{S^c} are independent. A1': $\{X_k\}_{k \in S}$ are independent.

A2: $\|f_0\|_\infty \leq sB$ A2': f_0 is convex, twice-differentiable, and L-Lipschitz.

A3: Suppose ϵ is mean-zero sub-Gaussian, independent of X, with sub-Gaussian scale σ, i.e. for all $t \in \mathbb{R}$, $\mathbb{E}e^{t\epsilon} \leq e^{\sigma^2 t^2/2}$.

A4: For all $k = 1, ..., s$, $\mathbb{E}(f_s^*(X_k))^2 \geq \alpha$ for some positive constant α.

We will use assumptions A1, A2, A3 to control the probability of false positives and the stronger assumptions A1', A2', A3, A4 to control the probability of false negatives. Assumption A4 can be weakened so that the relevant functions satisfy $\mathbb{E}(f_s^*(X_k))^2 \geq \alpha_n$ for α_n decaying to zero at an appropriate rate.

Remark 2.5. Assumption A4 ensures that the relevant variables are "relevant enough". Under A4, the population risk of an additive function with $s-1$ components is at least α larger than the population risk of the optimal additive function with s components. See lemma 5.1 in section 5.3 of the appendix.

Theorem 2.3. *(Controlling false positives) Suppose assumptions A1, A2, A3 hold. Suppose also that we run optimization* (2.11) *with the B-boundness constraint. Let c, C be absolute constants. Suppose $\lambda_n \geq cb(sB+\sigma)\sqrt{\frac{s}{n}\log n \log(pn)}$. Then with probability at least $1-\frac{C}{n}$, for all j,k, $\lambda_n > |\frac{1}{n}\widehat{r}^{\mathsf{T}}\Delta_{k,j}|$. Therefore, any solution to the full regression (2.11), with boundedness constraint, is zero on S^c.*

The proof of Theorem 2.3 exploits independence of $\widehat{r}$ and $\Delta_{k,j}$ from A1, and then uses concentration of measure results to argue that $|\frac{1}{n}\widehat{r}^{\mathsf{T}}\Delta_{k,j}|$ concentrates around zero at a desired rate. The fact that $\widehat{r}$ is a centered vector is crucial to our proof, and our theory thus further illustrates the importance of imposing the centering constraints in optimization (2.11). Our proof uses the concentration of the average of data sampled *without* replacement [16]. The full proof of Theorem 2.3 is in Section 5.2 of the Appendix.

Theorem 2.4. *(Controlling false negatives) Suppose assumptions A1', A2', A3, A4 hold. Let $\widehat{f} = \{\widehat{d}_k, \widehat{c}_k\}_{k\in S}$ be any solution to the restricted regression with both the B-boundedness and L-Lipschitz constraint. Let c, C be absolute constants. Suppose $sL\lambda_n \to 0$ and $Lb(sB+\sigma)sB\sqrt{\frac{s}{n^{4/5}}\log sn} \to 0$. Then, for sufficiently large n, $\widehat{f}_k = (\widehat{d}_k, \widehat{c}_k) \neq 0$ for all $k \in S$ with probability at least $1-\frac{C}{n}$.*

This is a finite sample version of Theorem 2.1. We need stronger assumptions in Theorem 2.4 to use our additive faithfulness result, Theorem 2.1. We also include an extra Lipschitz constraint so that we can use existing covering number results [2]. Recent work [6] shows that the Lipschitz constraint is not required with more advanced empirical process theory techniques; we leave the incorporation of this development as future work. We give the full proof of Theorem 2.4 in Section 5.3 of the Appendix.

Combining Theorem 2.3 and 2.4 and ignoring dependencies on b, B, L, σ, we have the following result.

Corollary 2.2. *Assume A1', A2', A3, A4. Let $\lambda_n = \Theta\left(\sqrt{\frac{s^3}{n}\log n \log(pn)}\right)$. Suppose $s\lambda_n \to 0$ and $\sqrt{\frac{s^5}{n^{4/5}}\log sn} \to 0$. Let $\widehat{f}_n$ be a solution to (2.11) with boundedness and Lipschitz constraints. Then $\mathbb{P}(\mathit{supp}(\widehat{f}_n) = \mathit{supp}(f_0)) \to 1$.*

The above corollary implies that sparsistency is achievable at the same exponential scaling of the ambient dimension $p = O(\exp(n^c)), c < 1$ rate as parametric models. The cost of nonparametric modeling is reflected in the scaling with respect to s, which can only scale at $o(n^{4/25})$.

Remark 2.6. Comminges and Dalalyan [4] have shown that under tradtional smoothness constraints, variable selection is achievable only if $n > O(e^s)$. It is interesting to observe that because of additive faithfulness, the convexity assumption enables a much better scaling of

$n = O(\text{poly}(s))$, demonstrating that geometric constraints can be quite different from the previously studied smoothness conditions.

2.7 Experiments

We first illustrate our methods using a simulation of the following regression problem

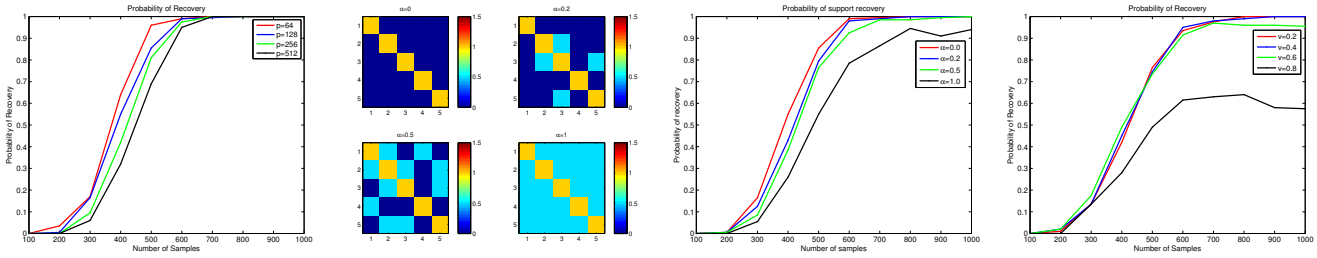
$$y_i = \boldsymbol{x}_{iS}^\top \boldsymbol{Q} \boldsymbol{x}_{iS} + \epsilon_i \quad (i = 1, 2, \ldots, n).$$

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

In the first simulation, we set $\boldsymbol{Q} = \boldsymbol{I}_{|S|}$ (the additive case), and choose $n = 100, 200, \ldots, 1000$ and $p = 64, 128, 256, 512$. For each (n, p) combination, we generate 200 independent data sets. For each data set we use SCAM to infer the model parameterized by $\boldsymbol{h}$ and $\boldsymbol{\beta}$; see equation (2.9). If $\|\boldsymbol{\beta}_{k\cdot}\|_\infty < 10^{-8}$ $(\forall k \notin S)$ and $\|\boldsymbol{\beta}_{k\cdot}\|_\infty > 10^{-8}$ $(\forall k \in S)$, then we declare correct support recovery. We then plot the probability of support recovery over the 200 data sets in Figure 2(a). We observe that SCAM performs consistent variable selection when the true function is convex additive. To give the reader a sense of the running speed, the code runs in about 2 minutes on one data set with $n = 1000$ and $p = 512$, on a MacBook with 2.3 GHz Intel Core i5 CPU and 4 GB memory.

In the second simulation, we study the case in which the true function is convex but not additive. We generate four $\boldsymbol{Q}$ matrices plotted in Figure 2(b), where the diagonal elements are all 1 and the off-diagonal elements are 0.5 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 again run the SCAM optimization on 200 independently generated data sets and plot the probability of recovery in Figure 2(c). The results demonstrate that SCAM performs consistent variable selection even if the true function is not additive (but still convex).

In the third simulation, we study the case of correlated design, where $\boldsymbol{x}_i$ is drawn from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ instead of $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$, with $\Sigma_{ij} = \nu^{|i-j|}$. We use the non-additive $\boldsymbol{Q}$ with $\alpha = 0.5$ and fix $p = 128$. The recovery curves for $\nu = 02, 0.4, 0.6, 0.8$ are depicted in Figure 2(d). As can be seen, for design of moderate correlation, SCAM can still select relevant variables well.



(a) additive model (b) four $\boldsymbol{Q}$ matrices (c) non-additive models (d) correlated design

Figure 2: Support recovery results where the additive assumption is correct (a), incorrect (b), (c), and with correlated design (d).

3 Proposed Work

3.1 More on Additive Faithfulness

Many important questions remain unanswered about the additive faithfulness property of convex functions.

1. Could we extend additive faithfulness to distributions supported on $\mathbb{R}^d$? Is additive faithfulness specific to the L2-loss or are convex functions additively faithful under other loss functions?
2. Under what general conditions can we guarantee that the optimal additive approximations of convex functions are themselves convex? When could we guarantee that the additive approximation has high precision, i.e., contains no false positives? These two properties are easy to prove under a product measure, but independence of the variables is too strong of an assumption.
3. The additive faithfulness result also implies that nonparametric marginal regression is faithful. It is then necessary to study cases under which marginal regression perform just as well as the additive model and cases under which the additive model performs better.
4. It would be interesting to extend the additive faithfulness result beyond univariate additive models. Could we for instance select bivariate components and retain the faithfulness property?

3.2 Convex-plus-Concave Regression

Convexity is often too strong of an assumption on the underlying regression function. To remedy this, we propose to model functions that are convex-plus-concave, i.e., functions that can be decomposed into the sum of a convex and a concave function. This function class is very general, as shown by the following theorem.

Theorem 3.1. *Any function $h : \mathbb{R}^d \to \mathbb{R}$ with a bounded Hessian can be decomposed as $h(x) = f(x) + g(x)$ where $f(x)$ is convex and $g(x)$ is $-cx^\mathsf{T}x$ for some $c \geq 0$.*

Proof. The Hessian of $f(x) = h(x) - g(x)$ is $H_x + cI$ where H is the Hessian of $h(x)$. For large enough c then, it must be that $H_x + cI$ is positive semidefinite. The function f is thus convex. □

The optimization program for learning additive convex functions can easily be modified to learn additive convex-plus-concave functions:

$$\begin{aligned}
\min_{\boldsymbol{h},\boldsymbol{f},\boldsymbol{g},\boldsymbol{\beta},\boldsymbol{\gamma},\mu,} \quad & \frac{1}{2n}\sum_{i=1}^{n}\Big(Y_i - \sum_{k=1}^{p} h_{ki} - \mu\Big)^2 + \lambda \sum_{k=1}^{p} \|\boldsymbol{\beta}_{k\cdot} + \boldsymbol{\gamma}_{k\cdot}\|_\infty \\
\text{subject to} \quad & f_{k(i+1)} = f_{k(i)} + \beta_{k(i)}(x_{k(i+1)} - x_{k(i)}), \\
& g_{k(i+1)} = g_{k(i)} + \gamma_{k(i)}(x_{k(i+1)} - x_{k(i)}) \\
& h_{ki} = f_{ki} + g_{ki} \qquad \sum_{i=1}^{n} h_{ki} = 0, \\
& \beta_{k(i+1)} \geq \beta_{k(i)} \qquad \gamma_{k(i+1)} \leq \gamma_{k(i)} \\
& \gamma_{ki} - \gamma_{k(i+1)} < c \quad \forall k, i
\end{aligned} \tag{3.1}$$

We need constraint 3.1 because unlike convex regression, convex-plus-concave regression can easily overfit–as shown by theorem 3.1, it is possible to represent almost arbitrary functions with the sum of a convex and a concave function. The constraint 3.1 is a lower bound on the second derivative of the estimated functions and thus reduces overfitting. The parameter c thus acts like the smoothing bandwidth in kernel regression.

Convex-plus-concave additive models have several potential advantages over current methods for additive models such as backfitting [14] or RKHS kernels [13]. It is easy and efficient to optimize and possibly shares the adaptivity of convex regression. The constraint 3.1 re-introduces the nuisances of parameter tuning and empirical studies will be needed to determine whether this parameter is easier or harder to tune than the traditional smoothing parameter.

Preliminary experiments with convex-plus-concave regression (albeit a different optimization from the one presented) shows that it is quite versatile; see figure 3.2. Smoothness tuning does become important however.

3.3 Variable Selection via Convex-Concave Separation

Because the convex-plus-concave functions are very general, they are in general not additively faithful. The tilting slope (example 2.1) is an example of an additively unfaithful convex-plus-concave function.

Nevertheless, we may still be able to take advantage of additive faithfulness if we first separate a general function h into a convex part f and a concave part g. We propose a two step variable selection procedure on a convex-plus-concave function h.

Step 1. We learn low-dimensional *non-additive* convex function f and concave function g

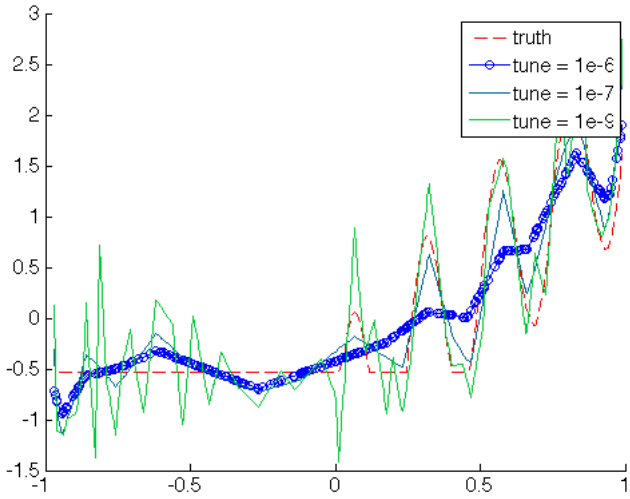


Figure 3: One-dimensional convex-plus-concave fit for many different smoothness parameter.

such that $h = f + g$:

$$\min_{\boldsymbol{f},\boldsymbol{g},\boldsymbol{\beta},\boldsymbol{\gamma}} \frac{1}{n}\|y - (f+g)\|_2^2 + \lambda \sum_{j=1}^{p} \|\beta_j + \gamma_j\|_\infty$$
$$\text{s.t.} f_{i'} \geq f_i + \beta_i^{\mathsf{T}}(x_{i'} - x_i)$$
$$g_{i'} \leq g_i + \gamma_i^{\mathsf{T}}(x_{i'} - x_i)$$
$$\sum_i f_i = 0 \quad \sum_i g_i = 0$$

The penalty on the subgradient of f, g encourages f, g to be low-dimensional. This optimization is not effective for variable selection but it may be enough for low-dimensional denoising and decomposition.

Step 2. We apply additive model variable selection on the f, g produced from step 1. If f, g are both dependent only on the relevant variable set S, i.e. $f(x) = f(x_S)$ and $g(x) = g(x_S)$, then the second step additive model will succeed.

Preliminary experiments have shown that this approach can successfully select the relevant variables for the tilting slope function $f(x) = x_1 x_2$. Although the proposed method is difficult to analyze with full generality, it is possible to give a partial analysis by imposing Lipschitz assumptions on the component functions f and g. The method is currently tractable for $p = 100$ and $n = 1000$ but speed improvement is still a major concern.

3.4 Graphical Learning on Shape-constrained Distributions

An undirected conditional independence graph can be defined on a multivariate distribution $(X_1, ..., X_p)$ where nodes j, k form an edge if $p(x_j \,|\, x_k, x_{-(j,k)}) \neq p(x_j \,|\, x_{-(j,k)})$ or vice versa.

The conditional independence graph is a natural way to adapt sparsity to density estimation: we can assume that the underlying distribution has a sparse graph.

For a Gaussian distribution, the structure of the graph is encoded into the sparsity pattern of its inverse covariance (precision) matrix. It is therefore possible to either directly estimate a sparse precision matrix or to perform neighborhood search, that is, to perform sparse regression with all variables x_{-j} on x_j and set the variables with non-zero coefficients to be the neighbors of variable j. It is not known how to perform efficient graph estimation beyond the Gaussian assumption. [3]

One way of generalizing the Gaussian assumption is to impose shape-constraint assumptions on the density. One could for example work with the family of log-concave distributions which includes the Gaussian, the Laplacian, the Dirichlet, and many more distributions. We propose then to apply the machinery for high dimensional regression with shape-constrained functions toward the problem of high-dimensional graph estimation with shape-constrained density.

There has been exciting development in the estimation of a low-dimensional log-concave density [5], but the problem of estimating the graph of a high-dimensional log-concave density seems impenetrable despite extensive effort on our part. One potentially interesting research direction is to consider distribution whose *conditional* density is log-concave. This is a subclass of log-concave distribution but it does encompass the Gaussian distribution, the Dirichlet distribution (under certain parameter settings), and possibly many other commonly used multivariate distributions. The graph of a conditionally log-concave distribution could be estimated via a neighborhood search procedure using some of convex regression at each step.

4 Timeline

1. Convex-plus-concave regression. June 2014.
2. Variable selection via separation. October 2014.
3. Graph structure learning. Uncertain.
4. **Thesis Defense** Spring 2015.

Acknowledgement

The author gratefully acknowledges collaboration with Minhua Chen who contributed invaluable work and ideas to the research described in this thesis.

References

[1] Karine Bertin and Guillaume Lecué. Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electronic Journal of Statistics*, 2:1224–1241, 2008. 1

[3] And certain other structure assumptions such as the tree structure assumption

[2] E. M. Bronshtein. ε-entropy of convex sets and functions. *Siberian Mathematical Journal*, 17:393–398, 1976. 2.6

[3] T Tony Cai and Mark G Low. A framework for estimation of convex functions. Technical report, Technical report, 2011. 3, 1

[4] Laëtitia Comminges and Arnak S. Dalalyan. Tight conditions for consistency of variable selection in the context of high dimensionality. *The Annals of Statistics*, 40(5):2667–2696, 2012. 1, 2.6

[5] M. Cule, R. Samworth, and M. Stewart. Maximum likelihood estimation of a multi-dimensional log-concave density (with discussion). *J. R. Stat. Soc. Ser. B, Stat. Methodol.*, 72:545–600, 2010. 1, 3.4

[6] A. Guntuboyina and B. Sen. Covering numbers for convex functions. *IEEE Trans. Info. Theory*, 59:1957–1965, 2013. 2.6

[7] Adityanand Guntuboyina and Bodhisattva Sen. Global risk bounds and adaptation in univariate convex regression. *arXiv preprint arXiv:1305.1648*, 2013. 3

[8] L. A. Hannah and D. B. Dunson. Ensemble methods for convex regression with applications to geometric programming based circuit design. In *International Conference on Machine Learning (ICML)*, 2012. 1

[9] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press; Reprint edition, 1990. 2

[10] John Lafferty and Larry Wasserman. Rodeo: sparse, greedy nonparametric regression. *The Annals of Statistics*, 36(1):28–63, 2008. 1

[11] Eunji Lim and Peter W. Glynn. Consistency of multidimensional convex regression. *Operations Research*, 60(1):196–208, 2012. 2

[12] Natalya Pya. *Additive models with shape constraints*. PhD thesis, University of Bath, 2010. 1

[13] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.*, 13: 389–427, 2012. 3.2

[14] Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society, Series B, Methodological*, 71(5):1009–1030, 2009. 3.2

[15] Emilio Seijo and Bodhisattva Sen. Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics*, 39(3):1633–1657, 2011. 1, 2

[16] Robert J Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 2(1):39–48, 1974. 2.6, 5.2

[17] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010. 5.4.1, 5.1

[18] Martin Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information*

Theory, 55(5):2183–2202, May 2009. 2.5, 2.5, 2.4, 5.1

5 Appendix

5.1 Proof of the Deterministic Condition for Sparsistency

We restate Theorem 2.2 first for convenience.

Theorem 5.1. *The following holds regardless of whether we impose the boundedness and smoothness condition in optimization 2.11 or not.*

For $k \in \{1, ..., p\}$, let $\Delta_{k,j}$ denote the n-dimensional vector $\max(X_k - X_{k(j)}\mathbf{1}, 0)$.

Let $\{\widehat{d}_k, \widehat{c}_k\}_{k \in S}$ be the minimizer of the restricted regression optimization program 2.11. Let $\widehat{d}_k = 0$ and $\widehat{c}_k = 0$ for $k \in S^c$.

Let $\widehat{r} \coloneqq Y - \bar{Y}\mathbf{1}_n - \sum_{k \in S}(\Delta_k \widehat{d}_k - \widehat{c}_k \mathbf{1}_n)$ be the residual.

Suppose for all $j = 1, ..., n, k \in S^c$, $\lambda_n > |\frac{1}{n}\widehat{r}^\mathsf{T}\Delta_{k,j}|$, then $\widehat{d}_k, \widehat{c}_k$ for $k = 1, ..., p$ is an optimal solution to the full regression 2.11.

Furthermore, any solution to the optimization program 2.11 must be zero on S^c.

Proof. We will omit the boundedness and smoothness constraints in our proof here. It is easy to add those in and check that the result of the theorem still holds.

We will show that with $\widehat{d}_k, \widehat{c}_k$ as constructed, we can set the dual variables to satisfy complementary slackness and stationary conditions: $\nabla_{d_k, c_k} L(\widehat{d}) = 0$ for all k.

we can re-write the Lagrangian L, in term of just d_k, c_k, as the following.

$$\min_{d_k, c_k} \frac{1}{2n}\|r_k - \Delta_k d_k + c_k \mathbf{1}\|_2^2 + \lambda \sum_{i=2}^{n} d_{ki} + \lambda |d_{k1}| - \mu_k^\mathsf{T} d_k + \gamma_k (c_k - \mathbf{1}_n^\mathsf{T} \Delta_k d_k)$$

where $r_k \coloneqq Y - \bar{Y}\mathbf{1}_n - \sum_{k' \in S, k' \neq k}(\Delta_{k'} d_{k'} - c_{k'}\mathbf{1}_n)$, and $\mu_k \in \mathbb{R}^{n-1}$ is a vector of dual variables where $\mu_{k,1} = 0$ and $\mu_{k,i} \geq 0$ for $i = 2, ..., n-1$.

First, note that by definition as solution of the restricted regression, for $k \in S$, $\widehat{d}_k, \widehat{c}_k$ satisfy stationarity with dual variables that satisfy complementary slackness.

Now, let us fix $k \in S^c$ and prove that $\widehat{d}_k = 0, \widehat{c}_k = 0$ is an optimal solution.

$$\begin{aligned}
\partial d_k : \quad & -\frac{1}{n}\Delta_k^\mathsf{T}(r_k - \Delta_k d_k + c_k \mathbf{1}) + \lambda \mathbf{u}_k - \mu_k - \gamma_k \Delta_k^\mathsf{T} \mathbf{1} \\
\partial c_k : \quad & -\frac{1}{n}\mathbf{1}^\mathsf{T}(r_k - \Delta_k d_k + c_k \mathbf{1}) + \gamma_k
\end{aligned}$$

In the derivatives, $\mathbf{u}_k$ is a $(n-1)$-vector whose first coordinate is $\partial |d_{k1}|$ and all other coordinates are 1.

We now substitute in $d_k = \widehat{d}_k = 0, c_k = \widehat{c}_k = 0, r_k = \widehat{r}_k = \widehat{r}$ and show that the duals can be set in a way to ensure that the derivatives are equal to 0.

$$-\frac{1}{n}\Delta_k^{\mathsf{T}}\widehat{r} + \lambda \mathbf{u}_k - \mu_k - \gamma_k \Delta_k^{\mathsf{T}}\mathbf{1} = 0$$
$$-\frac{1}{n}\mathbf{1}^{\mathsf{T}}\widehat{r} + \gamma_k = 0$$

where $\mathbf{u}_k$ is 1 in every coordinate except the first, where it can take any value in $[-1, 1]$.

First, we observe that $\gamma_k = 0$ because $\widehat{r}$ has empirical mean 0. All we need to prove then is that

$$\lambda \mathbf{u}_k - \mu_k = \frac{1}{n}\Delta_k^{\mathsf{T}}\widehat{r}.$$

Suppose

$$\lambda \mathbf{1} > |\frac{1}{n}\Delta_k^{\mathsf{T}}\widehat{r}|,$$

then we easily see that the first coordinate of $\mathbf{u}_k$ can be set to some value in $(-1, 1)$ and we can set $\mu_{k,i} > 0$ for $i = 2, ..., n-1$.

Because we have strict inequality in the above equation, Lemma 1 from [18] show that all solutions must be zero on S^c. □

5.2 Proof of False Positive Control

Note: 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 5.2. *Suppose assumptions A1, A2, A3 hold.*

Suppose $\lambda_n \geq cb(sB + \sigma)\sqrt{\frac{s}{n}\log n \log(pn)}$*, then with probability at least* $1 - \frac{C}{n}$*, for all* $j = 1, ..., n, k \in S^c$*,*

$$\lambda_n > |\frac{1}{n}\widehat{r}^{\mathsf{T}}\Delta_{k,j}|$$

And therefore, the solution to the optimization 2.11 is zero on S^c*.*

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.

We remind the reader that $\Delta_{k,j} = \max(X_k, -X_{k(j)}\mathbf{1}_n, 0)$. Because $\widehat{r}$ is empirically centered,

$$\begin{aligned}\frac{1}{n}\widehat{r}^{\mathsf{T}}\Delta_{k,j} &= \frac{1}{n}\widehat{r}^{\mathsf{T}}\max(X_k, X_{k(j)}\mathbf{1}_n) - \frac{1}{n}\widehat{r}^{\mathsf{T}}\mathbf{1}_n X_{k(j)} \\ &= \frac{1}{n}\widehat{r}^{\mathsf{T}}\max(X_k, X_{k(j)}\mathbf{1}_n)\end{aligned}$$

Our goal in this proof is to bound $\frac{1}{n}\widehat{r}^{\mathsf{T}}\max(X_k, X_{k(j)})$ from above.

Step 1. We first get a high probability bound on $\|\widehat{r}\|_\infty$.

$$\begin{aligned}\widehat{r}_i &= Y_i - \bar{Y} - \sum_{k\in S}\widehat{f}_k(X_k^{(i)}) \\ &= f_0(X_S^{(i)}) + \epsilon_i - \bar{f}_0 - \bar{\epsilon} - \sum_{k\in S}\widehat{f}_k(X_k^{(i)}) \\ &= f_0(X_S^{(i)}) - \bar{f}_0 - \sum_{k\in S}\widehat{f}_k(X_k^{(i)}) + \epsilon_i - \bar{\epsilon}\end{aligned}$$

Where $\bar{f}_0 = \frac{1}{n}\sum_{i=1}^n f_0(X_S^{(i)})$ and likewise for $\bar{\epsilon}$.

ϵ_i is subgaussian with subgaussian norm σ. For a single ϵ_i, we have that $P(|\epsilon_i| \geq t) \leq C\exp(-c\frac{1}{\sigma^2}t^2)$. Therefore, with probability at least $1-\delta$, $|\epsilon_i| \leq \sigma\sqrt{\frac{1}{c}\log\frac{C}{\delta}}$.

By union bound, with probability at least $1-\delta$, $\max_i|\epsilon_i| \leq \sigma\sqrt{\frac{1}{c}\log\frac{2nC}{\delta}}$.

Also, $|\bar{\epsilon}| \leq \sigma\sqrt{\frac{c}{n}\log\frac{C}{\delta}}$ with probability at least $1-\delta$.

We know that $|f_0(x)| \leq sB$ and $|\widehat{f}_k(x_k)| \leq B$ for all k.

Then $|f_0(\bar{x})| \leq sB$ as well, and $|f^*(X_S^{(i)}) - \bar{f}^* - \sum_{k\in S}\widehat{f}_k(X_k^{(i)})| \leq 3sB$.

Therefore, taking an union bound, we have that with probability at least $1-\frac{C}{n}$,

$$\|\widehat{r}\|_\infty \leq (3sB + c\sigma\sqrt{\log n})$$

Step 2. We now bound $\frac{1}{n}\widehat{r}^\mathsf{T}\max(X, X_{k(j)}\mathbf{1})$.

$$\frac{1}{n}\widehat{r}^\mathsf{T}\max(X_k, X_{k(j)}\mathbf{1}) = \frac{1}{n}\sum_{i=1}^n \widehat{r}_i\max(X_{ki}, X_{k(j)}) = \frac{1}{n}\sum_{i=1}^n \widehat{r}_i X_{ki}\delta(ord(i)\geq j) + \frac{1}{n}X_{k(j)}\mathbf{1}_A^\mathsf{T}\widehat{r}_A$$

Where $A = \{i : ord(i) \geq j\}$ and $ord(i)$ is the order of sample i where (1) is the smallest element.

We will bound both terms.

Term 1.

$$\text{Want to bound} \qquad F(X_{k1}, ..., X_{kn}) \coloneqq \frac{1}{n}\sum_{i=1}^n \widehat{r}_i X_{ki}\delta(ord(i)\geq j)$$

First, we note that X_{ki} is bounded in the range $[-b, b]$.

We claim then that F is coordinatewise-Lipschitz. Let $X_k = (X_{k1}, X_{k2}, ..., X_{kn})$ and $X_k' = (X_{k1}', X_{k2}, ..., X_{kn})$ differ only on the first coordinate.

The order of coordinate i in X_k and X_k' can change by at most 1 for $i \neq 1$. Therefore, of the $j-1$ terms of the series, at most 2 terms differ from $F(X_k)$ to $F(X_k')$ and

$$|F(X_{k1}, ..., X_{kn}) - F(X_{k1}', ..., X_{kn}')| \leq \frac{4b\|\widehat{r}\|_\infty}{n}$$

By McDiarmid's inequality therefore,

$$P(|F(X_k) - \mathbb{E}F(X_k)| \geq t) \leq C\exp(-cn\frac{t^2}{(4b\|\widehat{r}\|_\infty)^2})$$

By symmetry and the fact that $\widehat{r}$ is centered, $\mathbb{E}F(X_k) = 0$.

We can fold the 4 into the constant c. With probability $1-\delta$, $|F(X_k)| \leq b\|\widehat{r}\|_\infty\sqrt{\frac{1}{cn}\log\frac{C}{\delta}}$.

Term 2:

$$\text{Want to bound}\frac{1}{n}X_{k(j)}\mathbf{1}_A^\mathsf{T}\widehat{r}_A$$

A is a random set and is probabilistically independent of $\widehat{r}$. $\mathbf{1}_A^\mathsf{T}\widehat{r}_A$ is the sum of a sample of $\widehat{r}$ without replacement. Therefore, according to Serfling's theorem (Corollary 5.2), with probability at least $1-\delta$, $|\frac{1}{n}\mathbf{1}_A^\mathsf{T}\widehat{r}_A|$ is at most $\|\widehat{r}\|_\infty\sqrt{\frac{1}{cn}\log\frac{C}{\delta}}$.

Since $|X_{k(j)}|$ is at most b, we obtain that with probability at least $1-\delta$, $|\frac{1}{n}X_{k(j)}\mathbf{1}_A^\mathsf{T}\widehat{r}_A| \leq b\|\widehat{r}\|_\infty\sqrt{\frac{1}{cn}\log\frac{C}{\delta}}$.

Now we put everything together.

Taking union bound across p and n, we have that with probability at least $1-\delta$,

$$|\frac{1}{n}\max(X_k, X_{k(j)}\mathbf{1})^\mathsf{T}\widehat{r}| \leq b\|\widehat{r}\|_\infty\sqrt{\frac{1}{c}\frac{1}{n}\log\frac{npC}{\delta}}$$

Taking union bound and substituting in the probabilistic bound on $\|\widehat{r}\|_\infty$, we get that with probability at least $1-\frac{C}{n}$,

$|\frac{1}{n}\max(X_k, X_{k(j)}\mathbf{1})^\mathsf{T}\widehat{r}|$ is at most

$$cb(sB+\sigma)\sqrt{\frac{s}{n}\log n\log(pn)}$$

□

5.3 Proof of False Negative Control

Note: 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 will use covering number and uniform convergence and will thus need to first introduce some notations.

5.3.1 Notation

Given samples $X^{(1)}, ..., X^{(n)}$, let f, g be a function and w be a n-dimensional random vector, then we denote $\|f-g+w\|_n^2 := \frac{1}{n}\sum_{i=1}^n(f(X^{(i)}) - g(X^{(i)}) + w_i)^2$. We will also abuse notation and let $\|f+c\|_n^2 = \frac{1}{n}\sum_{i=1}^n(f(X^{(i)})+c)^2$ if c is a scalar.

We let $\langle f, g\rangle_n := \frac{1}{n}\sum_{i=1}^n f(X^{(i)})g(X^{(i)})$. It then follows that:

1. $\|f+g\|_n^2 = \|f\|_n^2 + 2\langle f, g\rangle_n + \|g\|_n^2$
2. $\langle f, g\rangle_n \leq \|f\|_n \|g\|_n$

For a function $g : \mathbb{R}^s \to \mathbb{R}$, define $\widehat{R}_s(g) \coloneqq \|f_0 + w - \bar{f}_0 - \bar{w} - g\|_n^2$ as the objective of the *restricted* regression and define $R_s(g) \coloneqq \mathbb{E}|f_0(X) + w - \mu - g(X)|^2$ as the population risk, where $\bar{f}_0 = \frac{1}{n}\sum_i f_0(X^{(i)})$ and $\bar{w} = \frac{1}{n}\sum_{i=1}^n w_i$ and $\mu = \mathbb{E}f_0(X)$. Note that we *subtract* out the mean in the two risk definitions.

For an additive function g, define $\rho_n(g) = \sum_{k=1}^s \|\partial g_k\|_\infty$. Because we always use the secant linear piece-wise function in our optimization program, we define

$$\|\partial g_k\|_\infty \coloneqq \max_{i=1,\ldots,n-1} |\frac{g_k(X^{(i)}) - g_k(X^{(i+1)})}{X^{(i)} - X^{(i+1)}}|$$

Let $\mathcal{C}[b, B, L]$ be the set of 1 dimensional convex functions on $[-b, b]$ that are bounded by B and L-Lipschitz.

Let $\mathcal{C}[s, b, B, L]$ be the set of additive functions with s components each of which is in $\mathcal{C}[b, B, L]$.

$$\mathcal{C}[s, b, B, L] \coloneqq \{f : \mathbb{R}^s \to \mathbb{R} \,:\, f = \sum_{k=1}^s f_k(x_k),\ f_k \in \mathcal{C}[b, B, L]\}$$

Define $f^{*s} = \arg\min\{R_s(f) \,|\, f \in \mathcal{C}^s[b, B, L],\ \mathbb{E}f_k(X_k) = 0\}$.

Define $f^{*(s-1)} = \arg\min\{R_s(f) \,|\, f \in \mathcal{C}^{(s-1)}[b, B, L],\ \mathbb{E}f_k(X_k) = 0\}$, the optimal solution with only $s-1$ components.

Note: By definition of the Lipschitz condition, $f_k \in \mathcal{C}[b, B, L]$ implies that $\|\partial f_k\|_\infty \leq L$. $f = \sum_k f_k \in \mathcal{C}[s, b, B, L]$ implies that $\rho_n(f) \leq sL$.

5.3.2 Proof

We first start with a lemma that converts assumption A4 into a more easily applicable condition.

Lemma 5.1. *Suppose assumptions A1' and A4 hold.*

Then $R(f^{*(s-1)}) - R(f^{*s}) \geq \alpha$, where α lower bounds the norm of the population optimal additive components as defined in assumpation A4.

Proof.

$$\begin{aligned}
&R(f^{*(s-1)}) - R(f^{*s}) \\
&= \mathbb{E}\Big(f^{*(s-1)}(X) - f_0(X) + \mu\Big)^2 - \mathbb{E}\Big(f^{*s}(X) - f_0(X) + \mu\Big)^2 \\
&= \mathbb{E}\Big(f^{*(s-1)}(X) - f^{*s}(X) + f^{*s}(X) - f_0(X) + \mu\Big)^2 - \mathbb{E}\Big(f^{*s}(X) - f_0(X) + \mu\Big)^2 \\
&= \mathbb{E}\Big(f^{*(s-1)}(X) - f^{*s}(X)\Big)^2 - 2\mathbb{E}\Big[(f^{*(s-1)}(X) - f^{*s}(X))(f^{*s}(X) - (f_0(X) - \mu))\Big]
\end{aligned}$$

We will argue that all the components of the additive function $f^{*(s-1)}$ are also in f^{*s}. Let us denote the components of $f^{*s} = \sum_{k=1}^s f_k^*$. We will now invoke Corollary 2.1, which is valid because we assume $X_1, ..., X_s$ are independent by assumption A1'. By Corollary 2.1, if we set

$f_k^* = 0$, the resulting additive function $\sum_{k' \neq k} f_{k'}^*$ minimizes the population risk subject to the constraint that $f_k = 0$. By definition, f^{*s} is $\arg\min_k \sum_{k' \neq k} f_{k'}^*$ and thus share components with f^{*s}.

Therefore, there exist some k such that $f^{*(s-1)} - f^{*s} = f_k^*$, and we can continue the bound

$$\begin{aligned} R(f^{*(s-1)}) - R(f^{*s}) &= \mathbb{E} f_k^*(X_k)^2 - 2\mathbb{E}[f_k^*(X_k)(f^{*s}(X) - (f_0(X) - \mu))] \\ &= \mathbb{E} f_k^*(X_k)^2 - 2\mathbb{E}[f_k^*(X_k) f^{*s}(X)] + 2\mathbb{E}[f_k^*(X_k)(f_0(X) - \mu)] \end{aligned}$$

$$\mathbb{E}[f_k^*(X_k) f^{*s}(X)] = \mathbb{E}\Big[f_k^*(X_k)\mathbb{E}[f^{*s}(X) \,|\, X_k]\Big] = \mathbb{E} f_k^*(X_k)^2$$
$$\mathbb{E}[f_k^*(X_k)(f_0(X) - \mu)] = \mathbb{E}\Big[f_k^*(X_k)\mathbb{E}[f_0(X) - \mu \,|\, X_k]\Big] = \mathbb{E} f_k^*(X_k)^2$$

Where we used the fact that $\mathbb{E} f_{k'}^*(X_{k'}) = 0$ for all k' and the fact that $\mathbb{E}[f_0(X) - \mu \,|\, X_k] = f_k^*(X_k)$ (Lemma 2.1).

Thus, $R(f^{*(s-1)}) - R(f^{*s}) \geq \mathbb{E} f_k^*(X_k)^2 \geq \alpha$ by Assumption A4.

□

We now restate the theorem in our newly defined notation.

Theorem 5.3. *Suppose assumptions A1', A2', A3, A4 hold.*
Let $\widehat{f} \coloneqq \arg\min\{\widehat{R}_s(f) + \lambda_n \rho_n(f) \,:\, f \in \mathcal{C}[s, b, B, L], f_k \text{ centered}\}$.
Suppose that $csL\lambda_n \to 0$ *and* $cLb(sB + \sigma)sB\sqrt{\frac{s}{n^{4/5}} \log sn} \to 0$.
Then, for all large enough n*, with probability at least* $1 - \frac{C}{n}$, $\widehat{f}_k \neq 0$ *for all* $k = 1, ..., s$.

Proof. Let us first sketch out the rough idea of the proof. We know that in the population setting, the best approximate additive function f^{*s} has s non-zero components. We also know that the empirical risk approaches the population risk uniformly. Therefore, it cannot be that the empirical risk minimizer maintains a zero component for all n; if that were true, then we can construct a feasible solution to the empirical risk optimization, based on f^{*s}, that achieves lower empirical risk.

Step 1: f^{*s} is not directly a feasible solution to the empirical risk minimization program because it is not empirically centered. Given n samples, $f^{*s} - \bar{f}^{*s}$ is a feasible solution where $\bar{f}^{*s} = \sum_{k=1}^s \bar{f}_k^{*s}$ and $\bar{f}_k^{*s} = \frac{1}{n}\sum_{i=1}^n f_k^{*s}(X^{(i)})$.

$$\begin{aligned} |\widehat{R}_s(f^{*s} - \bar{f}^{*s}) - \widehat{R}_s(f^{*s})| &\leq \|f_0 + w - \bar{f}_0 - \bar{w} - f^{*s} + \bar{f}^{*s}\|_n^2 - \|f_0 + w - \bar{f}_0 - \bar{w} - f^{*s}\|_n^2 \\ &\leq 2\langle f_0 + w - \bar{f}_0 - \bar{w} - f^{*s}, \bar{f}^{*s}\rangle_n + \|\bar{f}^{*s}\|_n^2 \\ &\leq 2\|f_0 + w - \bar{f}_0 - \bar{w} - f^{*s}\|_n \|\bar{f}^{*s}\|_n + \|\bar{f}^{*s}\|_n^2 \\ &\leq 2|\bar{f}^{*s}| \|f_0 + w - \bar{f}_0 - \bar{w} - f^{*s}\|_n + |\bar{f}^{*s}|^2 \end{aligned}$$

Because each f^{*s} is bounded by sB and $\mathbb{E} f^{*s}(X) = 0$, by Hoeffding inequality, with probability at least $1 - \frac{C}{n}$, $|\bar{f}^{*s}| \leq sB\sqrt{\frac{1}{cn} \log n}$.

$$\|f_0 + w - \bar{f}_0 - \bar{w} - f^{*s}\|_n \leq \|f_0 - f^{*s}\|_n + \|w\|_n + |\bar{f}_0| + |\bar{w}|$$

$\|f_0 - f^{*s}\|_n \leq \|f_0 - f^{*s}\|_\infty$ is bounded by $2sB$ and w_i is zero-mean subgaussian with scale σ. Therefore, $\|w\|_n$ is at most $c\sigma$ with probability at least $1 - \frac{C}{n}$ for all $n > n_0$.

$|\bar{f}_0| \leq sB$ and $|\bar{w}| \leq c\sigma\sqrt{\frac{1}{n}}$ with probability at least $1 - \frac{C}{n}$ by Chernoff bound.

So we derive that, with probability at least $1 - \frac{C}{n}$, for all $n > n_0$,

$$|\widehat{R}_s(f^{*s} - \bar{f}^{*s}) - \widehat{R}_s(f^{*s})| \leq 2c(sB + \sigma)sB\sqrt{\frac{1}{cn}\log n}$$

Step 2: Now that we bounded the cost of approximating f^{*s} with the empirically centered $f^{*s} - \bar{f}^{*s}$, we move on to the proof of the main result.

Suppose $\widehat{f}$ has at most $s - 1$ non-zero components. Then

$$\begin{aligned}
\widehat{R}_s(\widehat{f}) &\geq R_s(\widehat{f}) - \tau_n \\
&\geq R_s(f^{*(s-1)}) - \tau_n \\
&\geq R_s(f^{*s}) + \alpha - \tau_n \\
&\geq \widehat{R}_s(f^{*s}) + \alpha - 2\tau_n \\
&\geq \widehat{R}_s(f^{*s} - \bar{f}^{*s}) - \tau_n' + \alpha - 2\tau_n
\end{aligned}$$

The third line follows from Lemma 5.1. τ_n is the deviation between empirical risk and true risk and τ_n' is the approximation error incurred by empirically sampling f^{*s}.

Adding and subtracting $\lambda_n\rho_n(f^{*s} - \bar{f}^{*s})$ and $\lambda_n\rho_n(\widehat{f})$, we arrive at the conclusion that

$$\widehat{R}_s(\widehat{f}) + \lambda_n\rho_n(\widehat{f}) \geq \widehat{R}_s(f^{*s} - \bar{f}^{*s}) + \lambda_n\rho_n(f^{*s} - \bar{f}^{*s}) - (\lambda_n\rho_n(f^{*s} - \bar{f}^{*s}) + \lambda_n\rho_n(\widehat{f})) - \tau_n' + \alpha - 2\tau_n$$

Because we assume that we impose the Lipschitz constraint in our optimization, $\rho_n(\widehat{f}), \rho_n(f^{*s} - \bar{f}^{*s})$ are at most sL and so $|\lambda_n\rho_n(\widehat{f}) - \lambda_n\rho_n(f_s^*)| \leq 2sL\lambda_n$.

By Theorem 5.4, we know that under the condition of the theorem, $\tau_n \leq Lb(sB+\sigma)sB\sqrt{\frac{s}{cn^{4/5}}\log n}$.

τ_n', as shown above, is at most $2(sB + \sigma)sB\sqrt{\frac{1}{cn}\log sn}$ with probability at least $1 - \frac{C}{n}$ for $n > n_0$.

For n large enough such that

$$csL\lambda_n < \frac{\alpha}{2} \text{ and } LbsB(sB + \sigma)\sqrt{\frac{s}{n^{4/5}}\log sn}) < \frac{\alpha}{4}$$

we get that $\widehat{R}_s(\widehat{f}) + \lambda_n\rho_n(\widehat{r}) > \widehat{R}_s(f_s^*) + \lambda_n\rho_n(f_s^*)$, which is a contradiction since we assumed that $\widehat{f}$ minimizes the regularized empirical risk.

□

Theorem 5.4. *(Uniform Risk Deviation) For all $n > n_0$, we have that, with probability at least $1 - \frac{C}{n}$,*

$$\sup_{f \in \mathcal{C}^s[b,B,L]} |\widehat{R}_s(f) - R_s(f)| \leq Lb(sB + \sigma)sB\sqrt{\frac{s}{cn^{4/5}} \log sn}$$

Proof. This proof uses a standard covering number argument.

Let $\mathcal{C}_\epsilon[s, b, B, L]$ be an ϵ-cover of $\mathcal{C}[s, b, B, L]$ such that for all $f \in \mathcal{C}[s, b, B, L]$, there exists $f' \in \mathcal{C}_\epsilon[s, b, B, L]$ such that $\|f - f'\|_\infty \leq \epsilon$.

For all $f \in \mathcal{C}[s, b, B, L]$,

$$\widehat{R}_s(f) - R_s(f) = \widehat{R}_s(f) - \widehat{R}_s(f') + \widehat{R}_s(f') - R_s(f') + R_s(f') - R_s(f)$$

where $f' \in \mathcal{C}_\epsilon[s, b, B, L]$ and $\|f - f'\|_\infty \leq \epsilon$.

Step 1. We first bound $\widehat{R}_s(f) - \widehat{R}_s(f')$.

$$\begin{aligned}
|\widehat{R}_s(f) - \widehat{R}_s(f')| &= |\|f_0 - \bar{f}_0 + w - \bar{w} - f\|_n^2 - \|f_0 - \bar{f}_0 + w - \bar{w} - f'\|_n^2| \\
&\leq 2\langle f_0 - \bar{f}_0 + w - \bar{w}, f' - f\rangle_n + \|f\|_n^2 - \|f'\|_n^2 \\
&\leq 2\|f_0 - \bar{f}_0 + w - \bar{w}\|_n \|f' - f\|_n + (\|f\|_n - \|f'\|_n)(\|f\|_n + \|f'\|_n)
\end{aligned}$$

We now want to bound $\|f_0 - \bar{f}_0 + w - \bar{w}\|_n \leq \|f_0\|_n + \|w\|_n + |\bar{f}| + |\bar{w}|$.

$\|w\|_n^2 = \frac{1}{n}\sum_{i=1}^n w_i^2$ is the average of subexponential random variables. Therefore, for all n larger than some absolute constant n_0, with probability at least $1 - \frac{C}{n}$, $|\|w\|_n^2 - \mathbb{E}|w|^2| < \sigma^2\sqrt{\frac{1}{cn}\log n}$. The absolute constant n_0 is determined so that for all $n > n_0$, $\sqrt{\frac{1}{cn}\log n} < 1$. Since $\mathbb{E}w^2 \leq \sigma^2$, for all $n > n_0$, with probability at least $1 - \frac{C}{n}$, for some constant c, $\|w\|_n^2 \leq c\sigma^2$.

By Chernoff bound and the fact that $\mathbb{E}w = 0$, we know that $|\bar{w}| \leq c\sigma\sqrt{\frac{1}{n}}$ with high probability.

$\|f_0\|_n^2 = \frac{1}{n}\sum_{i=1}^n f_0(X^{(i)})^2$. Since $f_0(X^{(i)})^2 \leq s^2B^2$, $\|f_0\|_n^2 \leq s^2B^2$.

$|\bar{f}_0| = |\frac{1}{n}\sum_{i=1}^n f_0(X^{(i)})| \leq sB$.

Combining these together, We have that for all $n \geq n_0$, with probability at least $1 - \frac{C}{n}$, $\|f_0 - \bar{f}_0 + w - \bar{w}\|_n^2 \leq c(s^2B^2 + \sigma^2)$, and so

$$\|f_0 - \bar{f}_0 + w - \bar{w}\|_n^2 \leq c(sB + \sigma)$$

$\|f' - f\|_\infty \leq \epsilon$ implies that $\|f' - f\|_n \leq \epsilon$. And therefore, $\|f\|_n - \|f'\|_n \leq \|f - f'\|_n \leq \epsilon$.

f, f' are all bounded by sB, and so $\|f\|_n, \|f'\|_n \leq sB$.

Thus, we have that, for all $n > n_0$,

$$|\widehat{R}_s(f) - \widehat{R}_s(f')| \leq \epsilon c(sB + \sigma) \tag{5.1}$$

with probability at least $1 - \frac{C}{n}$.

Step 2: Now we bound $R_s(f') - R_s(f)$. The steps follow the bounds before, and we have that

$$|R_s(f') - R_s(f)| \leq \epsilon c(sB + \sigma) \tag{5.2}$$

Step 3: Lastly, we bound $\sup_{f' \in \mathcal{C}_\epsilon[s,b,B,L]} \widehat{R}_s(f') - R_s(f')$.
For a fixed f', we have that, by definition

$$\begin{aligned}
\widehat{R}_s(f') &= \|f_0 + w - \bar{f}_0 - \bar{w} - f'\|_n^2 \\
&= \|f_0 - \bar{f}_0 - f'\|_n^2 + 2\langle w - \bar{w}, f_0 - \bar{f}_0 - f'\rangle_n + \|w - \bar{w}\|_n^2 \\
R_s(f') &= \mathbb{E}\big(f_0(X) + w - \mu - f'(X)\big)^2 \\
&= \mathbb{E}\big(f_0(X) - \mu - f'(X)\big)^2 + \mathbb{E}w^2
\end{aligned}$$

Therefore:

$$\begin{aligned}
\widehat{R}_s(f') - R_s(f') =& \|f_0 - \bar{f}_0 - f'\|_n^2 - \mathbb{E}(f_0(X) - \mu - f'(X))^2 \\
&+ \|w - \bar{w}\|_n^2 - \mathbb{E}w^2 \\
&+ 2\langle w - \bar{w}, f_0 - \bar{f}_0 - f'\rangle_n
\end{aligned}$$

Step 3.1: We first bound $2\langle w - \bar{w}, f_0 - \bar{f}_0 - f'\rangle_n$.

$$\langle w - \bar{w}, f_0 - \bar{f}_0 - f'\rangle_n = \langle w, f_0 - \bar{f}_0 - f'\rangle_n - \langle \bar{w}, f_0 - \bar{f}_0 - f'\rangle_n$$

The first term, fully expanded, is $\frac{1}{n}\sum_{i=1}^n w_i(f_0(X^{(i)}) - \bar{f}_0 - f'(X^{(i)}))$. Since w_i and $X^{(i)}$ are independent, we use the sub-Gaussian concentration inequality. Note that $|f_0(X^{(i)}) - \bar{f}_0 - f'(X^{(i)})| \leq 3sB$, and so $|\langle w, f_0 - \bar{f}_0 - f'\rangle_n| > t$ with probability at most $C\exp\big(-cnt^2\frac{1}{\sigma^2(sB)^2}\big)$.

The second term, fully expanded, is $\bar{w}\bar{f}'$. $\bar{f}' \leq sB$ and so $|\bar{w}\bar{f}'| > t$ with probability at most $C\exp\big(-cnt^2\frac{1}{\sigma^2(sB)^2}\big)$ as well.

Step 3.2 We now bound $\|w - \bar{w}\|_n^2 - \mathbb{E}w^2$.

$$\begin{aligned}
\|w - \bar{w}\|_n^2 &= \|w\|_n^2 - 2\langle w, \bar{w}\rangle_n + \|\bar{w}\|_n \\
&= \|w\|_n^2 - \bar{w}^2
\end{aligned}$$

Using sub-Exponential concentration, we know that $|\|w\|_n^2 - \mathbb{E}w^2| \geq t$ with probability at most $C\exp(-cn\frac{1}{\sigma^2})$.

$\bar{w} \leq \sigma\sqrt{\frac{1}{cn}}$ with probability at least $1 - \frac{C}{n}$. Thus, $|\bar{w}|^2 \leq \sigma^2\frac{1}{cn}$ with high probability, has a second order effect, and can be safely ignored in the bound.

Step 3.3: We now bound $\|f_0 - \bar{f}_0 - f'\|_n^2 - \mathbb{E}(f_0(X) - \mu - f'(X))^2$.

$$\begin{aligned}
\|f_0 - \bar{f}_0 - f'\|_n^2 &= \|f_0 - \mu + \mu - \bar{f}_0 - f'\|_n^2 \\
&= \|f_0 - \mu - f'\|_n^2 + 2\langle \mu - \bar{f}_0, f_0 - \mu - f'\rangle_n + \|\mu - \bar{f}_0\|_n^2
\end{aligned}$$

Using similar reasoning as before, we know that $|\langle \mu - \bar{f}_0, f_0 - \mu - f' \rangle_n| \geq t$ with probability at most $C \exp(-cnt^2 \frac{1}{(sB)^4})$.

Likewise, $|\mu - \bar{f}_0| \leq sB\sqrt{\frac{1}{cn}}$ with probability at least $1 - \frac{C}{n}$. Thus, $|\mu - \bar{f}_0|^2 \leq (sB)^2 \frac{1}{cn}$, has a second order effect, and can be safely ignored in the bound.

Because $f_0(X^{(i)}) - \mu - f'^{(}X^{(i)})$ is bounded by $3sB$, $\|f_0 - \mu - f'\|_n^2$ is the empirical average of n random variables bounded by $9(sB)^2$.

Using Hoeffding Inequality then, we know that the probability $\left| \|f_0 - \mu - f'\|_n^2 - \mathbb{E}(f_0(X) - \mu - f'(X))^2 \right| \geq t$ is at most $C \exp(-cnt^2 \frac{1}{(sB)^4})$.

Applying union bound, we have that $\sup_{f' \in \mathcal{C}_\epsilon[s,b,B,L]} |\widehat{R}_s(f') - R_s(f')| \geq t$ occurs with probability at most

$$C \exp\left(s \left(\frac{bBLs}{\epsilon}\right)^{1/2} - cnt^2 \frac{1}{\sigma^2 (sB)^2 + (sB)^4}\right)$$

for all $n > n_0$.

Restating, we have that with probability at most $1 - \frac{1}{n}$, the deviation is at most

$$(sB + \sigma) sB \sqrt{\frac{1}{cn} \left(\log Cn + s(\frac{bBLs}{\epsilon})^{1/2}\right)} \tag{5.3}$$

Substituting in $\epsilon = \frac{bBLs}{n^{2/5}}$, expression 5.3 can be upper bounded by $sB(\sigma + sB)\sqrt{\frac{s}{cn^{4/5}} \log Cn}$.

Expressions 5.1 and 5.2 from **Step 1** and **Step 2** become $\sqrt{\frac{(bBLs)^2}{cn^{4/5}}}(sB + \sigma)$.

We can arrive at the statement of the theorem by summing these up and absorbing any constants into the symbols c and C.

□

5.4 Supporting Technical Material

5.4.1 Concentration of Measure

Sub-Exponential random variable is the square of a subgaussian random variable[17].

Proposition 5.1. *(Subexponential Concentration [17]) Let $X_1, ..., X_n$ be zero-mean independent subexponential random variables with subexponential scale K.*

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

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 at most

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

Corollary 5.1. *Let $w_1, ..., w_n$ be n independent subgaussian random variables with subgaussian 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 $1-\frac{1}{n}$,

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

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 $1-\frac{1}{n}$,

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

□

5.4.2 Sampling Without Replacement

Lemma 5.2. *(Serfling [16]) Let $x_1, ..., x_N$ be a finite list, $\bar{x} = \mu$. 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

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

Corollary 5.2. *Suppose $\mu = 0$.*

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

And, by union bound, we have that

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

A simple restatement. 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.

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

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

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

This completes the proof.

□

5.4.3 Covering Number for Lipschitz Convex Functions

Definition 5.1. $\{f_1, ..., f_N\} \subset \mathcal{C}[b, B, L]$ is an ϵ-covering of $\mathcal{C}[b, B, L]$ if for all $f \in \mathcal{C}[b, B, L]$, there exist f_i such that $\|f - f_i\|_\infty \leq \epsilon$.

We define $N_\infty(\epsilon, \mathcal{C}[b, B, L])$ as the size of the minimum covering.

Lemma 5.3. *(Bronshtein 1974)*

$$\log N_\infty(\epsilon, \mathcal{C}[b, B, L]) \leq C\left(\frac{bBL}{\epsilon}\right)^{1/2}$$

For some absolute constant C.

Lemma 5.4.

$$\log N_\infty(\epsilon, \mathcal{C}^s[b, B, L]) \leq Cs\left(\frac{bBLs}{\epsilon}\right)^{1/2}$$

For some absolute constant C.

Proof. Let $f = \sum_{k=1}^s f_k$ be a convex additive function. Let $\{f'_k\}_{k=1,..,s}$ be k functions from a $\frac{\epsilon}{s}$ L_∞ covering of $\mathcal{C}[b, B, L]$.

Let $f' \coloneqq \sum_{k=1}^s f'_k$, then

$$\|f' - f\|_\infty \leq \sum_{k=1}^s \|f_k - f'_k\|_\infty \leq s\frac{\epsilon}{s} \leq \epsilon$$

Therefore, a product of s $\frac{\epsilon}{s}$-coverings of univariate functions induces an ϵ-covering of the additive functions. □