A Linear Fit Gets the Correct Monotonicity Directions[*]

Malik Magdon-Ismail
CS Dept., RPI, Rm 207 Lally,
110 8th Street, Troy, NY 12180, USA.
Email: magdon@cs.rpi.edu.

Joseph Sill
Citadel Investment Group, LLC,
131 South Dearborn Street,
Chicago, IL 60603, USA.
Email: joe_sill@yahoo.com.

September 17, 2007

Abstract

Let f be a function on $\mathbb{R}^d$ that is monotonic in every variable. There are 2^d possible assignments to the directions of monotonicity (two per variable). We provide sufficient conditions under which the optimal linear model obtained from a least squares regression on f will identify the monotonicity directions correctly. We show that when the input dimensions are independent, the linear fit correctly identifies the monotonicity directions. We provide an example to illustrate that in the general case, when the input dimensions are not independent, the linear fit may not identify the directions correctly. However, when the inputs are jointly Gaussian, as is often assumed in practice, the linear fit will correctly identify the monotonicity directions, even if the input dimensions are dependent. Gaussian densities are a special case of a more general class of densities (Mahalanobis densities) for which the result holds. Our results hold when f is a classification or regression function.

If a finite data set is sampled from the function, we show that if the exact linear regression would have yielded the correct monotonicity directions, then the sample regression will also do so asymptotically (in a probabilistic sense). This result holds even if the data are noisy.

[*] This work was partially presented as "Using a Linear Model to Determine Monotonicity Directions", at the 16th Conference on Learning Theory (COLT 2003).

1 Introduction and Background

In a typical learning setting, one wishes to determine a target function f from a representative data set. Here, we consider the case when f is monotonic, and so in order to "learn" f one would like to enforce the constraint that the learned function be monotonic. For example, the expected creditworthiness of an individual would be a monotonic function of variables such as income, [18]. The likelihood of a heart condition (as could be measured by (say) the probability of a heart attack within the next year) should be a monotonic function of cholesterol level. The function learned from the finite data set is typically used for predictive purposes. In such a case, incorporating the monotonicity constraint can significantly enhance the performance of the resulting predictor, because the capacity[1] of monotonic functions can be considerably less than the capacity of an unrestricted class, which has consequences on the generalization ability of the learned function, [16, 18, 20]. The immediate challenge is to determine in which direction the function is monotonically increasing.

In a more general setting, inference with monotonicity constraints has been referred to as order restricted inference [2, 14] (inference when restrictions are placed on the ordering of the predicted target values). One example of order restriction is that the target values are monotonically increasing. A survey of research in this area reveals that the most interesting and important open problems are

(i) Testing whether a monotone relation exists.

(ii) If there is a monotone relationship, how to estimate the regression function.

(iii) The asymptotic or finite sample properties of isotonic estimators.

The problem we address sits within area (ii). Specifically, a monotone relationship is known to exist, however in order to estimate the regression function, we need to first determine the directions of monotonicity. In particular, we study the properties of using a particular method for determining the monotonicity directions using a linear model. We show that in a many cases of practical interest, this approach works.

[1] The capacity of a set of functions is the number of data points for which a random dichotomy is separable with probability $\frac{1}{2}$. The capacity is related to the expected number of dichotomies that the set of functions can implement on a set of points. Some basic properties of capacity,VC dimension, etc., can be found in [16, 20].

Testing for the monotonocity of a regression function has been considered in the literature, see for example [5, 15]. Approximate testing of monotonicity has also been considered; for example, in the domain of boolean functions, the task is to determine with high probability if a function is approximately monotonic [1, 7]. Algorithms for enforcing monotonicity have also been considered, for example [3, 9, 10, 11, 12, 13, 17, 18]. Most of these (especially the nonparametric regression approaches) focus on the single variable case, and it is always assumed that the monotonicity direction is known (usually positive).

For the credit and heart problems above, it is reasonable to guess that the direction of the monotonicity is positive. However, it can often be the case that while monotonicity is known to hold, the direction of monotonicity is not known, and needs to be determined. An example is when the identity of the variables is kept secret for privacy or propriety reasons. We draw on an example considered in [18] as motivation for our result. A problem considered in [18] is that of predicting credit quality from a set of indicator variables. For privacy reasons, the identity of the variables (as well as identifiers of the individuals) were hidden. However, one expects that credit quality should be a monotonic function of most reasonable indicator variables. Some specific examples are that the probability of default on a loan should be decreasing in the individuals income, decreasing in the number of years of higher education, increasing in the level of outstanding debt. Given that the indicator variables have been standardized, and we do not know which variables are which, if we wish to incorporate monotonicity into the learning, we are faced with the task of determining the monotonicity directions of the target function to be learned, in each of the indicator variables. In fact, this was exactly the problem faced in [18] and the approach which they adopted was to estimate the monotonicity directions using a linear model. One important observation, which will be relevant later is that the indicator variables need not be independent, as would be the case with income and number of years of education. Another example we mention is the general multi-level problem studied in [9]. This is also a domain in which monotonicity is expected to hold. The multilevel problem is a classification problem with some structure among the classes. Specifically, each class is broken into subclasses or levels, which intuitively represent the "severity" of the class. Two examples are disease prediction (for example cardiac disease) and fault prediction in machinary. Focussing on fault prediction, the classes in this problem represent the types of faults. The levels represent the severities of the faults. The task is to predict both the fault and its severity. The severity should be monotonic in the indicators: for example, if increasing one particular indicator variable changes the classification from normal operation to the mild case of a particular fault, then one expects that increasing that same variable should increase the severity of the fault. This is a an example where the target function (severity level) in a classification problem is expected to be monotonic in its indicator variables, though it may not be known *apriori* what the monotonicity directions are.

In d-dimensions, each indicator variable has two possible monotonicity directions, yielding 2^d possible choices for all the monotonicity directions. It is therefore not practical to enforce monotonicity in each of these 2^d choices and then pick the "best fit" among these 2^d choices, especially when d is large. Rather, one would like to determine a specific direction in which to enforce the monotonicity, before performing the non-linear regression or learning. The approach which we investigate is to use a simple model, the linear model, to determine the monotonicity directions. Having determined the monotonicity directions, they can then be used as constraints in the more complex regression.

1.1 Preliminary Definitions and Results

Definition 1.1 *A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is said to be monotonic with positive direction in dimension i if*

$$f(x_1, \ldots, x_{i-1}, x_i + \Delta, x_{i+1}, \ldots, x_d) \geq f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d), \tag{1}$$

for all $\Delta > 0$ and all $\mathbf{x} \in \mathbb{R}^d$. The direction of monotonicity is negative if the condition $\Delta > 0$ is replaced by $\Delta < 0$.

When the context is clear, we will use the notation $f_i(x_i; \mathbf{x}_{-i})$ to denote the function f of x_i with all other variables held constant (the vector $\mathbf{x}_{-i}$ refers to the values at which the other variables are held constant). When the context is clear, we will drop the dependence on $\mathbf{x}_{-i}$. We assume throughout that we are in $\mathbb{R}^d$.

A function is *monotonic* if it is monotonic in every dimension [2]. If f is only defined on some subset of $\mathbb{R}^d$, then the monotonicity conditions need hold only in this subset. One can also extend the definition of monotonicity to one which incorporates a probability distribution $p(\mathbf{x})$, specifically,

[2] Such functions are sometimes referred to as unate.

f need only satisfy the monotonicity condition on the support of $p(\mathbf{x})$.

We can represent the monotonicity directions of such a function by a d dimensional vector $\mathbf{m}$ of ±1's. There are 2^d possible choices for $\mathbf{m}$. A classification function, $f : \mathbb{R}^d \mapsto \{+1, -1\}$ is monotonic if it can be represented as $f(\mathbf{x}) = \text{sign}(g(\mathbf{x}))$, where g is a monotonic function. Condition (1) can now be more compactly written as $f_i(x_i + m_i\Delta; \mathbf{x}_{-i}) \geq f_i(x_i; \mathbf{x}_{-i})$, for all $\Delta \geq 0$ and for all $\mathbf{x}_{-i}$.

A linear function l is defined by $l(\mathbf{x}; \mathbf{w}, w_0) = \mathbf{w}^T\mathbf{x} + w_0$. Since a linear model is monotonic, one approach to determining the monotonicity directions of f would be to fit a linear model to the data, and use the monotonicity direction implied by the optimal linear model as an estimate of the monotonicity direction of f. Such an approach was used in [18]. The purpose here is to show that such an approach is valid. Assume that the inputs are distributed according to $p_{\mathbf{x}}(\mathbf{x})$. The expected mean square error $\mathcal{E}$ of the linear function $l(\mathbf{x}; \mathbf{w}, w_0)$ is given by

$$\mathcal{E}(\mathbf{w}, w_0) = \int d\mathbf{x}\ p_{\mathbf{x}}(\mathbf{x})\ (\mathbf{w}^T\mathbf{x} + w_0 - f(\mathbf{x}))^2. \tag{2}$$

The optimal linear fit (which we will refer to more simply as the linear fit) is given by the choice of $\mathbf{w}$ and w_0 that minimize $\mathcal{E}(\mathbf{w}, w_0)$. We will assume throughout that the linear fit exists. Without loss of generality, we can also assume that $E[\mathbf{x}] = \mathbf{0}$ (this is formally justified in Lemmas 2.2, 2.3). First we state how to obtain the linear fit.

Lemma 1.2 (Linear fit.) *Let* $\mathbf{\Sigma} = \int d\mathbf{x}\ p_{\mathbf{x}}(\mathbf{x})\ \mathbf{x}\mathbf{x}^T$ *be invertible. The linear fit is then given by*

$$\mathbf{w}^l = \mathbf{\Sigma}^{-1} \int d\mathbf{x}\ p_{\mathbf{x}}(\mathbf{x})\ f(\mathbf{x})\mathbf{x}, \qquad w_0^l = \int d\mathbf{x}\ p_{\mathbf{x}}(\mathbf{x})\ f(\mathbf{x}) \tag{3}$$

Proof: We refer to any standard book on statistics for a proof, for example [6]. ■

From now on, we will assume that $\mathbf{\Sigma}$ is invertible to simplify the analysis. This is a reasonable assumption as long as the input distribution has support with finite measure. For degenerate input distributions, which include distributions like delta functions, this assumption will not hold. However, such input distributions are not typical from the point of view of learning.

Practically, from the learning perspective, one does not have access to the target function $f(\mathbf{x})$, which is assumed to be monotonic, nor does one have access to the input distribution $p_{\mathbf{x}}(\mathbf{x})$. Rather, one has a data set, $\mathcal{D}_N = \{\mathbf{x}_i, y_i\}_{i=1}^N$. The particular way in which the data set was sampled defines the regression model. The model we will assume is the standard homoskedastic[3] regression model. $\mathbf{x}_i$ are sampled independently from $p_{\mathbf{x}}(\mathbf{x})$ and $y_i = f(\mathbf{x}_i) + \epsilon_i$, where ϵ_i is noise. In the regression case, we assume that the ϵ_i are independent zero mean noise, with bounded fourth moments.

$$E[\epsilon_i|\mathbf{x}_i] = 0, \quad E[\epsilon_i\epsilon_j|\mathbf{x}_i, \mathbf{x}_j] = \sigma^2\delta_{ij}, \tag{4}$$

where δ_{ij} is the Kronecker delta function. Often, one assumes the noise to be Gaussian, but this is not a necessary requirement. For technical reasons, we will generally assume that all fourth moments that include powers of the noise variable, powers of $\mathbf{x}$ and powers of f are bounded. For example, $E[f^2(\mathbf{x})\mathbf{x}\mathbf{x}^T] < \infty$, etc. Some of these restrictions can be dropped, however for simplicity, we maintain them. For the classification case, we assume that the noise ϵ_i is independent flip noise, i.e., independent flips of the output values from y_i to $-y_i$ with some probability $p < \frac{1}{2}$.

$$\epsilon_i = \begin{cases} 0 & \text{w.p. } 1-p, \\ -2f(\mathbf{x}_i) & \text{w.p. } p. \end{cases} \tag{5}$$

Define the augmented input vector by

$$\hat{\mathbf{x}}_i = \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix},$$

and define $\mathbf{X}_N$ by

$$\mathbf{X}_N = \frac{1}{N}\sum_{i=1}^N \hat{\mathbf{x}}_i\hat{\mathbf{x}}_i^T = \frac{1}{N}\sum_{i=1}^N \begin{bmatrix} 1 & \mathbf{x}_i^T \\ \mathbf{x}_i & \mathbf{x}_i\mathbf{x}_i^T \end{bmatrix}.$$

An approximation to the linear fit is given by the Ordinary Least Squares (OLS) estimator, which minimizes the sample average of the squared error. The OLS estimator is given in the following lemma,

Lemma 1.3 *The OLS estimates $w_0^*, \mathbf{w}^*$, of $w_0^l, \mathbf{w}^l$ are given by*

$$\boldsymbol{\beta}^* = \begin{bmatrix} w_0^* \\ \mathbf{w}^* \end{bmatrix} = \frac{\mathbf{X}_N^{-1}}{N}\sum_{i=1}^N y_i\hat{\mathbf{x}}_i.$$

[3] *iid* noise random variables ϵ_i added to the measured function value for each data point $\mathbf{x}_i$.

Proof: See any standard book on statistics, for example [6]. ■

The statement of Lemma 1.3 assumes the existence of $\mathbf{X}_N^{-1}$. If this inverse does not exist, then typically one uses the pseudo-inverse, which corresponds to constructing the best fit (which is not unique) having the smallest norm. From now on, we will always assume that $\mathbf{X}_N^{-1}$ exists to avoid unnecessary technical dificulties. Asymptotically, this is a negligible assumption because it is true with probability approaching 1 since $\mathbf{X}_N^{-1} \to \hat{\mathbf{\Sigma}}^{-1}$ (see Lemma 1.5).

Under reasonable conditions, when $N \to \infty$, sample averages converge to their expectations. The next two lemmas summarize these facts more precisely. We use the standard notation $\xrightarrow{P}$ to denote convergence in probability.

Lemma 1.4 *Let Y_N, Z_N be random variables such that $Y_N \xrightarrow{P} Z_N$, and let g be a continuous function. Then $g(Y_N) \xrightarrow{P} g(Z_n)$. Further, if Z_N is the constant z, then g need only be continuous at z.*

Proof: See for example [4]. ■

Lemma 1.5

$$\frac{1}{N}\sum_i \mathbf{x}_i \xrightarrow{P} E[\mathbf{x}] = \mathbf{0}, \quad \mathbf{X}_N \xrightarrow{P} \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{\Sigma} \end{bmatrix}, \quad \mathbf{X}_N^{-1} \xrightarrow{P} \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{\Sigma}^{-1} \end{bmatrix}, \quad \sum_{i=1}^{N} y_i \hat{\mathbf{x}}_i \xrightarrow{P} \begin{bmatrix} E[f(\mathbf{x})] \\ E[f(\mathbf{x})\mathbf{x}] \end{bmatrix}.$$

Proof: The first, second and fourth limits follow by the weak law of large numbers because the fourth order moments are bounded (see for example [4]). Since $\mathbf{\Sigma}$ is invertible, the function $\mathbf{X}_N^{-1}$ is continuous at $\mathbf{X}_N = \mathbf{\Sigma}$. Therefore, by Lemma 1.4, the third limit also holds. ■

Thus, the OLS estimates should converge to the true linear fit. This will be made more precise in Section 3.

1.2 Contribution

The main contribution of this paper is to determine conditions under which $\mathbf{w}^l$ from the linear fit in (3) will produce the correct monotonicity directions for the function $f(\mathbf{x})$. First we will give the result for independent input variables, which essentially states that when the input density factors into a product of marginals, the linear fit reproduces the correct monotonicity directions, even though f may not resemble a linear function in any way (see Theorem 2.1). Further, note that Theorem 2.1 does not differentiate between classification or regression functions, and thus the optimal linear fit for a classification problem will also yield the correct directions of monotonicity. An immediate corollary of this theorem is that when the input dimension is $d = 1$, the linear fit will always yield the correct monotonicity direction. This is a conclusion implied by standard results in statistics, because in one dimension, if $E[x] = 0$, then $\mathbf{w}^l$ is proportional to $cov(x, f(x))$, and it is well known that if f is monotonically increasing, then $cov(x, f(x)) > 0$. Thus, Theorem 2.1 can be viewed as a generalization of this result to multi-dimensional independent input variables. An important special case is when the function is defined on a hyper–rectangle, and the measure is uniform on the rectangle.

Independence in the input variables is quite a strong restriction, and much of the benefit of the monotonicity constraint is due to the fact that the input variables are *not independent* (for example number of years of higher education and income would be highly correlated in the credit default application). This is evident from the fact that the VC-dimencion of the class of monotonic classification functions is ∞, but the capacity of this class is heavily dependent on the input distribution (see [16] for a discussion of this issue). When the input variables are independent, the capacity of the class of monotonic functions grows exponentially in d [16], but when the input variables are dependent, the capacity can be a much more slowly growing function. Such issues are discussed in greater detail in [16], where the author gives efficient algorithms for computing the capacity and gives some experimental results on the growth of the capacity with d for independent and dependent inputs. The basic conclusion is that the if there is enough dependence in the inputs, the capacity can grow much more slowly than in the independent case. There are (to our knowledge) no theoretical results computing the capacity for dependent inputs. From the learning point of view, the impact of a monotonicity constraint in improving the generalization capability of a learning model is greatest when the inputs are dependent, thus being able to extract the monotonicity directions when the inputs are dependent is important. We show by example (in Proposition 2.5) that if we remove the independence requirement, then we cannot guarantee that the optimal linear fit will induce the correct monotonicity directions.

While we cannot remove the independence restriction in general, for certain special classes of input densities, we show that our result applies even when there is dependence. In particular, a common assumption is that the inputs are jointly Gaussian. In this case, the linear fit will correctly induce the monotonicity directions even when there is dependence in the inputs. This result is a special case of a more general one dealing with a class of input densities which we call *Mahalanobis densities.*

Definition 1.6 *A density $p_{\mathbf{x}}(\mathbf{x})$ is a Mahalanobis density if it can be written as*

$$p_{\mathbf{x}}(\mathbf{x}) = g\left((\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

The mean vector and covariance matrix are given by $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively. $g(x)$ is a function defined on $\mathbb{R}_+$ that is the derivative of a non-decreasing function $G(x) < 0$, i.e., $g(x) = G'(x)$. By definition, $\boldsymbol{\Sigma} = \int d\mathbf{x}\ g(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})\mathbf{x}\mathbf{x}^T$. Further, we require the following constraints on $G(x)$: $\lim_{|x|\to\infty} G(x^2)x = 0$; $\int d\mathbf{x}\ G'(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}^T) = 1$; $\int d\mathbf{x}\ G(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}^T) = -2$. $G(x)$ is called the associated Mahalanobis distribution function.

The first constraint on G is merely technical, stating that G decays "quickly" to zero.[4]. The second ensures that $p_{\mathbf{x}}$ is a legitimate density, integrating to 1. The third merely enforces the consistency constraint that $\boldsymbol{\Sigma} = \int d\mathbf{x}\ g(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})\mathbf{x}\mathbf{x}^T$. The Gaussian density function is defined by

$$N(\mathbf{x};\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}e^{-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix for $\mathbf{x}$ and the mean is zero. A Gaussian distribution with mean $\boldsymbol{\mu}$ has a density function given by $N(\mathbf{x}-\boldsymbol{\mu};\boldsymbol{\Sigma})$. It is easily verified that every Gaussian density is a Mahalanobis density with Mahalanobis distribution function $G(x) = -2e^{-\frac{1}{2}x}/(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}$. Another important special case which is easily verified to be Mahalanobis is the uniform density on an ellipsoid (eg the unit spheroid). Note that for the uniform density on the unit spheroid, the input variables are not independent. We show that that the linear fit produces the correct monotonicity directions of $f(\mathbf{x})$ whenever the input density is a Mahalanobis density (see Theorem 2.6).

Since the Gaussian density is a Mahalanobis density, Theorem 2.6 applies, and an immediate corollary is that the linear fit will induce the correct monotonocity directions, provided a certain

[4] This is not a serious constraint if moments of $p_{\mathbf{x}}(\mathbf{x})$ are to exist. In fact, since $p_{\mathbf{x}}(\mathbf{x})$ integrates to 1, this constraint becomes vacuous when $d \geq 3$.

technical condition regarding the growth of f is met. The technical condition essentially amounts to the fact that $\log |f(\mathbf{x})| = o(\mathbf{x}^T\mathbf{x})$, which is a reasonable assumption if the moments of f are to exist. Other Mahalanobis densities are given in Appendix A. Stein's Lemma [19] states that for a univariate normal random variable $X \sim N(\mu, \sigma^2)$ and a differentiable function f, $cov(X, f(X)) = \sigma^2 E[f'(X)]$, which he also generalizes to multi-dimensional normal random variables with identity covariance. Theorem 2.6 can be viewed as a further generalization of Stein's result.

In a practical setting, one has access to a data set sampled according to the regression models discussed in Section 1.1, not to the input distribution $p_{\mathbf{x}}$ or target function f. We show that in the limit of many data points, the monotonicity directions given by the OLS estimate of the linear fit converges to the monotonicity directions given by the true linear fit (see Theorem 3.3). Thus, whenever the true linear fit will give the correct monotonicity directions (for example independent input dimensions or Mahalonobis densities), the OLS estimator will also do so in the limit of many data points. We then give a brief discussion of the sample complexity required to obtain an accurate estimate of the true monotonicity directions. We show in Theorem 3.11 that the probability of making a mistake in monotonicity directions using OLS is $O(de^{-\frac{1}{2}\kappa^2 N}/\kappa\sqrt{N})$, where κ is a constant depending on the noise in the data, and the "degree of monotonicity" of the target function f. When the noise is large or the true linear fit has has small magnitude components, more samples are required, as could be expected.

1.3 Paper Outline

The remainder of the paper is arranged as follows. Next, in Section 2, we prove our main result which gives conditions under which the linear fit constructs the correct monotonicity directions. We then extend the analysis to the OLS estimator and consider the sample complexity in Section 3. We conclude in Section 4 where we also discuss some open questions.

2 The Linear Fit Produces the Correct Monotonicity Directions

We now present our main result which gives conditions under which the linear fit gives the correct monotonicity directions. In particular, when the input variables are independent or when the input density Mahalanobis, the linear fit works.

2.1 Independent Input Variables

We will establish the following theorem which states that when the input variables are independent, and hence the input density factors into a product of marginals, the linear fit extracts the correct monotonicity directions.

Theorem 2.1 (Independent Input Variables.) *Let $f(\mathbf{x})$ be monotonic with monotonicity direction $\mathbf{m}$, and let the input variables be independent* [5]. *Let $\mathbf{w}^l, w_0^l$ be given by the linear fit. Then $m_i = \text{sign}(w_i^l)$ for all i such that $w_i^l \neq 0$. Further, if $f_i(x_i)$ is non-constant for all $\mathbf{x}$ in a compact set of positive probability, then $w_i^l \neq 0$.*

Before we prove Theorem 2.1, we will need some preliminary results which we establish in a sequence of lemmas. The first two state that the monotonicity directions of f and the monotonicity directions that would be induced by a linear fit are unchanged under scaling and translation of the input space. The third states a useful property of monotonic functions.

Lemma 2.2 (Monotonicity direction is scale and translation invariant) *Let $f(\mathbf{x})$ be monotonic with direction $\mathbf{m}$. Let $\mathbf{A}$ be any invertible diagonal matrix and $\mathbf{b}$ be any vector. Then, $g(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$ is monotonic. Further, the monotonicity direction of g is $\text{sign}(\mathbf{A})\mathbf{m}$.*

Proof: Suppose $m_i = +1$ and let $\Delta > 0$. $g_i(x_i + \Delta) = f_i(A_{ii}x_i + b_i + A_{ii}\Delta)$. If $A_{ii} > 0$, then $f_i(A_{ii}x_i + b_i + A_{ii}\Delta) \geq f_i(A_{ii}x_i + b_i) = g_i(x_i)$. Similarily if $A_{ii} < 0$, then $f_i(A_{ii}x_i + b_i + A_{ii}\Delta) \leq f_i(A_{ii}x_i + b_i) = g_i(x_i)$. An analogous argument with $m_i = -1$ and $\Delta < 0$ completes the proof. ■

Lemma 2.3 *Let $\mathbf{w}, w_0$ be the linear fit for $f(\mathbf{x})$ with respect to input density $p_{\mathbf{x}}(\mathbf{x})$. Let $\mathbf{A}$ be any invertible diagonal matrix and $\mathbf{b}$ be any vector. Let $\mathbf{x}' = \mathbf{A}\mathbf{x} + \mathbf{b}$ be a scaled and translated coordinate system, with respect to $\mathbf{x}$. In the $\mathbf{x}'$ coordinate system, let $\mathbf{v}, v_0$ be the linear fit. Then $\mathbf{w} = \mathbf{A}\mathbf{v}$.*

Proof: $\mathbf{w}, w_0$ are minimizers of $\int d\mathbf{x}\ p_{\mathbf{x}}(\mathbf{x})\ (\mathbf{w}^T\mathbf{x} + w_0 - f(\mathbf{x}))^2$, and $\mathbf{v}, v_0$ are minimizers of $\int d\mathbf{x}'\ p_{\mathbf{x}'}(\mathbf{x}')(\mathbf{v}^T\mathbf{x}' + v_0 - f(\mathbf{A}^{-1}(\mathbf{x}' - \mathbf{b})))^2$ where $p_{\mathbf{x}'}(\mathbf{x}') = p_{\mathbf{x}}(\mathbf{A}^{-1}(\mathbf{x}' - \mathbf{b}))/|\mathbf{A}|$. Making a change of variables to $\mathbf{x} = \mathbf{A}^{-1}(\mathbf{x}' - \mathbf{b})$ we have that $\mathbf{v}, v_0$ are minimizers of $\int d\mathbf{x}\ p_{\mathbf{x}}(\mathbf{x})\ (\mathbf{v}^T\mathbf{A}\mathbf{x} + \mathbf{v}^T\mathbf{b} + v_0 - f(\mathbf{x}))^2$. Consequently, we identify $\mathbf{w}^T = \mathbf{v}^T\mathbf{A}$, and since $\mathbf{A}$ is diagonal, the lemma follows. ■

[5] i.e., $p_{\mathbf{x}}(\mathbf{x}) = p_1(x_1)p_2(x_2)\cdots p_d(x_d)$.

Lemma 2.4 *Let $f(\mathbf{x})$ be monotonic with direction $\mathbf{m}$. Then,*

$$m_i x_i f_i(x_i) \geq m_i x_i f_i(0).$$

Further, if $f_i(x_i)$ is non-constant, then $\exists x_i^- < 0$ such that the inequality is strict $\forall x_i \leq x_i^-$, or $\exists x_i^+ \geq 0$ such that the inequality is strict $\forall x_i \geq x_i^+$.

Proof: Let $m_i = +1$. If $x_i \geq 0$, then $f_i(x_i) \geq f_i(0)$ therefore $x_i f_i(x_i) \geq x_i f_i(0)$. If $x_i < 0$, then $f_i(x_i) \leq f_i(0)$ therefore $x_i f_i(x_i) \geq x_i f_i(0)$. An exactly analogous argument holds with inequalities reversed when $m_i = -1$. Further, suppose that $f_i(x_i)$ is non-constant, and that $m_i = 1$. Then one of the following two cases must hold.

(i) $\exists x_i^+ > x \geq 0$ such that $f_i(x_i^+) > f_i(x) \geq f_i(0)$.

(ii) $\exists x_i^- < x \leq 0$ such that $f_i(x_i^+) < f_i(x) \leq f_i(0)$.

In both cases, it is easy to see that the inequality becomes strict in the respective ranges for x_i as claimed. An analogous argument with $m_i = -1$ and the inequality signs reversed completes the proof of the lemma. ■

We now give the proof of our main result.

Proof: **(Theorem 2.1.)** By Lemmas 2.2 and 2.3, after suitable scaling and translation, we can assume, without loss of generality, that $E[\mathbf{x}] = \mathbf{0}$ and that $E[\mathbf{x}\mathbf{x}^T] = \mathbf{I}$ (note that we are excluding the case of degenerate distributions which have zero variance in any given dimension). Then, using Lemma 1.2, we have that

$$\mathbf{w} = \int_{-\infty}^{\infty} d\mathbf{x}\, p_{\mathbf{x}}(\mathbf{x})\mathbf{x} f(\mathbf{x}) \qquad\qquad w_0 = \int_{-\infty}^{\infty} d\mathbf{x}\, p_{\mathbf{x}}(\mathbf{x})\ f(\mathbf{x}). \tag{6}$$

It remains to show that $m_i w_i \geq 0$, as follows.

$$\begin{aligned} m_i w_i &\overset{(a)}{=} \int d\mathbf{x}'_{-i} \int_{-\infty}^{\infty} dx_i\ p_{\mathbf{x}}(\mathbf{x})\ m_i x_i f_i(x_i; \mathbf{x}'_{-i}), \\ &\overset{(b)}{\geq} \int d\mathbf{x}'_{-i}\ p_{\mathbf{x}'_{-i}}(\mathbf{x}'_{-i}) \int_{-\infty}^{\infty} dx_i\ p_{x_i}(x_i) m_i x_i f_i(0; \mathbf{x}'_{-i}), \\ &= \int d\mathbf{x}'\ p_{\mathbf{x}'}(\mathbf{x}') m_i f_i(0; \mathbf{x}'_{-i}) \int_{-\infty}^{\infty} dx_i\ p_{x_i}(x_i) x_i, \\ &\overset{(c)}{=} 0, \end{aligned}$$

where $\mathbf{x}' = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d)$. (a) follows since the measure is independent, (b) by Lemma 2.4 and (c) because $E[\mathbf{x}] = \mathbf{0}$, concluding the proof. Note that if $w_i = 0$, the result is ambiguous. However, from Lemma 2.4 we see that if f_i is non-constant for all $\mathbf{x}$ in a compact set of positive probability, then the $x_i^{\pm}$ can be chosen so as to specify sets of positive probability with the inequality being strict, and hence the result is that $m_i w_i > 0$, concluding the proof of the theorem. ■

Note that the ambiguous case is when $w_i^l = 0$. This may occur only if f is strictly increasing on a set of probability 0 and constant elsewhere. Thus, such an outcome tends to be an artefact of the measure – i.e. in the input region of positive probability, f is actually a constant, hence from point of view of this learning problem, the function is not strictly increasing.

2.1.1 A Counterexample for General Input Densities

In general, one cannot expect the linear fit to extract the correct monotonicity directions. The following proposition establishes this fact by constructing an explicit example. The essential idea behind the counter example is to choose a function like $f(\mathbf{x}) = x_1^3 - x_2$ which is increasing in one dimension and decreasing in the other. By suitably choosing the correlation between x_1 and x_2, the linear regression can be "tricked" into believing that the function is increasing in the x_2 dimension, because the x_1 behavior of the function dominates. The details are given in the proof.

Proposition 2.5 *There exist monotonic functions f and input densities $p_{\mathbf{x}}(\mathbf{x})$ for which the optimal linear fit induces the incorrect monotonicity directions.*

Proof: It suffices to construct an example where the optimal linear fit gives the wrong monotonicity directions. We use a two dimensional example $f(\mathbf{x}) = x_1^3 - x_2$, and for the input density, we use a mixture of Gaussians,

$$p_{\mathbf{x}}(x_1, x_2) = \frac{1}{2}N(x_1 - a_1)N(x_2 - 1) + \frac{1}{2}N(x_1 + a_1)N(x_2 + 1).$$

where $N(x)$ is the standard Gaussian density function, $N(x) = e^{-\frac{1}{2}x^2}/\sqrt{2\pi}$. Notice that $E[\mathbf{x}] = \mathbf{0}$. Denote the covariance matrix of this distribution by $\mathbf{\Sigma}$. Using the moments of the Gaussian distribution, see for example [6], we find that

$$\mathbf{\Sigma} = \begin{bmatrix} 1+a_1^2 & a_1 \\ a_1 & 2 \end{bmatrix}, \qquad E[x_1^4] = a_1^4 + 6a_1^2 + 3, \qquad E[x_1^3 x_2] = a_1^3 + 3a_1.$$

The optimal linear fit is given by

$$\begin{aligned} \mathbf{w} &= \mathbf{\Sigma}^{-1} E[f(\mathbf{x})\mathbf{x}] = \mathbf{\Sigma}^{-1} \begin{bmatrix} E[x_1^4] - E[x_1 x_2] \\ E[x_1^3 x_2] - E[x_2^2] \end{bmatrix}, \\ &= \frac{1}{2+a_1^2} \begin{bmatrix} a_1^4 + 9a_1^2 + 6 \\ -2a_1^3 - a_1^2 - 2 \end{bmatrix}. \end{aligned}$$

The monotonicity direction of f is $\mathbf{m} = [1, -1]$. The first component is always positive, which is consistent with $\mathbf{m}$, however for sufficiently negative a_1, for example $a_1 < -2$, the second component becomes positive which is inconsistent with $\mathbf{m}$, thus concluding the proof. ■

2.2 Mahalanobis Densities

We now obtain the analogous result of Section 2.1 for Mahalanobis input densities. Remember that a density $p_{\mathbf{x}}(\mathbf{x})$ is a Mahalanobis density if it can be written as $p_{\mathbf{x}}(\mathbf{x}) = g\left((\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)$ for some function g defined on $\mathbb{R}_+$. Let $p_{\mathbf{x}}(\mathbf{x})$ be a density that depends on $\mathbf{x}$ only through $\mathbf{x}^T \Sigma^{-1} \mathbf{x}$, where $\mathbf{\Sigma}$ is the covariance matrix for $\mathbf{x}$ under density $p_{\mathbf{x}}$. Thus, $p_{\mathbf{x}}(\mathbf{x}) = g(\mathbf{x}^T \mathbf{\Sigma}^{-1} \mathbf{x})$ is a Mahalanobis density with zero mean. We begin by discussing some properties of $p_{\mathbf{x}}$ before stating the main theorem of this section.

By construction, $E[\mathbf{x}] = \mathbf{0}$, since $p_{\mathbf{x}}$ is a symmetric function. Let G be the indefinite integral of g, so $G'(x) = g(x)$. Assume that $G(x) \leq 0$, $\forall x \geq 0$, and that G is a sufficiently decreasing function such that

$$\lim_{|x|\to\infty} G(x^2)x = 0 \tag{7}$$

Note that g must satisfy some constraints. It must normalize to 1, and the covariance must be $\mathbf{\Sigma}$.

Thus,

$$\begin{aligned}
\mathbf{\Sigma} &= \int d\mathbf{x}\ g(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\mathbf{x}\mathbf{x}^T \\
&= \mathbf{\Sigma}\int d\mathbf{x}\ g(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\mathbf{\Sigma}^{-1}\mathbf{x}\mathbf{x}^T \\
&= \frac{1}{2}\mathbf{\Sigma}\int d\mathbf{x}\ \left[\nabla_{\mathbf{x}}G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\right]\mathbf{x}^T \\
&= \frac{1}{2}\mathbf{\Sigma}\int d\mathbf{x}\ \nabla_{\mathbf{x}}\big(G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\mathbf{x}^T\big) - G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\nabla_{\mathbf{x}}\mathbf{x}^T \\
&= -\frac{1}{2}\mathbf{\Sigma}\int d\mathbf{x}\ G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})
\end{aligned}$$

where the last line follows because, using the fundamental theorem of calculus and (7), the first term is zero, and, $\nabla_{\mathbf{x}}\mathbf{x}^T = \mathbf{I}$. Thus, we have the two constraints,

$$\int d\mathbf{x}\ G'(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}) = 1 \qquad \int d\mathbf{x}\ G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}) = -2. \tag{8}$$

The first constraint can always be effected by multiplying G by some positive scalar. The second then leads to a constraint on G. The solid angle in d-dimensions is given by $\Omega_d = 2\pi^{d/2}/\Gamma(d/2)$. Using a transformation into polar coordinates, $\int d\mathbf{x}f(\mathbf{x}^T\mathbf{x}) = \Omega_d\int_0^\infty ds\ s^{d-1}f(s^2)$. Thus, using a coordinate transformation to $\mathbf{z} = \mathbf{\Sigma}^{-1/2}\mathbf{x}$, these two constraints can be reduced to

$$\int_0^\infty ds\ s^{d-1}G'(s^2) = \frac{\Gamma(d/2)}{2|\mathbf{\Sigma}|^{1/2}\pi^{d/2}} \qquad \int_0^\infty ds\ s^{d-1}G(s^2) = -\frac{\Gamma(d/2)}{|\mathbf{\Sigma}|^{1/2}\pi^{d/2}}, \tag{9}$$

where the Gamma function is defined by $\Gamma(x) = \int_0^\infty ds\ s^{x-1}e^{-s}$. In terms of $g(x)$, these constraints become

$$\int_0^\infty ds\ s^{d-1}g(s^2) = \frac{\Gamma(d/2)}{2|\mathbf{\Sigma}|^{1/2}\pi^{d/2}} \qquad \int_0^\infty ds\ s^{d+1}g(s^2) = \frac{\Gamma(d/2+1)}{|\mathbf{\Sigma}|^{1/2}\pi^{d/2}}. \tag{10}$$

The classification boundary with respect to dimension x_i is a function $f_i^c(\mathbf{x}') : \mathbb{R}^{d-1} \mapsto \mathbb{R}\cup\{\infty, -\infty\}$, that determines the point at which $f_i(x_i)$ changes sign. Here $\mathbf{x}' = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$. Thus,

$$f_i(x_i) = \begin{cases} m_i & x_i \geq f_i^c(x') \\ -m_i & x_i < f_i^c(x') \end{cases}$$

An interesting fact about the classification boundary is that it is a monotonic function. In fact, its monotonicity directions $\mathbf{m}^c$ can be obtained from the original monotonicity directions by $\mathbf{m}^c = -m_i\mathbf{m}'$. We are now ready to give our main result for Mahalanobis densities.

Theorem 2.6 (Mahalanobis Densities.) *Let $f(\mathbf{x})$ be monotonic with monotonicity direction $\mathbf{m}$, and let the input probability density be a Mahalanobis density. In the regression case, assume that f is differentiable and does not grow too quickly, i.e.,*

$$\lim_{|x_i|\to\infty} G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}^T)f(\mathbf{x}) = 0 \qquad \forall i = 1,\ldots,d. \tag{11}$$

Let $\mathbf{w}^l$ be given by the linear fit. Then $m_i = \operatorname{sign}(w_i^l)$ for all i such that $w_i^l \neq 0$. Further, if $f_i(x_i)$ is non-constant for all $\mathbf{x}$ in some compact set of positive measure, then $w_i^l \neq 0$.

Proof: Let $p_{\mathbf{x}}(\mathbf{x}) = g(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}^T)$ satisfy the properties described above. Let f be a monotonic function with monotonicity direction $\mathbf{m}$ satisfying[6]

$$\lim_{|x_i|\to\infty} G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}^T)f(\mathbf{x}) = 0 \qquad \forall i = 1,\ldots,d. \tag{12}$$

Let's first consider the regression case, then $\mathbf{w}$ from the linear fit is given by

$$\begin{aligned}
\mathbf{w} &= \int d\mathbf{x}\; g(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\mathbf{\Sigma}^{-1}\mathbf{x}f(\mathbf{x}),\\
&\overset{(a)}{=} \frac{1}{2}\int d\mathbf{x}\;\left[\nabla_{\mathbf{x}}G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\right]f(\mathbf{x}),\\
&= \frac{1}{2}\int d\mathbf{x}\;\nabla_{\mathbf{x}}\big(G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})f(\mathbf{x})\big) - G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\nabla_{\mathbf{x}}f(\mathbf{x}),\\
&\overset{(b)}{=} -\frac{1}{2}\int d\mathbf{x}\;G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\nabla_{\mathbf{x}}f(\mathbf{x}),\\
&\overset{(c)}{=} -\frac{1}{2}\left(\int d\mathbf{x}\;G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\mathbf{\Lambda}(\mathbf{x})\right)\mathbf{m},\\
&\overset{(d)}{=} \mathbf{\Lambda}\mathbf{m},
\end{aligned} \tag{13}$$

where $\mathbf{\Lambda}(\mathbf{x})$ and $\mathbf{\Lambda}$ are a non-negative diagonal matrices. (a) follows by the definition of G; (b) follows by using the fundamental theorem of calculus and (12); (c) follows because f is monotonic with direction $\mathbf{m}$, therefore $\nabla_{\mathbf{x}}f(\mathbf{x})$ must have the same sign as $\mathbf{m}$ and hence can be written as $\mathbf{\Lambda}(\mathbf{x})\mathbf{m}$; (d) follows because $-G$ is non-negative. Thus all the non-zero components of $\mathbf{w}$ have the same sign

[6] Note that for the classification case this restriction is vacuous as $|f(\mathbf{x})| = 1$.

as $\mathbf{m}$ and the theorem follows. Note that if for each i and some $\epsilon > 0$, $|G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\mathbf{\Lambda}_{ii}(\mathbf{x})| \geq \epsilon$ holds in some set of measure greater than zero, then every component of $\mathbf{w}$ will be non-zero. Certainly this will be the case if $f_i(x_i)$ is non-constant for all $\mathbf{x}$ in a compact set of positive probability.

For the classification case, (13) gives

$$\begin{aligned} w_i &\overset{(a)}{=} -\frac{m_i}{2}\int d\mathbf{x}' \left[\int_{-\infty}^{f_i^c(\mathbf{x}')} \frac{\partial}{\partial x_i} G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}) - \int_{f_i^c(\mathbf{x}')}^{\infty} \frac{\partial}{\partial x_i} G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\right], \\ &\overset{(b)}{=} -m_i \int d\mathbf{x}'\; G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\big|_{x_i=f_i^c(\mathbf{x}')} \\ &\overset{(c)}{=} \lambda_i m_i \end{aligned}$$

where $\lambda_i \geq 0$. (a) follows by definition of $f_i^c(\mathbf{x}')$; (b) follows by the fundamental theorem of calculus; and (c) follows because $G(x) < 0$. If $f_i^c(\mathbf{x}')$ is bounded on a compact positive probability set, which will happen if $f_i(x_i)$ is non-constant for all $\mathbf{x}$ in a compact set of positive probability, then $\lambda_i > 0$, and the theorem follows. ■

3 The OLS Linear Fit on a Finite Sample

Since the linear fit is not accesible in practice, we need to consider the effects that the variability of a finite sample drawn from the input distribution has on the quality of the estimates of the monotonicity directions. We will consider the regression models described in Section 1.1 which draw a finite *iid* sample from the input distribution $p_{\mathbf{x}}$ and add *iid* noise to the function value. Let the data be $\{\mathbf{x}_i, y_i\}_{i=1}^N$ and let $\mathbf{X}_N = \frac{1}{N}\sum_{i=1}^N \hat{\mathbf{x}}_i\hat{\mathbf{x}}_i^T$, and $y_i = f(\mathbf{x}_i) + \epsilon_i$. The noise ϵ_i satisfies (4) for regression, and (5) for classification.

3.1 Convergence of the OLS Estimator

Since sample averages converge to the their expectations, it should not be surprising that the OLS estimators (which are based on the sample averages) should converge to quantities related to the true linear fit. The following lemma is therefore not surprising.

Lemma 3.1 *Let $w_0^l, \mathbf{w}^l$ be the linear fit to $f(\mathbf{x})$ with respect to input density $p_{\mathbf{x}}(\mathbf{x})$. Assume that all fourth order moments of $p_{\mathbf{x}}$ with respect to $\mathbf{x}$, $f(\mathbf{x})$ and ϵ_i are bounded, and that $E[\mathbf{x}] = \mathbf{0}$. Suppose that N points $\{\mathbf{x}_i\}_{i=1}^N$ are sampled* i.i.d. *from $p_{\mathbf{x}}$ with $y_i = f(\mathbf{x}_i) + \epsilon_i$ where ϵ_i is independent noise.*

For regression, the noise satisfies (4), and for classification, the noise is independent flip noise (5). Let $\mathbf{w}^*$ *be the OLS estimator of* $\mathbf{w}^l$*. Then,*

$$\mathbf{w}^* \xrightarrow{P} \begin{cases} \mathbf{w}^l & \textit{regression}, \\ (1-2p)\mathbf{w}^l & \textit{classification}. \end{cases}$$

Notice that while $\mathbf{w}^*$ converges in probability to $\mathbf{w}^l$ for regression, it *does not* for classification, unless $p = 0$. However, since $p < \frac{1}{2}$, the sign of $\mathbf{w}^*$ converges in probability to the sign of $\mathbf{w}^l$ for both cases. Thus, if the linear fit $\mathbf{w}^l$ induces the correct monotonicity directions, then so will $\mathbf{w}^*$, asymptotically as $N \to \infty$. The basic idea behind the proof of Lemma 3.1 is to prove that the expectation of the OLS estimator converges to the result claimed in the Lemma, and the covariance converges to zero.

An immediate implication of Lemma 3.1 is that the OLS estimator and the linear fit are related (in the limit) by a positive scalar, and hence obtain the same monotonicity directions.

Corollary 3.2 $sign(w_i^*) \xrightarrow{P} sign(w_i^l)$ *whenever* $sign(w_i^l) \neq 0$.

Thus, the OLS estimator will obtain the correct monotonicity directions whenever the linear fit does, and the following theorem is therefore evident.

Theorem 3.3 (OLS) *Let* $f(\mathbf{x})$ *be monotonic with direction* $\mathbf{m}$*, and suppose that* N *points* $\{\mathbf{x}_i\}_{i=1}^N$ *are sampled* i.i.d. *from* $p_{\mathbf{x}}(\mathbf{x})$ *with* $y_i = f(\mathbf{x}_i)+\epsilon_i$ *where* ϵ_i *is independent noise. For classification,* ϵ_i *is flip noise with probability* $p < \frac{1}{2}$ *(5), otherwise it is a zero mean random variable with variance* σ^2 *(4). Assume all fourth order moments are finite. Let* $\mathbf{w}^l$ *be given by the exact linear fit, and let* $\mathbf{w}^*$ *be the OLS estimators for* $\mathbf{w}^l$*. Suppose further that the linear fit induces the correct monotonicity directions, i.e.,* $sign(\mathbf{w}^l) = \mathbf{m}$*. Then,*

$$\lim_{N\to\infty} P[sign(\mathbf{w}^*) = \mathbf{m}] = 1.$$

This theorem states that if the linear fit extracts the correct monotonicity directions, then with high probability (for large N), the OLS estimator will do so as well, even in the presence of noise. The theorem thus applies to independent input variables and Mahalanobis densities. The theorem is more general, in that it applies whenever the linear fit works.

To prove Lemma 3.1, we will need some intermediate results.

Lemma 3.4 (Expectation of the OLS estimator) *Let* $\mathbf{w}^*$ *be the OLS estimator and* $\mathbf{w}'$ *be the OLS estimator had the data been noiseless. Then*

$$E_\epsilon[\mathbf{w}^*] = \begin{cases} \mathbf{w}' & \text{regression,} \\ \mathbf{w}'(1-2p) & \text{classification,} \end{cases}$$

where $\mathbf{w}' = \frac{1}{N}\mathbf{X}_N^{-1}\sum_{i=1}^N f(\mathbf{x}_i)\hat{\mathbf{x}}_i$ *and the expectation is with respect to the noise.*

Proof: By Lemma 1.3,

$$\mathbf{w}^* = \mathbf{w}' + \frac{\mathbf{X}_N^{-1}}{N}\sum_{i=1}^N \epsilon_i \hat{\mathbf{x}}_i,$$

because $\mathbf{w}' = \frac{1}{N}\mathbf{X}_N^{-1}\sum_{i=1}^N f(\mathbf{x}_i)\hat{\mathbf{x}}_i$. Taking expectations, for regression noise we have $E[\epsilon_i] = 0$, and for the flip noise we have $E[\epsilon_i] = -2pf(\mathbf{x}_i)$, from which the lemma follows. ■

Note that for regression, the OLS estimator is unbiased, whereas for classification flip noise, it is not unbiased. This is because the classification flip noise is not unbiased noise.

Lemma 3.5 (Covariance of the OLS estimator) *Let* $\mathbf{w}^*$ *be the OLS estimator, then*

$$Cov(\mathbf{w}^*) = \begin{cases} \frac{\sigma^2\mathbf{X}_N^{-1}}{N} & \text{regression,} \\ \frac{4p(1-p)\mathbf{X}_N^{-1}}{N} & \text{classification.} \end{cases}$$

Proof: $Cov(\mathbf{w}^*) = E[(\mathbf{w}^* - \mathbf{E}[\mathbf{w}^*])(\mathbf{w}^* - \mathbf{E}[\mathbf{w}^*])^T]$. For regression,

$$\begin{aligned} Cov(\mathbf{w}^*) &= \frac{\mathbf{X}_N^{-1}}{N}\left(\sum_{i=1}^N\sum_{j=1}^N \hat{\mathbf{x}}_i\hat{\mathbf{x}}_j^T\mathbf{E}[\epsilon_i\epsilon_j]\right)\frac{\mathbf{X}_N^{-1}}{N}, \\ &\overset{(a)}{=} \frac{\sigma^2\mathbf{X}_N^{-1}}{N}, \end{aligned}$$

where (a) follows because $\mathbf{E}\left[\epsilon_i\epsilon_j\right] = \sigma^2\delta_{ij}$. For classification,

$$\begin{aligned} Cov(\mathbf{w}^*) &= \frac{\mathbf{X}_N^{-1}}{N}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{\mathbf{x}}_i\hat{\mathbf{x}}_j^T\mathbf{E}\left[(2pf(\mathbf{x}_i)+\epsilon_i)(2pf(\mathbf{x}_j)+\epsilon_j)\right]\right)\frac{\mathbf{X}_N^{-1}}{N}, \\ &\overset{(b)}{=} \frac{4p(1-p)\mathbf{X}_N^{-1}}{N}, \end{aligned}$$

where (b) follows because $f(\mathbf{x}_i)^2 = 1$ and so using (5) and the independence of the ϵ_i, we get that $\mathbf{E}\left[(2pf(\mathbf{x}_i)+\epsilon_i)(2pf(\mathbf{x}_j)+\epsilon_j)\right] = 4p(1-p)\delta_{ij}$, from which the lemma follows. ■

The following is a well known lemma about the distribution of the OLS estimator, essentially stating that it has an asymptotically Gaussian distribution.

Lemma 3.6 *The OLS estimator has a distribution that is asymptotically Gaussian, given by*

$$\boldsymbol{\beta}^* \xrightarrow{P} N(\bar{\boldsymbol{\beta}}; Q)$$

where $\bar{\boldsymbol{\beta}}$ is the mean of the estimator, given in Lemma 3.4 and the covariance matrix Q is given by Lemma 3.5. Therefore, $\boldsymbol{\beta}^* \xrightarrow{P} \bar{\boldsymbol{\beta}}$.

Proof: The fact that $\boldsymbol{\beta}^* \xrightarrow{P} N(\bar{\boldsymbol{\beta}}, Q)$ is a standard result, see for example [6]. By Lemmas 3.5, 1.5, we have that $Q \xrightarrow{P} \mathbf{0}$, and so $\boldsymbol{\beta}^* \xrightarrow{P} N(\bar{\boldsymbol{\beta}}, \mathbf{0})$, implying that $\boldsymbol{\beta}^* \xrightarrow{P} \bar{\boldsymbol{\beta}}$. ■

Proof: **(Lemma 3.1.)** Let $\hat{\boldsymbol{\Sigma}} = E[\hat{\mathbf{x}}\hat{\mathbf{x}}^T]$. By Lemma 1.5, $X_N^{-1} \xrightarrow{P} \hat{\boldsymbol{\Sigma}}^{-1}$. By the weak law of large numbers, $\frac{1}{N}\sum_i f(x_i)\hat{\mathbf{x}}_i \xrightarrow{P} E[f(\mathbf{x})\hat{\mathbf{x}}]$, so $\mathbf{w}' \xrightarrow{P} \boldsymbol{\Sigma}^{-1}E[f(\mathbf{x})\hat{\mathbf{x}}] = \mathbf{w}^l$. By Lemma 3.6, $\mathbf{w}^* \xrightarrow{P} \mathbf{w}'$ for regression, and $\mathbf{w}^* \xrightarrow{P} (1-2p)\mathbf{w}'$ for classification. Since $\mathbf{w}' \xrightarrow{P} \mathbf{w}^l$, we therefore conclude that $\mathbf{w}^* \xrightarrow{P} \mathbf{w}^l$ for regression and $\mathbf{w}^* \xrightarrow{P} (1-2p)\mathbf{w}^l$ for classification. ■

3.2 Sample Complexity

In the previous section, we established that the OLS estimator converges to a scalar multiple of the linear fit, and hence extracts the same monotonicity directions as the linear fit, in the limit as $N \to \infty$. In any practical setting, one has a finite number of samples. The natural question is how large a sample size is needed to ensure that the monotonicity directions are correctly predicted with high enough probability – i.e., how quickly (w.r.t. N) does the OLS estimator give the correct monotonicity directions.. These questions can be answered by appealing to Lemma 3.6. More precisely, write

$$\mathbf{X}_N = \hat{\mathbf{\Sigma}} - \frac{\mathbf{A}_N}{\sqrt{N}}, \qquad \mathbf{q}_N = \frac{1}{N}\sum_i f(\mathbf{x}_i)\hat{\mathbf{x}}_i = \mathbf{q} + \frac{\mathbf{b}_N}{\sqrt{N}}, \tag{14}$$

where $\mathbf{q} = E[f(\mathbf{x})\hat{\mathbf{x}}]$. We will need some results regarding the moments of $\mathbf{A}_N$ and $\mathbf{b}_N$, as well as some cross moments.

Lemma 3.7

(i) $E[\mathbf{A}_N] = E[\mathbf{b}_N] = \mathbf{0}$.

(ii) All second order moments involving $\mathbf{A}_N, \mathbf{b}_N$ *are* $O(1)$*. By all second order moments, we mean moments of the form* $E[\mathbf{A}_N^2]$, $E[\mathbf{b}_N\mathbf{b}_N^T]$, $\mathbf{A}_N\hat{\mathbf{\Sigma}}^{-1}\mathbf{b}_N$*, etc.*

(iii) $E[\mathbf{X}_N^{-1}] = \hat{\mathbf{\Sigma}}^{-1} + \frac{1}{N}\hat{\mathbf{\Sigma}}^{-1}E[\mathbf{A}_N\hat{\mathbf{\Sigma}}^{-1}\mathbf{A}_N]\hat{\mathbf{\Sigma}}^{-1} + O(\frac{1}{N^{3/2}})$ *(we assume that any necessary expectations exist).*

Proof:

(i) By taking expectations on both sides of the definitions in (14), we obtain $E[\mathbf{A}_N] = E[\mathbf{b}_N] = \mathbf{0}$.

(ii) The basic idea behind this result is that we can write $\mathbf{X}_N - \hat{\mathbf{\Sigma}} = -\frac{\mathbf{A}_N}{\sqrt{N}}$ and $\mathbf{b}_N - \mathbf{q} = -\frac{\mathbf{b}_N}{\sqrt{N}}$. Thus the second order moments of terms involving $\mathbf{A}_N$ and $\mathbf{b}_N$ correspond to second order central moments of $\mathbf{X}_N$ and $\mathbf{b}_N$. Let $\mathcal{L}$ be a second order central moment involving $\mathbf{X}_N, \mathbf{b}_N$. Then $\mathcal{L} = \mathcal{M}/N$, where $\mathcal{M}$ is a corresponding second moment involving $\mathbf{A}_N, \mathbf{b}_N$. Since $\mathbf{X}_N$ and $\mathbf{q}_N$ are sums of independent random matrices and vectors respectively, their second order central moments should decay proportionaly to $\frac{1}{N}$. Thus, $\mathcal{L} = O(\frac{1}{N})$, and hence $\mathcal{M} = O(1)$. We give the details for one specific case to illustrate the mechanics of the proof. The remaining cases are similar and we omit the details. Consider $E[\mathbf{A}_N\hat{\mathbf{\Sigma}}^{-1}\mathbf{A}_N]$. We have

$$\begin{aligned}
\frac{E[\mathbf{A}_N\hat{\mathbf{\Sigma}}^{-1}\mathbf{A}_N]}{N} &= E[(\mathbf{X}_N-\hat{\mathbf{\Sigma}})\hat{\mathbf{\Sigma}}^{-1}(\mathbf{X}_N-\hat{\mathbf{\Sigma}})], \\
&= E[\mathbf{X}_N\hat{\mathbf{\Sigma}}^{-1}\mathbf{X}_N-2\mathbf{X}_N+\hat{\mathbf{\Sigma}}], \\
&= \frac{1}{N^2}E[\sum_{i,j}\mathbf{x}_i\mathbf{x}_i^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{x}_j\mathbf{x}_j^T]-\hat{\mathbf{\Sigma}}, \\
&\overset{(a)}{=} \frac{1}{N^2}\left(NE[\mathbf{x}\mathbf{x}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{x}\mathbf{x}^T]+N(N-1)\hat{\mathbf{\Sigma}}\right)-\hat{\mathbf{\Sigma}}, \\
&= \frac{E[\mathbf{x}\mathbf{x}^T\hat{\mathbf{\Sigma}}^{-1}\mathbf{x}\mathbf{x}^T]-\hat{\mathbf{\Sigma}}}{N},
\end{aligned}$$

where the equality (a) follows by considering separately the cases $i=j$ and $i\neq j$. The result now follows because we have assumed that all exist (and are therefore $O(1)$).

(iii) Taking the Taylor expansion for $\mathbf{X}_N^{-1}$ followed by its expectation yields the final result.

■

In order to apply Lemma 3.6, we need $E[\mathbf{w}']$. Writing

$$\begin{aligned}
\mathbf{w}' &= \mathbf{X}_N^{-1}\mathbf{q}_N, \\
&= \left(\hat{\mathbf{\Sigma}}-\frac{\mathbf{A}_N}{\sqrt{N}}\right)^{-1}\left(\mathbf{q}+\frac{\mathbf{b}_N}{\sqrt{N}}\right), \qquad (15)
\end{aligned}$$

and taking expectations after expanding to second order, we obtain

Lemma 3.8 $E[\mathbf{w}']=\hat{\mathbf{\Sigma}}^{-1}\mathbf{q}+\frac{1}{N}\hat{\mathbf{\Sigma}}^{-1}E[\mathbf{A}_N\hat{\mathbf{\Sigma}}^{-1}\mathbf{A}_N\hat{\mathbf{\Sigma}}^{-1}\mathbf{q}+\mathbf{A}_N\hat{\mathbf{\Sigma}}^{-1}\mathbf{b}_N]+O(\frac{1}{N^{3/2}})=\mathbf{w}^l+O(\frac{1}{N})$.

The covariance of $\mathbf{w}^*$ is given by $cov(\mathbf{w}')$ plus the expectation with respect to $\mathbf{X}_N$ of the covariance given in Lemma 3.5 (in Lemma 3.5, we have only taken expectations with respect to the noise variables). After expanding (15), we find that

Lemma 3.9 *$cov(\mathbf{w}')=\frac{1}{N}\hat{\mathbf{\Sigma}}^{-1}E[(\mathbf{A}_N\hat{\mathbf{\Sigma}}^{-1}\mathbf{q}+\mathbf{b}_N)(\mathbf{A}_N\hat{\mathbf{\Sigma}}^{-1}\mathbf{q}+\mathbf{b}_N)^T]\hat{\mathbf{\Sigma}}^{-1}+O(\frac{1}{N^{3/2}})=\frac{R_N}{N}+O(\frac{1}{N^{3/2}})$, where $R_N=O(1)$.*

The fact that $R_N=O(1)$ follows from Lemma 3.7(ii). Taking the expectation of the result in Lemma 3.5 with respect to $\mathbf{X}_N$, and combining this result with Lemma 3.8 and Lemma 3.9, and using Lemma 3.6, we obtain

Lemma 3.10 $\mathbf{w}^* \sim N(\boldsymbol{\mu}, \mathbf{Q})$, *where*

$$\mu = \begin{cases} \mathbf{w}^l + O\left(\frac{1}{N}\right) & \textit{regression,} \\ (1-2p)\mathbf{w}^l + O\left(\frac{1}{N}\right) & \textit{classification.} \end{cases} \qquad \mathbf{Q} = \begin{cases} \dfrac{\sigma^2\hat{\boldsymbol{\Sigma}}^{-1} + R_N}{N} + O(\frac{1}{N^{3/2}}) & \text{regression,} \\ \dfrac{4p(1-p)\hat{\boldsymbol{\Sigma}}^{-1} + R_N}{N} + O(\frac{1}{N^{3/2}}) & \text{classification.} \end{cases}$$

Notice that $\boldsymbol{\mu} = \alpha\mathbf{w}^l + O(\frac{1}{N})$ and $N\mathbf{Q} = \gamma\hat{\boldsymbol{\Sigma}}^{-1} + R_N + o(1)$, where α and γ are constants depending on whether we have the regression or classification model,

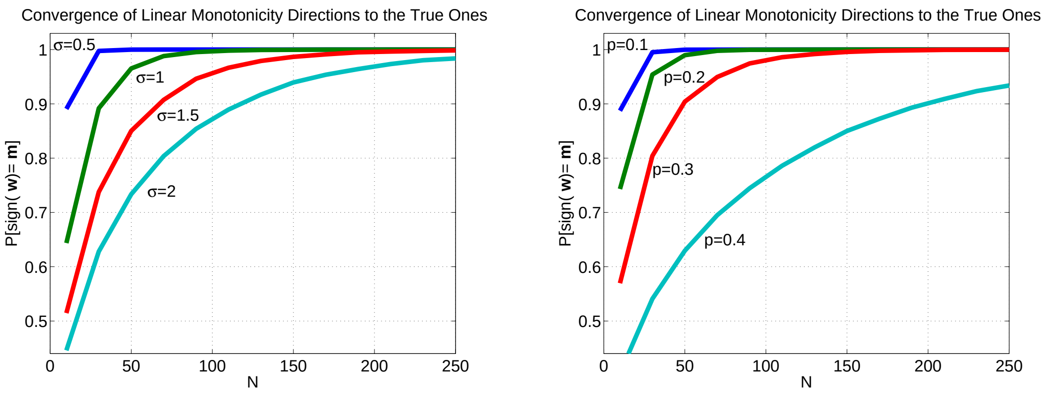
$$\alpha = \begin{cases} 1 & \text{regression,} \\ (1-2p) & \text{classification.} \end{cases} \qquad \gamma = \begin{cases} \sigma^2 & \text{regression,} \\ 4p(1-p) & \text{classification.} \end{cases}$$

Assume $\mathbf{w}^l > 0$ (no loss of generality) and let $q_i^2 = (NQ)_{ii} = \gamma\hat{\boldsymbol{\Sigma}}_{ii}^{-1} + (R_N)_{ii} + o(1)$. Then an application of Chebyshev's inequality gives that $P[w_i^* > 0] \geq 1 - \frac{q_i^2}{N\mu_i^2} = 1 - \frac{\gamma\hat{\boldsymbol{\Sigma}}_{ii}^{-1} + (R_N)_{ii}}{N\alpha^2 {w_i^l}^2} + o(\frac{1}{N})$. The numerator in the second term represents the randomness (both in the noise and the sampling of the $\mathbf{x}_i$'s in the data set). The denominator reflects the tradeoff between N and w_i^l, specifically keeping all else constant, $N \propto 1/{w_i^l}^2$ for constant error probability in the i^{th} monotonicity direction. Taking a union bound, we see that $P[\mathbf{w}^* > 0] \geq 1 - \frac{1}{N\alpha^2}\sum_{i=1}^{d} \frac{\gamma\hat{\boldsymbol{\Sigma}}_{ii}^{-1} + (R_N)_{ii}}{{w_i^l}^2} + o(\frac{1}{N})$. Actually, the convergence is much faster than the Chebyshev bound would suggest. Using the fact that the marginals of a Normal distribution are Normal, we have:

$$\begin{aligned} P[w_i^* > 0] &= \int_{w>0} dw \, \frac{1}{\sqrt{2\pi Q_{ii}^2}} e^{-\frac{(w-\mu_i)(w-\mu_i)}{2Q_{ii}^2}}, \\ &= \int_{z>-\sqrt{N}\mu_i} dz \, \frac{1}{\sqrt{2\pi q_i^2}} e^{-\frac{w^2}{2q_i^2}}. \end{aligned}$$

The asymptotic expansion of the integral above is well known (see for example [8]). After some relatively straightforward manipulation, and applying the union bound, we arrive at the following result.

$$P[\mathbf{w}^* > 0] \geq 1 - \frac{1}{\pi\sqrt{2N}} \sum_{i=1}^{d} \frac{\sqrt{\gamma\Sigma_{ii}^{-1} + (R_N)_{ii}}}{\alpha w_i^l} \exp\left(-\frac{N\alpha^2 {w_i^l}^2}{2(\gamma\Sigma_{ii}^{-1} + (R_N)_{ii})}\right) \cdot (1 + o(1)).$$



Regression

Classification

Figure 1: Probability of obtaining the correct monotonocity directions vs. sample size.

Defining $\kappa^2 = \alpha^2 \min_i \left\{ \frac{{w_i^1}^2}{\gamma \Sigma_{ii}^{-1} + (R_N)_{ii}} \right\}$, we have

Theorem 3.11 $P[\mathbf{w}^* > 0] \geq 1 - \dfrac{d}{\kappa\pi\sqrt{2N}} e^{-\frac{N\kappa^2}{2}} \cdot (1 + o(1))$.

As can be noted, the convergence is exponential. Asymptotically, if we ignore the $o(1)$ term, then after some elementary manipulation, we obtain a result on the sample complexity.

Corollary 3.12 *Given* $\eta > 0$, *if* $N \geq \frac{1}{\kappa^2} \log \left(\frac{d^2}{2\pi^2 \kappa^2 \eta^2} \right)$, *then* $P[\mathbf{w}^* = \mathbf{m}] \geq 1 - \eta$.

Corollary 3.12 quantitatively captures the asymptotic behavior of the sample complexity on the dimension d, the error tolerence η, and κ which represents the regression setting (input distribution, noise in the data and target function). Qualitatively, κ is increasing in the magnitude of the true linear fit (the degree of monotonicity in f), increasing in the input variance and decreasing in the noise level; the larger κ is, the smaller the sample complexity. Note that the sample complexity is only logarithmic in $\frac{d}{\eta}$, and thus the curse of dimensionality does not apply to estimating the monotonicity directions in this way.

3.3 Experimental Evaluation

We performed an experimental evaluation of the theoretical results. We generated data uniformly from $[0,1]^2$ and used the target function $f(x,y) = e^{y-x} - 1$ for regression and the sign of this function for classification. The convergence of $sign(\mathbf{w}^*)$ to the true monotonicity directions $\mathbf{m} = [-1, 1]$ is

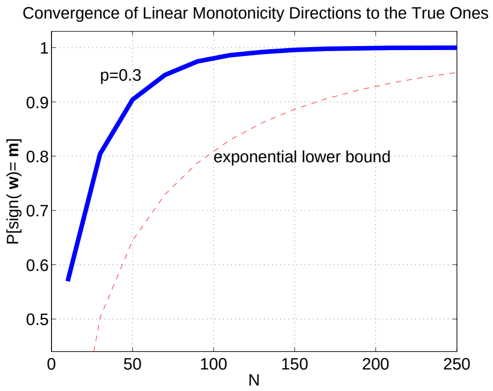


Figure 2: Comparison of the theoretical lower bound from Theorem 3.11 with the probability of error.

illustrated in Figure 1, where we show the dependence of $P[\text{sign}(\mathbf{w}) = \mathbf{m}]$ as a function of N for different noise levels.

For this particular test case, it is possible to analytically compute the parameters in the theoretical bounds of the previous section. In particular,

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{12} & 0 \\ 0 & 0 & \frac{1}{12} \end{bmatrix}, \quad \mathbf{q} = \begin{bmatrix} 0 \\ -\frac{1}{6} \\ \frac{1}{6} \end{bmatrix}, \quad \mathbf{w}^l = \begin{bmatrix} 0 \\ -2 \\ 2 \end{bmatrix}$$

and

$$\hat{\mathbf{A}}_N = -\frac{1}{\sqrt{N}} \begin{bmatrix} 0 & \sum_i x_i & \sum_i y_i \\ \sum_i x_i & \sum_i x_i^2 - \frac{N}{12} & \sum_i x_i y_i \\ \sum_i y_i & \sum_i x_i y_i & \sum_i y_i^2 - \frac{N}{12} \end{bmatrix}, \quad \mathbf{b}_N = \frac{1}{\sqrt{N}} \begin{bmatrix} \sum_i f(\mathbf{x}_i) \\ \sum_i f(\mathbf{x}_i) x_i + \frac{N}{6} \\ \sum_i f(\mathbf{x}_i) y_i - \frac{N}{6} \end{bmatrix}.$$

We can therefore explicitly compute R_N as given in Lemma 3.9 and use Theorem 3.11 to compute the exponential lower bound on the probability of error. The comparison of this lower bound with the experiments are shown in Figure 2 for the case of classification with noise parameter $p = 0.3$.

4 Conclusion

The premise of this work is that incorporating constraints into the learning process improves the generalization of the resulting learned function. We study monotonicity constraints, and before a monotonicity constraint can be enforced, most algorithms will require knowledge of the direction of the monotonicity. We have shown that under quite general assumptions, the correct monotonicity directions are induced by fitting a linear model to the data, in particular, when the inputs have a Gaussian distribution. We have assumed that the function f is monotonic in every dimension. It is possible for f to be monotonic in some dimensions and non-monotonic in others. The proofs do not require monotonicity in every dimension, i.e. it is straightforward to extend the proofs to the situation where only some of the dimensions are monotonic. In this case, the linear model will extract the correct monotonicity directions for those dimensions in which monotonicity is known to hold. Once the direction of monotonicity has been determined, it can be incorporated into more complicated learning models such as neural networks, a task that would have been considerably tougher had the monotonicity directions not been known.

The linear model has a number of appealing features: it is easy to implement; once it has been implemented, the monotonicity directions are easy to determine; the OLS linear fit to a finite data set obtains monotonicity directions which converge exponentially quickly to the monotonicity directions of the exact linear fit. As we see in Corollory 3.12 these convergence rates can be used to determine the sample complexity (how much data is needed to make an accurate determiniation of the monotonicity directions). The sample complexity is decreasing in the magnitude of the linear fit vector $\mathbf{w}^l$: as expected, it is easier to estimate the monotonicity when the "degree of monotonicity" of the true function is larger. The dependence of the sample complexity on the noise, and the input distribution can be deduced from Corollary 3.12. In particular, determining the monotonicity directions using the OLS estimator does not suffer from the curse of dimension. The main drawback of the linear model is that it is useful for certain classes of input densities, in particular when the inputs are independent or distributed according to a Mahalanobis (eg. Gaussian) density. Enlarging this class of densities would be useful progress.

Other approaches to determining the monotonicity of a function that are as simple and efficient as fitting a linear model would also be useful. There is potential that some non-parametric techniques could prove successful in this respect, for example regression approaches that are consistent, in that they approach the true function f in a distribution-independent manner. The main drawbacks of such a general approach are that the convergence will be much slower than for linear models, and the monotonicity directions of the resulting function may not be easy to determine – this function may not even be monotonic. Our motivation is that a simple, effective algorithm be used to obtain the monotonicity directions which can then be used to *constrain* more powerful models so that the more powerful model will attain a better generalization performance.

We end by formulating two open questions. The first relates to how bad can it be to follow the linear models monotonicity directions. Specifically, The example constructed in Proposition 2.5 required one dimension to dominate the other. In such a situation, one might suspect that this second dimension is not important in the implementation of the true monotonic fit. We thus formulate the following question.

Question 1 *Let f be a monotonic function with monotonicity directions $\mathbf{m}$ and suppose the linear fit implies monotonicity directions $\mathbf{m}'$. Consider a learning model $\mathcal{L}$ used to estimate the function f subject to the monotonicity constraints. Specifically, let $g \in \mathcal{L}$ be the best fit to f subject to the monotonicity constraints $\mathbf{m}$ and let $g' \in \mathcal{L}$ be the best fit to f subject to the monotonicity constraints $\mathbf{m}'$. The question is to construct bounds on $\parallel g - g' \parallel$, or show (non-trivial) examples where $\parallel g - g' \parallel$ can be arbitrarily large.*

Perhaps, if f is constrained somehow, for example to have bounded derivatives, then the answer to this question is in the affirmative, i.e., despite the linear fit not giving the correct monotonicity directions, using those monotonicity directions anyway do not lead one too far astray.

The second question is based on the observation that the target function may not be monotonic in any of the variables $\mathbf{x}$, but may be monotonic in some set of features built from these variables (for example linear combinations of the variables). In particular, we formulate the following question.

Question 2 *Do there exist efficient algorithms to construct monotonic features. Specifically suppose that f is a function of $\mathbf{x}$ and f is monotonic in the feature variables $\mathbf{y} = \mathbf{A}(\mathbf{x})$ where $\mathbf{A}$ is a linear transformation. Are there efficient algorithms to construct the monotonic features $\mathbf{y}$?*

Ofcourse, one might generalize this question by extending the possible form for $\mathbf{A}$ to something more general than linear. To define efficient algorithms, one must assume some form of oracle which evaluates f. The algorithm's complexity could then be measured by the number of calls to the oracle. Alternatively, one could assume that some data sampled according to some probability distribution are given, and measure complexity with respect to the number of data points. In both cases, one would like to get closer to the monotonic features $\mathbf{y}$ as the number of data points increases. The algorithmic efficiency of such procedures would also be a criterion to optimize.

A Some Mahalanobis Densities

We list some Mahalanobis densities, and their associated Mahalanobis distribution functions.

Name	$G(x)/g(x)$	$p(\mathbf{x})$
Gamma Density	$g(x) = Ax^k e^{-\alpha x^\rho}$	$A(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^k e^{-\alpha(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^\rho}$ $k > -\frac{d}{2},\ \rho > 0,\ \alpha = \left[\frac{\Gamma(\frac{d+2(k+1)}{2\rho})}{d\,\Gamma(\frac{d+2k}{2\rho})}\right]^\rho$ $A = \frac{\Gamma(\frac{d}{2})\rho\alpha^{(d+2k)/2\rho}}{\Gamma(\frac{d+2k}{2\rho})\lvert\boldsymbol{\Sigma}\rvert^{1/2}\pi^{d/2}}$
Gaussian	$G(x) = -\frac{2e^{-\frac{1}{2}x}}{(2\pi)^{d/2}\lvert\boldsymbol{\Sigma}\rvert^{1/2}}$	$N(\mathbf{x};\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}\lvert\boldsymbol{\Sigma}\rvert^{1/2}}e^{-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}}$
Exponential square root	$G(x) = -Ae^{-\sqrt{(d-1)x}}$	$A\sqrt{d-1}\frac{e^{-\sqrt{(d-1)\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}}}}{2\sqrt{\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}}}$ $d > 1$ $A = \frac{(d-1)^{d/2}\Gamma(\frac{d}{2})}{\Gamma(d)\lvert\boldsymbol{\Sigma}\rvert^{1/2}\pi^{d/2}}$
Polynomial ratio	$g(x) = \frac{Ax^p}{(1+ax)^{q+1}}$ For integer $p \geq 0$: $G(x) = -\frac{A}{a^{p+1}}\sum_{i=0}^{p} G_i(x),$ $G_i(x) = \frac{p!(q-i-1)!(ax)^{p-i}}{(p-i)!q!(1+ax)^{q-i}}$	$\frac{A(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^p}{(1+a\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})^{q+1}}$ $\frac{d}{2} + p > 0,\ q > \frac{d}{2} + p$ $a = \frac{1}{d}\left(\frac{\frac{d}{2}+p}{q-\frac{d}{2}-p}\right)$ $A = \frac{a^{d/2+p}\Gamma(q+1)\Gamma(\frac{d}{2})}{\lvert\boldsymbol{\Sigma}\rvert^{1/2}\pi^{d/2}\Gamma(\frac{d}{2}+p)\Gamma(q+1-\frac{d}{2}-p)}$
Linear combination	$G(x) = \sum_i A_i\alpha_i^{d/2}G_i(\alpha_i x)$ $G_i(x)$ are Mahalanobis	$\sum_i A_i\alpha_i^{d/2+1}g_i(\alpha_i\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})$ $\sum_i A_i = 1, \quad \sum_i A_i\alpha_i = 1,\ \alpha_i > 0,\ A_i > 0$ $g_i(x)$ are Mahalanobis

References

[1] Nir Ailon, Bernard Chazelle, Seshadhri Comandur, and Ding Liu. *Estimating the distance to a monotone function.* 2007.

[2] M. Banerjee and J. A. Wellner. Likelihood ratio tests for monotone functions. *Annals of Statistics*, 29:1699–1731, 2001.

[3] A. Ben-David. Monotonicity maintenance in information theoretic machine learning algorithms. *Machine Learning*, 19:29–43, 1995.

[4] P. Billingsley. *Probability and Measure.* Wiley Series in Probability and Mathematical Statistics. Wiley, 1986.

[5] A. W. Bowman, M. C. Jones, and I. Gubels. Testing monotonicity of regression. *Journal of Computational and Graphical Statistics*, 7(4):489–500, 1998.

[6] M. H. DeGroot. *Probability and Statistics.* Addison–Wesley, Reading, Massachusetts, 1989.

[7] Oded Goldreich, Shafi Goldwasser, Eric Lehman, Dana Ron, and Alex Samorodnitsky. Testing monotonicity. *Combinatorica*, 20(3):301–337, 2000.

[8] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series and Products, Corrected and Enlarged Edition.* Academic Press, San Diego, CA, 1980.

[9] Malik Magdon-Ismail, Justin (Hung-Ching) Chen, and Yaser S. Abu-Mostafa. The multilevel classification problem and a monotonicity hint. *Intelligent Data Engineering and Learning (IDEAL 02), Third International Conference*, August 2002.

[10] Enno Mammen. Estimating a smooth monotone regression function. *Annals of Statistics*, 19(2):724–740, June 1991.

[11] Hari Mukerjee. Monotone nonparametric regression. *Annals of Statistics*, 16(2):741–750, June 1988.

[12] Hari Mukerjee and Steven Stern. Feasible nonparametric estimation of multiargument monotone functions. *Journal of the American Statistical Association*, 89(425):77–80, March 1994.

[13] R. Potharst and A. J. Feelders. Classification trees for problems with monotonicity constraints. *SIGKDD Explorations*, 4(1):1–10, June 2002.

[14] Tim Robertson, F. T. Wright, and R. L. Dykstra. *Order Restricted Statistical Inference.* Wiley Series in Probability and Statistics. Wiley, new York, 1988.

[15] W. Schlee. Non-parametric tests of the monotony and convexity of regression. *in Non–Parametric Statistical Inference*, 2:823–836, 1982.

[16] J. Sill. The capacity of monotonic functions. *Discrete Applied Mathematics*, Special Issue on VC Dimension, 1998.

[17] J. Sill. Monotonic networks. In *Advances in Neural Information Processing Systems (NIPS)*, volume 10, 1998.

[18] J. Sill and Y. S. Abu-Mostafa. Monotonicity hints. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 9, pages 634–640. Morgan Kaufmann, 1997.

[19] Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6):1135–1151, November 1981.

[20] V. N. Vapnik. *Statistical Learning Theory.* Adaptive and Learning Systems for Signal Processing, Communications and Control. John Wiley & Sons, Inc., New york, 1998.