# Using A Linear Fit To Determine Monotonicity Directions

Malik Magdon-Ismail
Department of Computer Science, RPI
Rm 207 Lally Building,
110 8th Street, Troy, NY 12180, USA.
magdon@rpi.edu.
http://www.cs.rpi.edu/~magdon.

Joseph Sill
Plumtree Software,
500 Sansome Street
San Francisco, CA 94111, USA.
joe_sill@yahoo.com.

March 11, 2003

**Abstract**

Let $f$ be a monotonic function on $\mathbb{R}^d$. 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 dependent, 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.

## 1 Introduction and Results

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 $\mathbf{x} \in \mathbb{R}^d$. When the context is clear, we will use the notation $f_i(x_i)$ to denote the function $f$ of $x_i$ with all other variables held constant. We assume throughout that we are in $\mathbb{R}^d$. The direction of monotonicity is negative if the condition $\Delta > 0$ is replaced by $\Delta < 0$. A function is *monotonic* if it is monotonic in every dimension. If $f$ is only defined on some subset of $\mathbb{R}^d$, then the monotonicity conditions need hold only in this subset. We can represent the monotonicity directions of such a function by a $d$ dimensional vector $\mathbf{m}$ of $\pm 1$'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 $f_i(x_i + m_i\Delta) \geq f_i(x_i)$, for all $\Delta \geq 0$.

Monotonicity is a property that might be true of a function that one might like to determine on the basis of some data. For example, the credit worthiness of an individual would be a monotonic function of variables such as income, [13]. The severity of a heart condition should be a monotonic function of cholesterol level. One might wish to learn such a function from a finite data set, for predictive purposes. In such cases, incorporating the monotonicity constraint can significantly enhance the performance of the resulting predictor, because the capacity 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, [11, 13, 14]

Some tests for the monotonocity of a regression function have been considered in the literature, see for example [3, 10]. Algorithms for enforcing monotonicity have also been considered, for example [1, 5, 6, 7, 8, 9, 12, 13]. 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 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. Exactly such a problem was encountered in [13]. It can also be argued that the general multilevel classification problem admits a monotonicity constraint, even though the directions are not known *apriori* [5]. In such cases, it is not practical to enforce monotonicity in one of the $2^d$ possible directions, especially when $d$ is large. Rather, one would like to determine a specific direction in which to enforce the monotonicity.

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 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 [13]. 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 $E[\mathbf{x}] = \mathbf{0}$ (Lemmas 2.1, 2.2). In general we will postpone proofs to Section 2. First we state how to obtain the linear fit.

**Lemma 1.1 (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 [4]. ■

The main content of this paper is to determine conditions under which the linear fit in (3) will produce the correct monotonicity directions for the function $f(\mathbf{x})$. First we give the result for independent input densities.

**Theorem 1.2 (Independent densities.)** *Let $f(\mathbf{x})$ be monotonic with monotonicity direction $\mathbf{m}$, and let the input probability density be any independent density* [1]*. 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 $w_i^l \neq 0$.*

Thus, when the inputs are independent, the linear fit deduces the correct monotonicity directions for $f$, even though $f$ may not resemble a linear function in any way. Further, note that the theorem 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. An important special case is when the function is defined on a hyper–rectangle. In this case, we can take the measure to be uniform on the rectangle.

Independence in the input dimensions is quite a strong restriction, and much of the benefit of the monotonicity constraint is due to the fact that the input dimensions are *not independent*. This is evident from the fact that the VC-dimencion of the class of monotonic classification functions is $\infty$, but the capacity of this class is heavily dependent on the input distribution. When the input dimensions are independent, the capacity of the class of monotonic functions grows exponentially in $N$, but when the input dimensions are dependent, the capacity can be a much more slowly growing function. Such issues are discussed in greater detail in [11]. Unfortunately, if we remove the independence requirement, then we cannot guarantee that the optimal linear fit will induce the correct monotonicity directions. The following proposition establishes this fact, and an explicit example is constructed in the proof in Section 2.

**Proposition 1.3** *There exist monotonic functions $f$ and input densities $p_{\mathbf{x}}(\mathbf{x})$ for which the optimal linear fit induces the incorrect monotonicity directions.*

The essential idea is to choose a function like $f(\mathbf{x}) = x_1^3 - x_2$. 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.

We cannot remove the independence restriction in general, however, for certain special cases we can. In particular, a common assumption is that the inputs are jointly Gaussian. In this case, the linear fit will correctly induce the monotonicity directions. This result is a special case of a more general one dealing with a class of input densities which we call *Mahalanobis densities.*

**Definition 1.4** *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.*

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

The first constraint on $G$ is merely technical, stating that $G$ decays "quickly" to zero.[2]. The second ensures the $p_{\mathbf{x}}$ is a legitimate density, integrating to 1. The third merely enforces the consistency constraint that $\mathbf{\Sigma} = \int d\mathbf{x}\ g(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\mathbf{x}\mathbf{x}^T$. The Gaussian density function is defined by

$$N(\mathbf{x}; \mathbf{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}} \tag{4}$$

where $\mathbf{\Sigma}$ 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 $N(\mathbf{x} - \boldsymbol{\mu}; \mathbf{\Sigma})$. It is easily verified that every Gaussian density is a Mahalanobis density with Mahalanobis distribution function $G(x) = -2e^{-\frac{1}{2}x}/(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}$. The next theorem shows that the linear fit is faithful to the monotonicity directions of $f(\mathbf{x})$ whenever the input density is a Mahalanobis density.

**Theorem 1.5 (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.,*

$$\lim_{|x_i|\to\infty} G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}^T)f(\mathbf{x}) = 0 \qquad \forall i = 1, \ldots, d. \tag{5}$$

*Let $\mathbf{w}^l$ 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 some compact set of positive measure, then $w_i^l \neq 0$.*

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

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 regression model. $\mathbf{x}_i$ are sampled independently from $p_{\mathbf{x}}(\mathbf{x})$ and $y_i = f(\mathbf{x}_i) + \epsilon_i$, where $\epsilon_i$ is noise. In the regression case, we assume that the $\epsilon_i$ are independent zero mean noise, with bounded fourth moments.

$$E[\epsilon_i|\mathbf{x}_i] = 0, \quad E[\epsilon_i\epsilon_j|\mathbf{x}_i, \mathbf{x}_j] = \sigma^2\delta_{ij}, \tag{6}$$

where $\delta_{ij}$ 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 $\epsilon_i$ is independent flip noise, i.e., independent flips of the output values from $y_i$ to $-y_i$ with some probability $p < \frac{1}{2}$.

$$\epsilon_i = \begin{cases} 0 & \text{w.p. } 1 - p, \\ -2f(\mathbf{x}_i) & \text{w.p. } p. \end{cases} \tag{7}$$

[2]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 $d \geq 3$.

Define the augmented input vector by

$$\hat{\mathbf{x}}_i = \begin{bmatrix} 1 \\ \mathbf{x}_i \end{bmatrix}, \tag{8}$$

and define $\mathbf{X}_N$ by

$$\mathbf{X}_N = \frac{1}{N}\sum_{i=1}^{N} \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i^T = \frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} 1 & \mathbf{x}_i^T \\ \mathbf{x}_i & \mathbf{x}_i\mathbf{x}_i^T \end{bmatrix}. \tag{9}$$

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

$$\boldsymbol{\beta}^* = \begin{bmatrix} w_0^* \\ \mathbf{w}^* \end{bmatrix} = \frac{\mathbf{X}^{-1}}{N}\sum_{i=1}^{N} y_i \hat{\mathbf{x}}_i. \tag{10}$$

PROOF: See any standard book on statistics, for example [4]. ■

Under reasonable conditions, when $N \to \infty$, we expect sample averages to converge to expectations, i.e.,

$$\frac{1}{N}\sum_i \mathbf{x}_i \to E[\mathbf{x}] = \mathbf{0}, \qquad \mathbf{X}_N^{-1} \to \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \boldsymbol{\Sigma}^{-1} \end{bmatrix}, \qquad \sum_{i=1}^{N} y_i \hat{\mathbf{x}}_i \to \begin{bmatrix} E[f(\mathbf{x})] \\ E[f(\mathbf{x})\mathbf{x}] \end{bmatrix}. \tag{11}$$

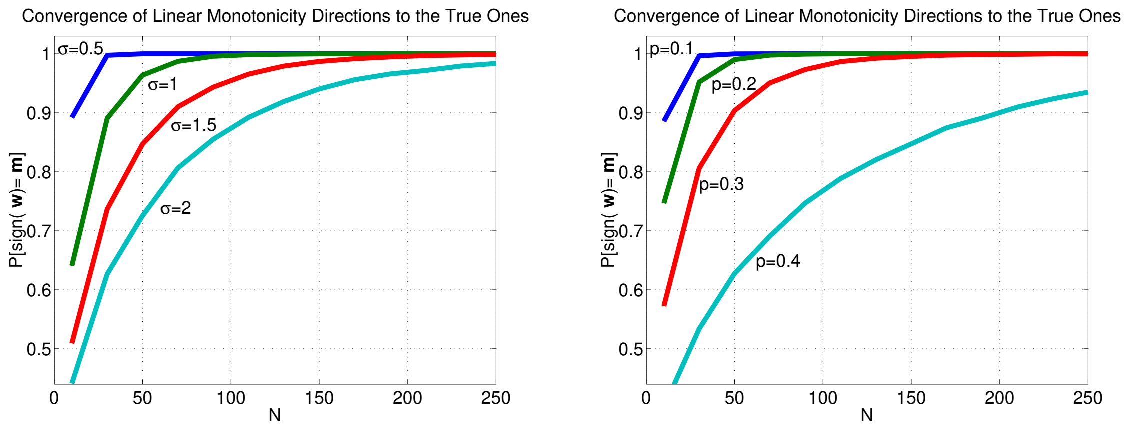
Thus, the OLS estimates should converge to the true linear fit. The following lemma is therefore not surprising.

**Lemma 1.7** *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 $\epsilon_i$ 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 $\epsilon_i$ is independent noise. For regression, the noise satisfies satisfying (6), and for classification, the noise is independent flip noise (7). Let $\mathbf{w}^l$ be given by the exact linear fit, and let $\mathbf{w}^*$ be the OLS estimator of $\mathbf{w}^l$. Then,*

$$\mathbf{w}^* \xrightarrow{P} \begin{cases} \mathbf{w}^l & \textit{regression}, \\ (1-2p)\mathbf{w}^l & \textit{classification}. \end{cases} \tag{12}$$

We use the standard notation $\xrightarrow{P}$ to denote convergence in probability. 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 following theorem is therefore evident.

**Theorem 1.8 (OLS)** *Let $f(\mathbf{x})$ be monotonic with monotonicity 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 $\epsilon_i$ is independent noise. Assume that all fourth order moments are finite. If $f$ is a classification function, then $\epsilon_i$ is a flip noise with probability $p < \frac{1}{2}$ (7), otherwise it is a zero mean random variable with variance $\sigma^2$ (6).*




Regression | Classification

Figure 1: Dependence of the probability of obtaining the correct monotonocity directions as a function of the sample size.

*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 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 densities and Mahalanobis densities. This convergence is 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 both classification and regression. The data were sampled uniformly from $[0,1]^2$ and $f(x,y) = e^{y-x} - \frac{1}{2}$ for regression and the sign of this function for classification.

## Discussion

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, this is so. Once the direction of monotonicity is known, 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 is appealing for at least two reasons. The first is the it is easy to implement. The second is that once it has been implemented, the monotonicity directions are easy to determine. We presented sufficient conditions under which using a linear model would work correctly. Necessary and sufficient conditions for such an algorithm are still missing. Other approaches to determining the monotonicity of a function that are as simple and efficient as fitting a linear model would also be a significant advance. There is potential that some non parametric techniques could prove useful in this respect.

Another interesting and unanswered question is, "How bad can the linear model be?". The example constructed in Proposition 1.3 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. To be more specific, if the linear model gave monotonicity direction $\mathbf{m}' \neq \mathbf{m}$, and one obtains the best monotonic fit subject to the incorrect monotonicity constraints $\mathbf{m}'$, then how bad can the expected fit error be? We leave such questions as food for future thought.

# 2 Proofs

The following lemmas will be useful in the proof of Theorem 1.2. 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 2.1 (Monotonicity direction is scale and translation invariant)** *Let $f(\mathbf{x})$ be monotonic with monotonicity 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 $\Delta > 0$. $\hat{g}(x_i + \Delta) = f_i(A_{ii}x_i + b_i + A_{ii}\Delta)$. If $A_{ii} > 0$, then $f_i(A_{ii}x_i + b_i + A_{ii}\Delta) \geq f_i(A_{ii}x_i + b_i) = \hat{g}(x_i)$. Similarly if $A_{ii} < 0$, then $f_i(A_{ii}x_i + b_i + A_{ii}\Delta) \leq f_i(A_{ii}x_i + b_i) = \hat{g}(x_i)$. An analogous argument with $m_i = -1$ and $\Delta < 0$ completes the proof. ■

**Lemma 2.2** *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. ■

**Lemma 2.3** *Let $f(\mathbf{x})$ be monotonic with monotonicity 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. ■

**Proof of Theorem 1.2** By Lemmas 2.1 and 2.2, after suitable scaling and translation, we can assume, without loss of generality, that $E[\mathbf{x}] = \mathbf{0}$ and that $E[\mathbf{x}\mathbf{x}^T] = \mathbf{I}$. Then, using Lemma 1.1, we have that

$$\mathbf{w} = \int_{-\infty}^{\infty} d\mathbf{x}\; p_{\mathbf{x}}(\mathbf{x})\mathbf{x}f(\mathbf{x}) \qquad\qquad w_0 = \int_{-\infty}^{\infty} d\mathbf{x}\; p_{\mathbf{x}}(\mathbf{x})\; f(\mathbf{x}). \tag{13}$$

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

$$\begin{aligned}
m_i w_i &\overset{(a)}{=} \int d\mathbf{x}' \int_{-\infty}^{\infty} dx_i\; p_{\mathbf{x}}(\mathbf{x})\; m_i x_i f_i(x_i), &(14)\\
&\overset{(b)}{\geq} \int d\mathbf{x}'\; p_{\mathbf{x}'}(\mathbf{x}') \int_{-\infty}^{\infty} dx_i\; p_{x_i}(x_i) m_i x_i f_i(0), &(15)\\
&= \int d\mathbf{x}'\; p_{\mathbf{x}'}(\mathbf{x}') m_i f_i(0) \int_{-\infty}^{\infty} dx_i\; p_{x_i}(x_i) x_i, &(16)\\
&\overset{(c)}{=} 0, &(17)\\
& &(18)
\end{aligned}$$

where $\mathbf{x}' = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$. (a) follows since the measure is independent, (b) by Lemma 2.3 and (c) because $E[\mathbf{x}] = \mathbf{0}$, concluding the proof. Note that if $w_i = 0$, the result is ambiguous and could be an artifact of the measure. However, from Lemma 2.3 we see that if $f_{ii}$ 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 strict, and hence the result is that $m_i w_i > 0$, concluding the proof of the theorem. ∎

**Proof of Proposition 1.3.** 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,

$$p_{\mathbf{x}}(x_1, x_2) = \frac{1}{2} N(x_1 - a_1) N(x_2 - 1) + \frac{1}{2} N(x_1 + a_1) N(x_2 + 1). \tag{19}$$

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 [4], we find that

$$\mathbf{\Sigma} = \begin{bmatrix} 1 + a_1^2 & a_1 \\ a_1 & 2 \end{bmatrix}, \qquad E[x_1^4] = a_1^4 + 6a_1^2 + 3, \qquad E[x_1^3 x_2] = a_1^3 + 3a_1. \tag{20}$$

The optimal linear fit is given by

$$\begin{aligned}
\mathbf{w} &= \mathbf{\Sigma}^{-1} E[f(\mathbf{x})\mathbf{x}] = \mathbf{\Sigma}^{-1} \begin{bmatrix} E[x_1^4] - E[x_1 x_2] \\ E[x_1^3 x_2] - E[x_2^2] \end{bmatrix}, &(21)\\
&= \frac{1}{2 + a_1^2} \begin{bmatrix} a_1^4 + 9a_1^2 + 6 \\ -2a_1^3 - a_1^2 - 2 \end{bmatrix}. &(22)
\end{aligned}$$

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

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, $p_{\mathbf{x}}(\mathbf{x}) = g(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})$ for some function $g$ defined on $\mathbb{R}_+$. 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

$$\lim_{|x|\to\infty} G(x^2)x = 0 \tag{23}$$

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

$$\begin{aligned}
\mathbf{\Sigma} &= \int d\mathbf{x}\ g(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\mathbf{x}\mathbf{x}^T && (24)\\
&= \mathbf{\Sigma}\int d\mathbf{x}\ g(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\mathbf{\Sigma}^{-1}\mathbf{x}\mathbf{x}^T && (25)\\
&= \frac{1}{2}\mathbf{\Sigma}\int d\mathbf{x}\ \left[\nabla_{\mathbf{x}}G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\right]\mathbf{x}^T && (26)\\
&= \frac{1}{2}\mathbf{\Sigma}\int d\mathbf{x}\ \nabla_{\mathbf{x}}\big(G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\mathbf{x}^T\big) - G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x})\nabla_{\mathbf{x}}\mathbf{x}^T && (27)\\
&= -\frac{1}{2}\mathbf{\Sigma}\int d\mathbf{x}\ G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}^T) && (28)
\end{aligned}$$

where the last line follows because, using the fundamental theorem of calculus and (23), the first term is zero, and, $\nabla_{\mathbf{x}}\mathbf{x}^T = \mathbf{I}$. Thus we have the two constraints,

$$\int d\mathbf{x}\ G'(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}^T) = 1 \qquad \int d\mathbf{x}\ G(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}^T) = -2. \tag{29}$$

The first constraint can always be effected by multiplying $G$ by some positive scalar. The second then leads to a constraint on $G$. Using some standard multidimensional integration techniques, these two constraints can be reduced to

$$\int_0^\infty ds\ s^{d-1}G'(s^2) = \frac{\Gamma(d/2)}{2|\mathbf{\Sigma}|^{1/2}\pi^{d/2}} \qquad \int_0^\infty ds\ s^{d-1}G(s^2) = -\frac{\Gamma(d/2)}{|\mathbf{\Sigma}|^{1/2}\pi^{d/2}}, \tag{30}$$

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

$$\int_0^\infty ds\ s^{d-1}g(s^2) = \frac{\Gamma(d/2)}{2|\mathbf{\Sigma}|^{1/2}\pi^{d/2}} \qquad \int_0^\infty ds\ s^{d+1}g(s^2) = \frac{\Gamma(d/2+1)}{|\mathbf{\Sigma}|^{1/2}\pi^{d/2}}. \tag{31}$$

The classification boundary with respect to dimension $x_i$ is a function $f_i^c(\mathbf{x}') : \mathbb{R}^{d-1} \mapsto \{\mathbb{R}, \infty, -\infty\}$, 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,

$$f_i(x_i) = \begin{cases} m_i & x_i \geq f_i^c(x') \\ -m_i & x_i < f_i^c(x') \end{cases} \tag{32}$$

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 $\mathbf{m}^c = -m_i\mathbf{m}'$.

**Proof of Theorem 1.5.** Let $p_{\mathbf{x}}(\mathbf{x}) = g(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}^T)$ satisfy the properties described above. Let $f$ be a monotonic function with monotonicity direction $\mathbf{m}$ satisfying[3]

$$\lim_{|x_i|\to\infty} G(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}^T)f(\mathbf{x}) = 0 \qquad \forall i = 1,\ldots,d. \tag{33}$$

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

$$\begin{aligned}
\mathbf{w} &= \int d\mathbf{x}\; g(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})\boldsymbol{\Sigma}^{-1}\mathbf{x}f(\mathbf{x}), && (34)\\
&\overset{(a)}{=} \frac{1}{2}\int d\mathbf{x}\;\left[\nabla_{\mathbf{x}}G(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})\right]f(\mathbf{x}), && (35)\\
&= \frac{1}{2}\int d\mathbf{x}\;\nabla_{\mathbf{x}}\big(G(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})f(\mathbf{x})\big) - G(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})\nabla_{\mathbf{x}}f(\mathbf{x}), && (36)\\
&\overset{(b)}{=} -\frac{1}{2}\int d\mathbf{x}\; G(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})\nabla_{\mathbf{x}}f(\mathbf{x}), && (37)\\
&\overset{(c)}{=} -\frac{1}{2}\left(\int d\mathbf{x}\; G(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})\boldsymbol{\Lambda}(\mathbf{x})\right)\mathbf{m}, && (38)\\
&\overset{(d)}{=} \boldsymbol{\Lambda}\mathbf{m}, && (39)
\end{aligned}$$

where $\boldsymbol{\Lambda}(\mathbf{x})$ and $\boldsymbol{\Lambda}$ are a non-negative diagonal matrices. (a) follows by the definition of $G$; (b) follows by using the fundamental theorem of calculus and (33); (c) follows because $f$ is monotonic with monotonicity direction $\mathbf{m}$, therefore $\nabla_{\mathbf{x}}f(\mathbf{x})$ must have the same sign as $\mathbf{m}$ and hence can be written as $\boldsymbol{\Lambda}(\mathbf{x})\mathbf{m}$; (d) follows because $-G$ is non-negative. Thus all the non-zero components of $\mathbf{w}$ have the same sign as $\mathbf{m}$ and the theorem follows. Note that if for each $i$ and some $\epsilon > 0$, $|G(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})\boldsymbol{\Lambda}_{ii}(\mathbf{x})| \geq \epsilon$ 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, (35) gives

$$\begin{aligned}
w_i &\overset{(a)}{=} -\frac{m_i}{2}\int d\mathbf{x}'\left[\int_{-\infty}^{f_i^c(\mathbf{x}')}\frac{\partial}{\partial x_i}G(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}) - \int_{f_i^c(\mathbf{x}')}^{\infty}\frac{\partial}{\partial x_i}G(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})\right], && (40)\\
&\overset{(b)}{=} -m_i\int d\mathbf{x}'\; G(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x})\big|_{x_i = f_i^c(\mathbf{x}')} && (41)\\
&\overset{(c)}{=} \lambda_i m_i && (42)
\end{aligned}$$

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. ∎

The following lemmas will prove useful in the proof of Theorem 1.8. 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 $\epsilon_i$ satisfies (6) for regression, and (7) for classification.

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

**Lemma 2.4 (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*

$$E_\epsilon[\mathbf{w}^*] = \begin{cases} \mathbf{w}'(1-2p) & \text{classification,} \\ \mathbf{w}' & \text{regression,} \end{cases} \tag{43}$$

*where the expectation is with respect to the noise.*

Proof: By Lemma 1.6,

$$\mathbf{w}^* = \mathbf{w}' + \frac{\mathbf{X}_N^{-1}}{N}\sum_{i=1}^{N}\epsilon_i\hat{\mathbf{x}}_i, \tag{44}$$

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. ∎

**Lemma 2.5 (Covariance of the OLS estimator)** *Let* $\mathbf{w}^*$ *be the OLS estimator, then*

$$Cov(\mathbf{w}^*) = \begin{cases} \dfrac{\sigma^2\mathbf{X}_N^{-1}}{N} & \text{regression,} \\ \dfrac{4p(1-p)\mathbf{X}_N^{-1}}{N} & \text{classification.} \end{cases} \tag{45}$$

Proof: For regression,

$$\begin{aligned} E[(\mathbf{w}^* - E[\mathbf{w}^*])(\mathbf{w}^* - E[\mathbf{w}^*])^T] &= \frac{\mathbf{X}_N^{-1}}{N}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{\mathbf{x}}_i\hat{\mathbf{x}}_j^T E[\epsilon_i\epsilon_j]\right)\frac{\mathbf{X}_N^{-1}}{N}, &(46)\\ &\overset{(a)}{=} \frac{\sigma^2\mathbf{X}_N^{-1}}{N}, &(47) \end{aligned}$$

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

$$\begin{aligned} E[(\mathbf{w}^* - E[\mathbf{w}^*])(\mathbf{w}^* - E[\mathbf{w}^*])^T] &= \frac{\mathbf{X}_N^{-1}}{N}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{\mathbf{x}}_i\hat{\mathbf{x}}_j^T E[(2pf(\mathbf{x}_i)+\epsilon_i)(2pf(\mathbf{x}_j)+\epsilon_j)]\right)\frac{\mathbf{X}_N^{-1}}{N}, \\ &\overset{(b)}{=} \frac{4p(1-p)\mathbf{X}_N^{-1}}{N}, &(48) \end{aligned}$$

where (b) follows because $f(\mathbf{x}_i)^2 = 1$ and so using (7) and the independence of the $\epsilon_i$, we get that $E[(2pf(\mathbf{x}_i)+\epsilon_i)(2pf(\mathbf{x}_j)+\epsilon_j)] = 4p(1-p)\delta_{ij}$, from which the lemma follows. ∎

**Lemma 2.6** *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 [2]. ∎

**Lemma 2.7**

$$\mathbf{X}_N \xrightarrow{P} \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{\Sigma} \end{bmatrix} \qquad\qquad \mathbf{X}_N^{-1} \xrightarrow{P} \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{\Sigma}^{-1} \end{bmatrix} \tag{49}$$

PROOF: The first result follows by the weak law of large numbers because the fourth order moments are bounded. Since $\mathbf{\Sigma}$ is invertible, the function $\mathbf{X}_N^{-1}$ is continuous at $\mathbf{X}_N = \mathbf{\Sigma}$. Therefore, by Lemma 2.6, the second result also holds. ∎

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

**Lemma 2.8** *The OLS estimator has a distribution that is asymptotically Gaussian, given by*

$$\boldsymbol{\beta}^* \xrightarrow{P} N(\bar{\boldsymbol{\beta}}; Q) \tag{50}$$

*where $\bar{\boldsymbol{\beta}}$ is the mean of the estimator, given in Lemma 2.4 and the covariance matrix $Q$ is given by Lemma 2.5. Therefore, $\boldsymbol{\beta}^* \xrightarrow{P} \bar{\boldsymbol{\beta}}$. 2.7.*

PROOF: The fact that $\boldsymbol{\beta}^* \xrightarrow{P} N(\bar{\boldsymbol{\beta}}, Q)$ is a standard result, see for example [4]. by Lemmas 2.5, 2.7, 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 of Lemma 1.7** Let $\hat{\mathbf{\Sigma}} = E[\hat{\mathbf{x}}\hat{\mathbf{x}}^T]$. By Lemma 2.7, $X_N^{-1} \xrightarrow{P} \hat{\mathbf{\Sigma}}^{-1}$. 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 $\mathbf{w}' \xrightarrow{P} \mathbf{\Sigma}^{-1}E[f(\mathbf{x})\hat{\mathbf{x}}] = \mathbf{w}^l$. By Lemma 2.8, $\mathbf{w}^* \xrightarrow{P} \mathbf{w}'$ 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. ∎

# References


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

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

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

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

[5] 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.

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

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

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

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

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

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

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

[13] 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.

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

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