Feature transforms

Story so far

$$\textbf{Predict } y \in \ \textbf{ as } \begin{cases} y = \mathbf{x}^T \mathbf{w} & \text{(prediction function)} \\ p(y \mid \mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}\left(y \mid \mathbf{x}^T \mathbf{w}, \sigma^2\right) & \text{(probabilistic view)} \end{cases}$$

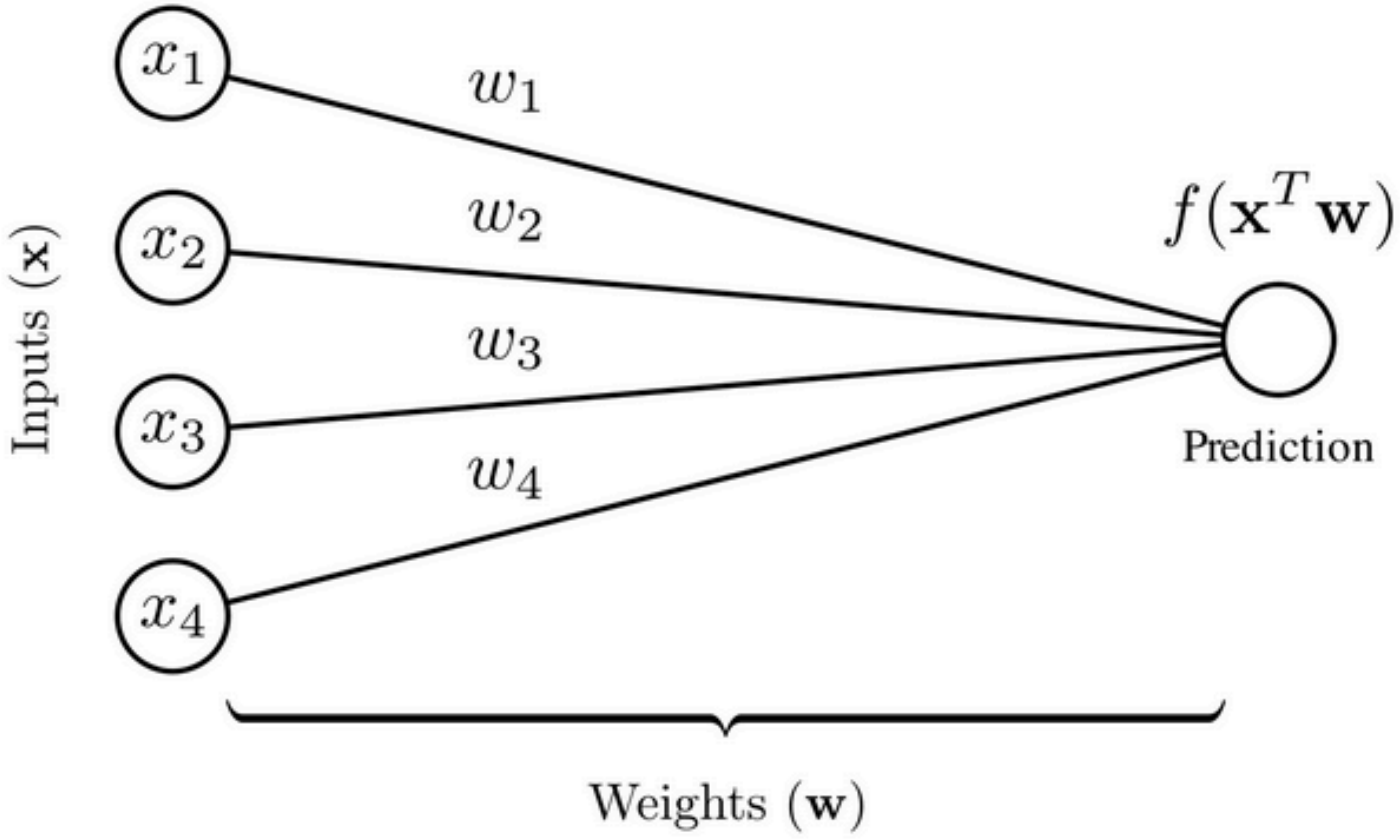
A reasonable model for *binary* outputs ($y \in \{0, 1\}$) is **logistic regression**:

$$\textbf{Predict } y \in \ \textbf{ as } \begin{cases} y = \mathbb{I}(\mathbf{x}^T \mathbf{w} > 0) & \text{(prediction function)} \\ p(y = 1 \mid \mathbf{x}, \mathbf{w}) = \sigma(\mathbf{x}^T \mathbf{w}) & \text{(probabilistic view)} \end{cases}$$

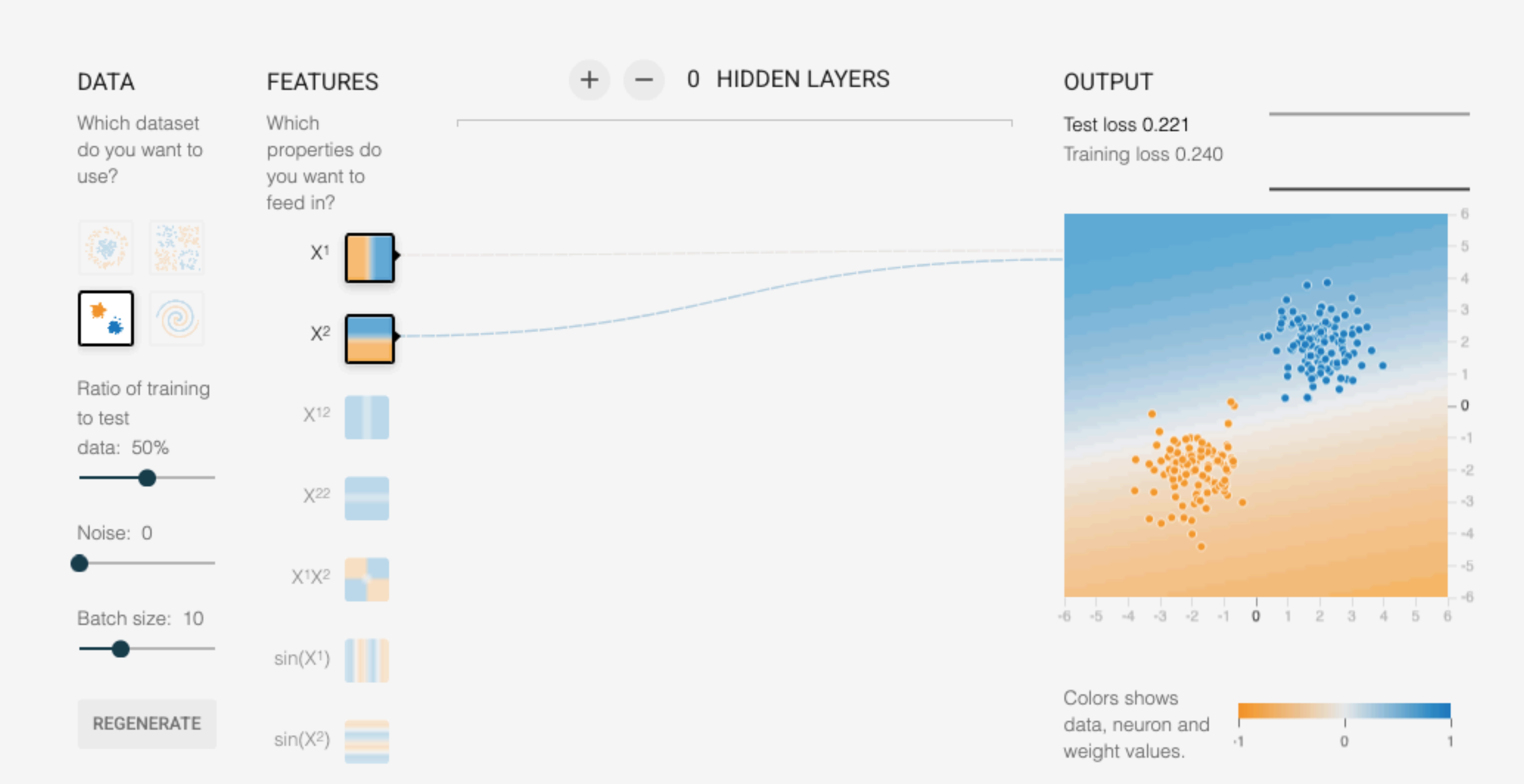
A reasonable model for *categorical* outputs ($y \in \{0, 1, \dots, C\}$) is **multinomial logistic regression**:

$$\textbf{Predict } y \in \ \textbf{ as } \begin{cases} y = \underset{c}{\operatorname{argmax}} \ \mathbf{x}^T \mathbf{w}_c & \text{(prediction function)} \\ p(y = c \mid \mathbf{x}, \mathbf{w}) = \operatorname{softmax}(\mathbf{x}^T \mathbf{W})_c & \text{(probabilistic view)} \end{cases}$$

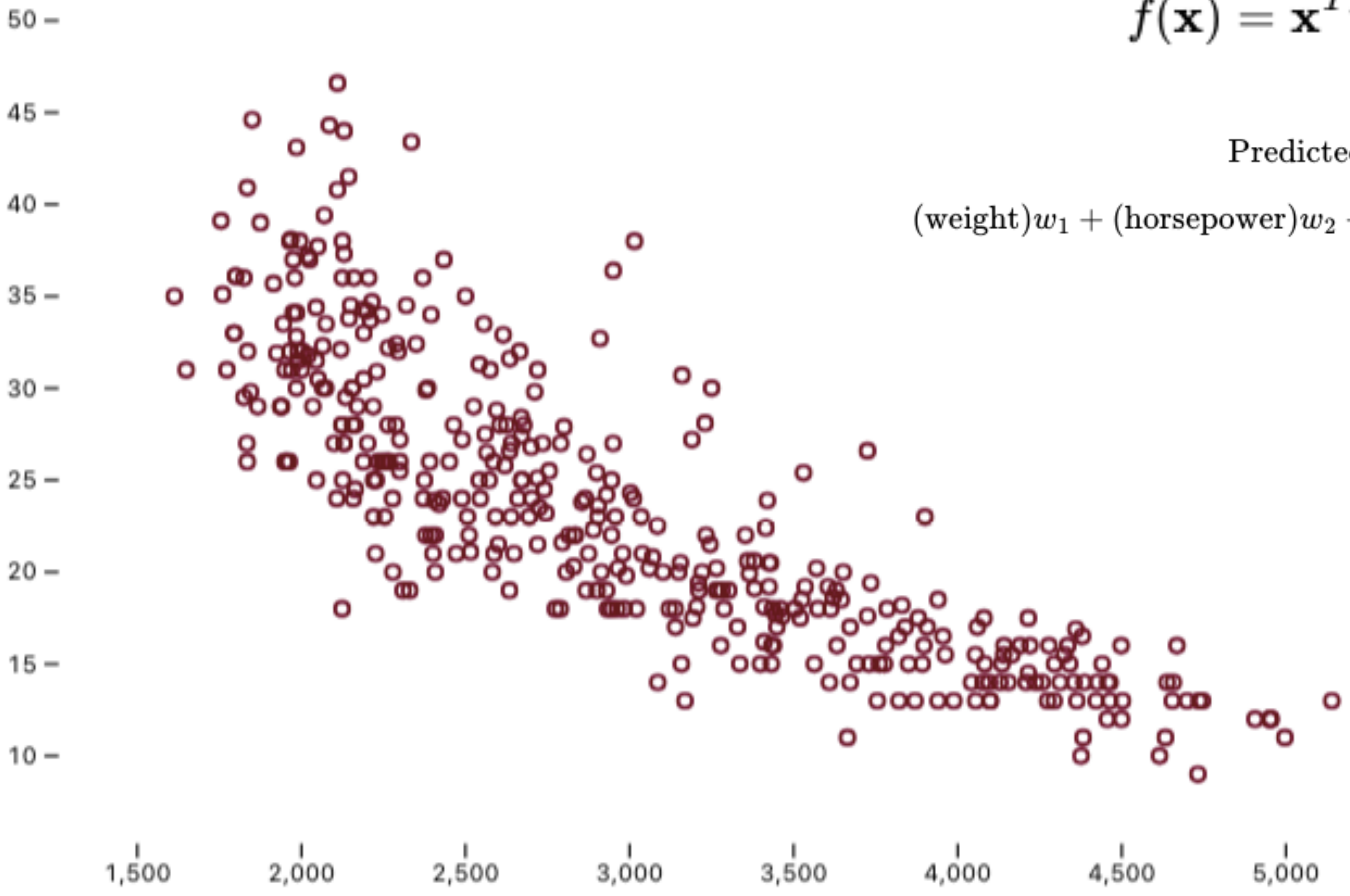
A new visualization



playground.tensorflow.org



(Approx.) Linear data

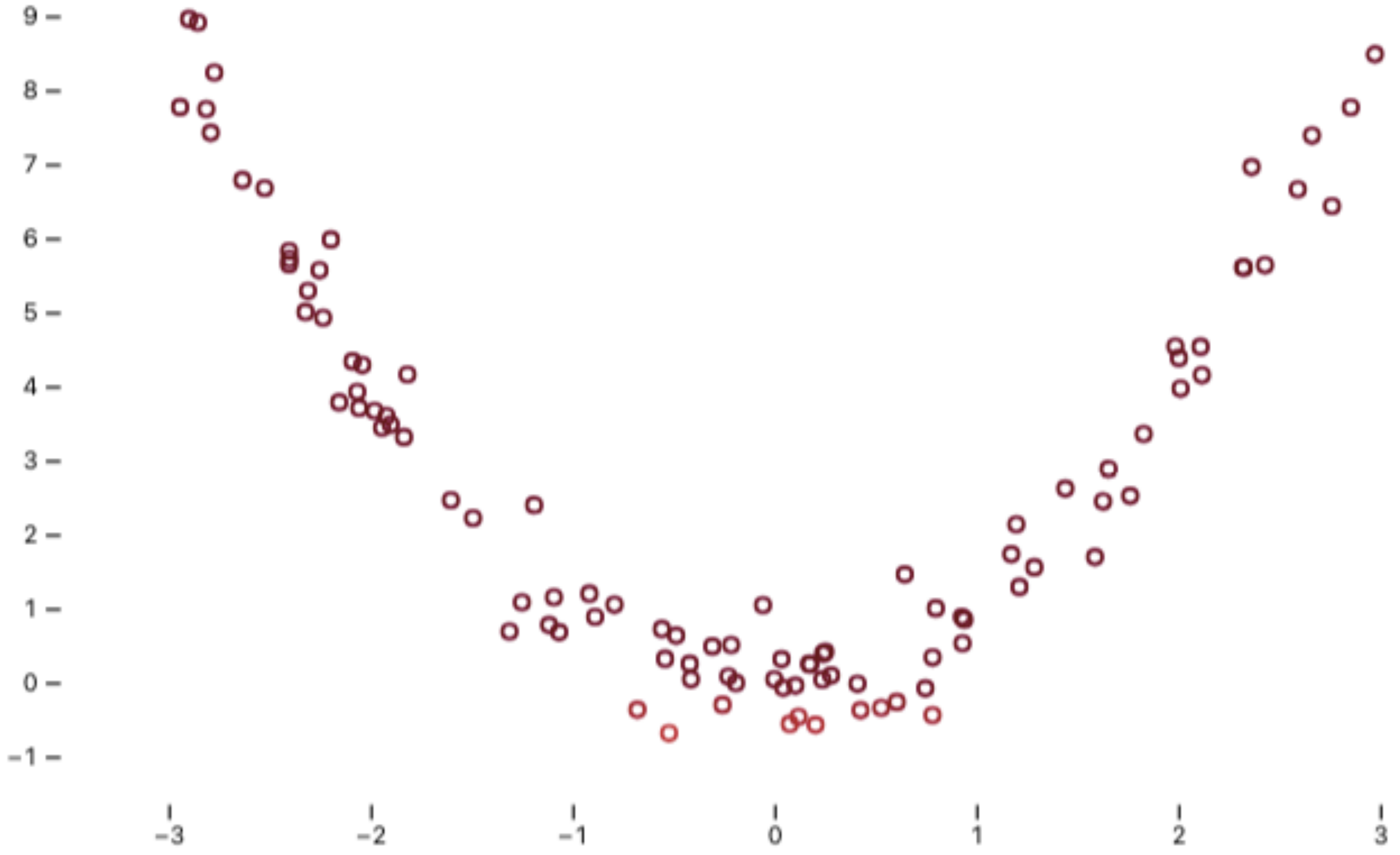


$$f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} = \sum_{i=1}^{n} x_i w_i$$

Predicted MPG $= f(\mathbf{x}) =$

$(\text{weight})w_1 + (\text{horsepower})w_2 + (\text{displacement})w_3 + (\text{0-60mph})w_4 + b$

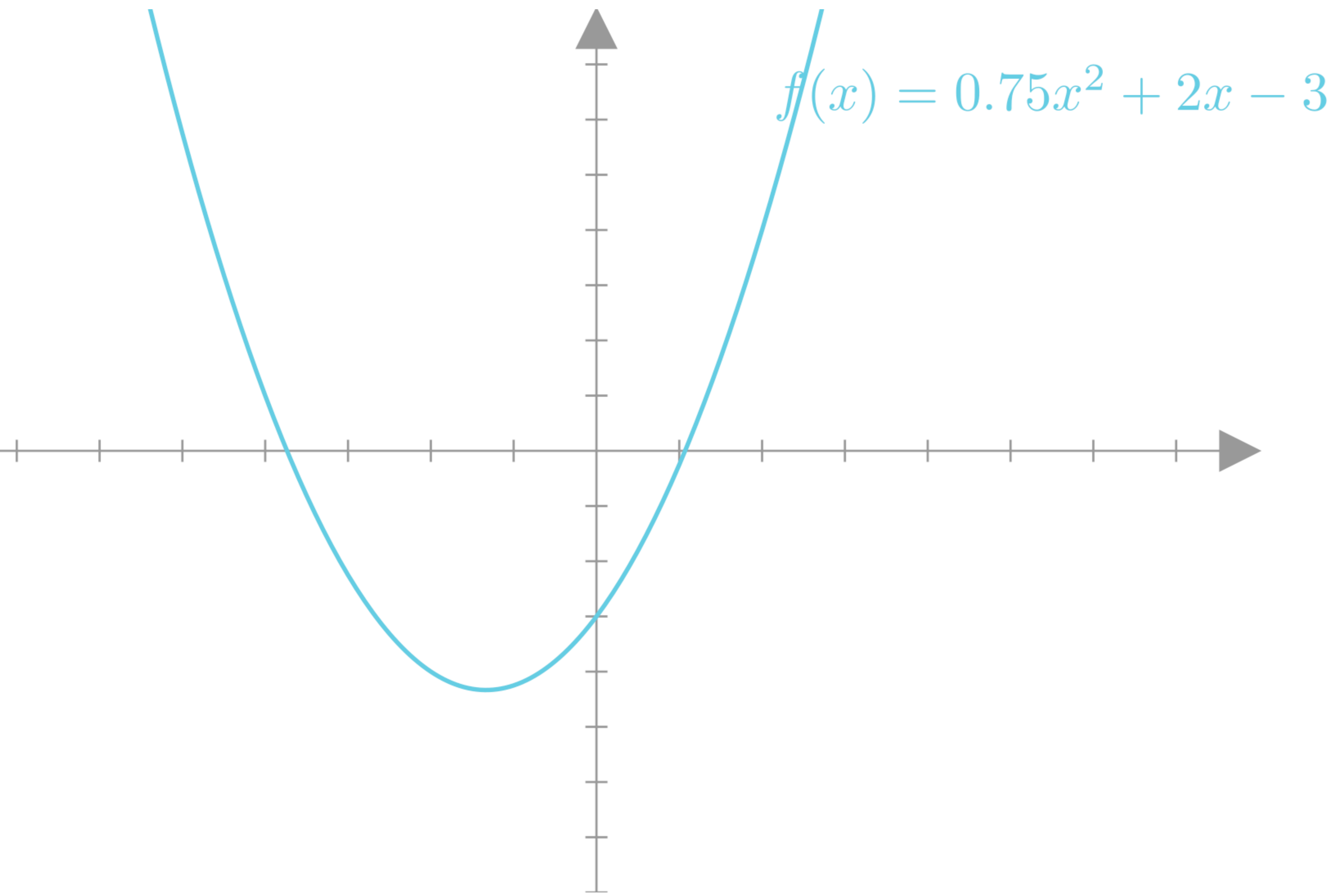
Non-Linear data



Polynomial functions

Quadratic function

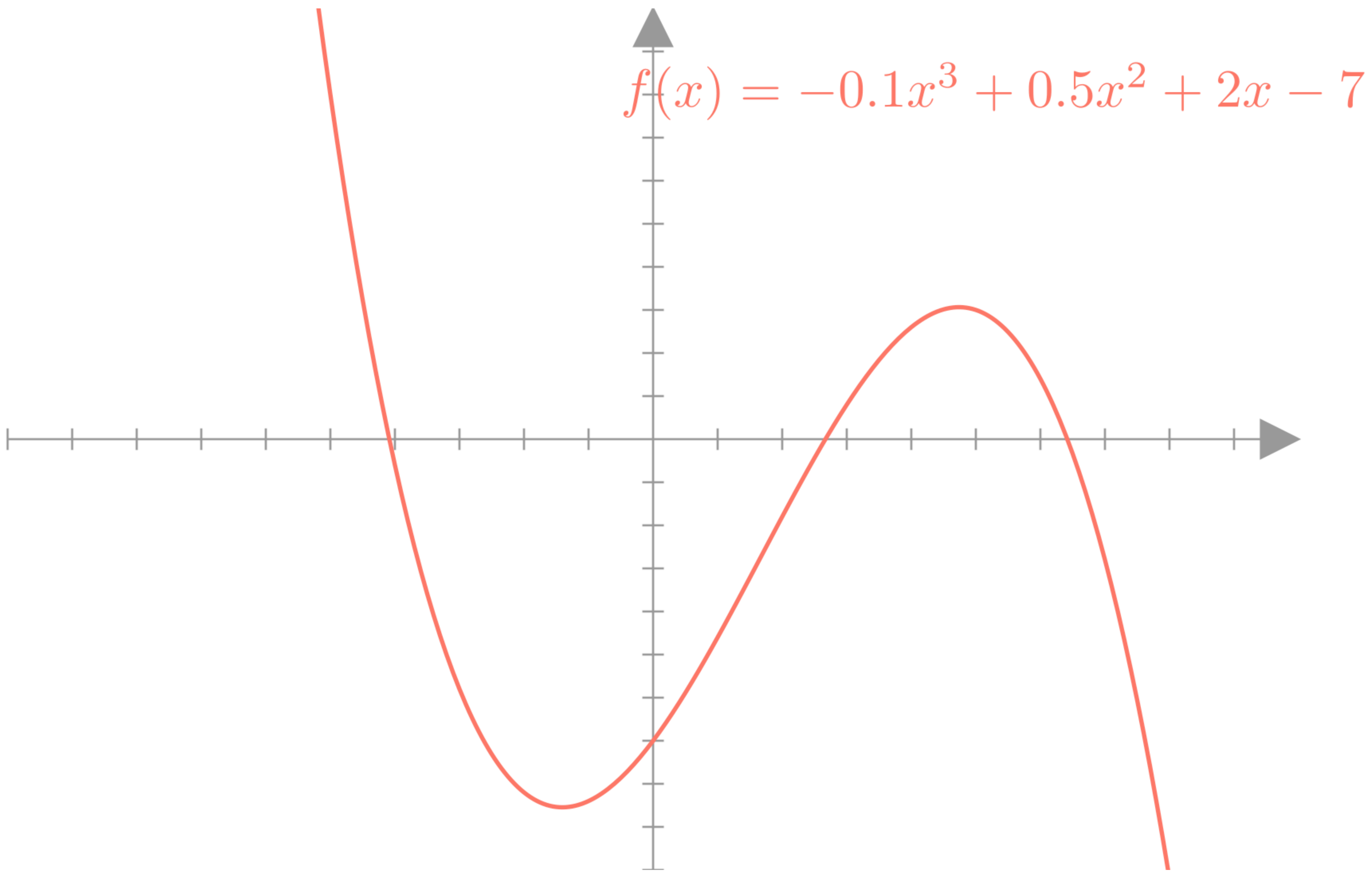
$$f(x) = w_2 x^2 + w_1 x + b$$



Degree (highest power): 2

Cubic function

$$f(x) = w_3 x^3 + w_2 x^2 + w_1 x + b$$



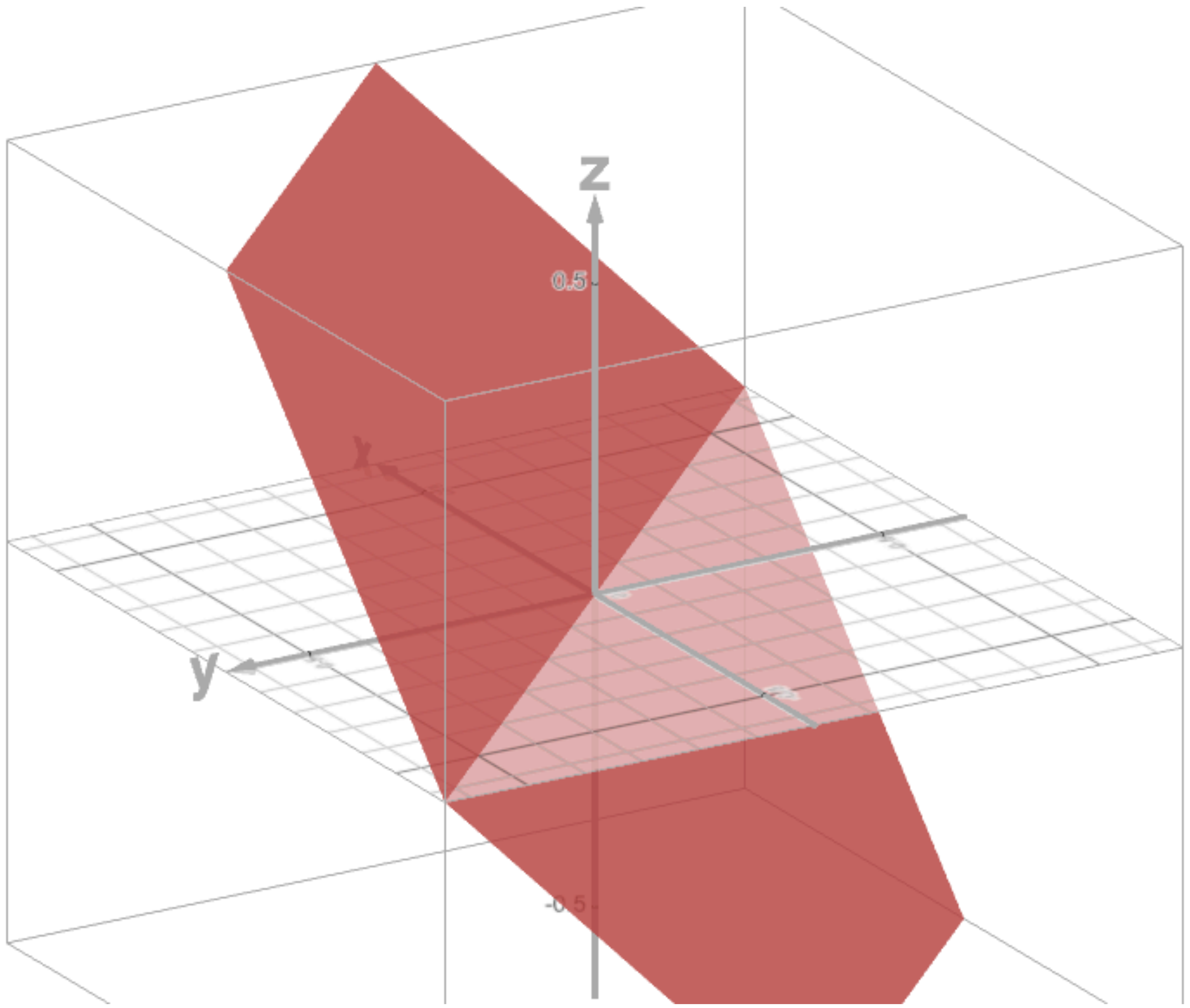
Degree (highest power): 3

Polynomial functions of multiple inputs

$$f(x, y) = w_5 x^2 + w_4 y^2 + w_3 xy + w_2 x + w_1 y + b$$

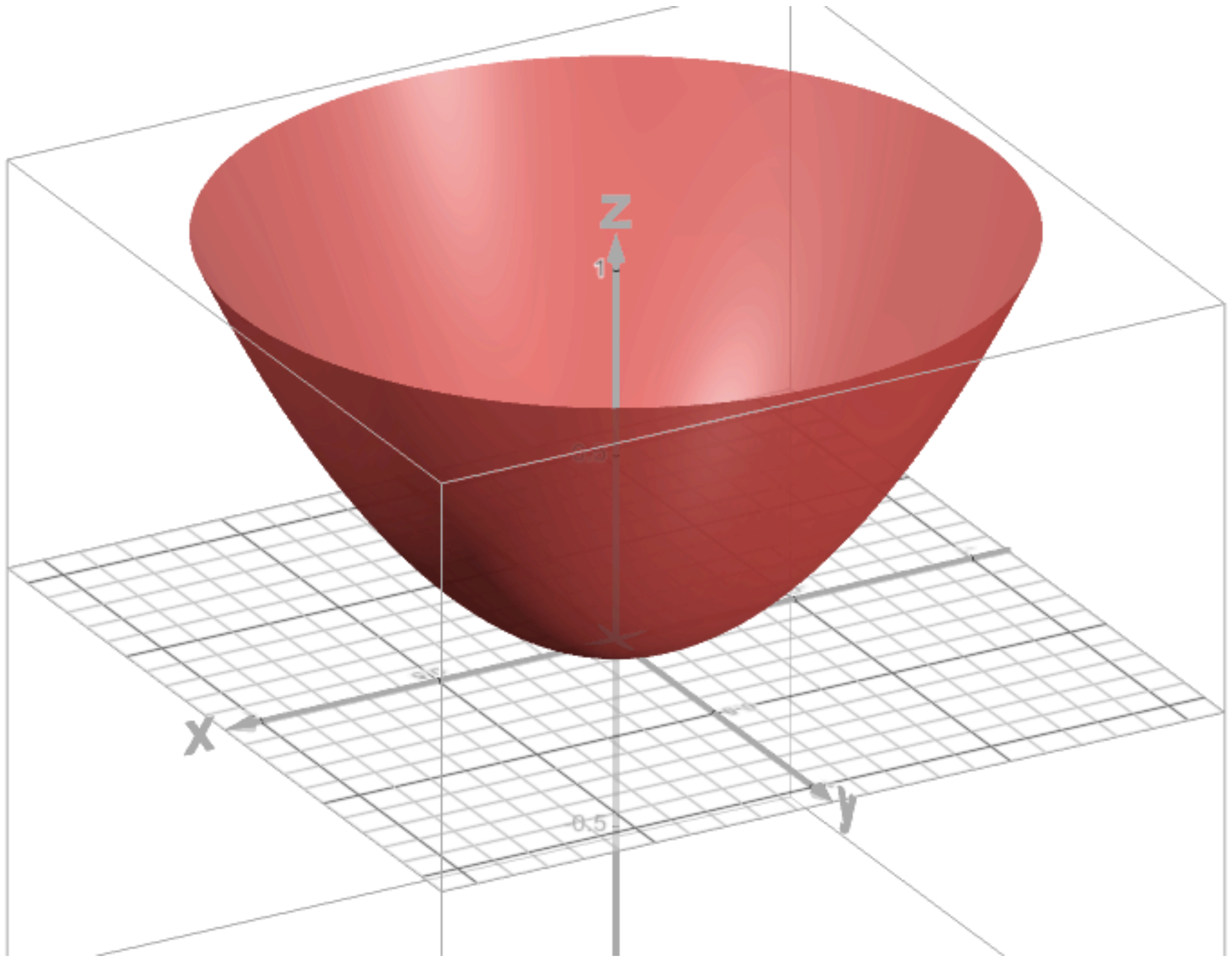
$$f(\mathbf{x}) = w_5 x_2^2 + w_4 x_1^2 + w_3 x_1 x_2 + w_2 x_2 + w_1 x_1 + b$$

Linear function of 2-inputs (plane)



Quadratic function of 2-inputs (paraboloid)

$$f(\mathbf{x}) = w_5 x_2^2 + w_4 x_1^2 + w_3 x_1 x_2 + w_2 x_2 + w_1 x_1 + b$$



Degree of a polynomial function

Largest (total) exponent in any term

$$f(x, y) = w_5 x^2 + w_4 y^2 + w_3 xy + w_2 x + w_1 y + b$$

$$f(x, y) = 3x^4 + 2xy + y - 2$$

$$f(x, y) = -2x^2 y^2 + 2x^3 + y^2 - 5$$

Polynomial functions as vector operations

$$f(\mathbf{x}) = w_5 x_2^2 + w_4 x_1^2 + w_3 x_1 x_2 + w_2 x_2 + w_1 x_1 + b$$

$$w_5 x_2^2 + w_4 x_1^2 + w_3 x_1 x_2 + w_2 x + w_1 y + b = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ b \end{bmatrix}$$

Polynomial functions as vector operations

$$w_5 x_2^2 + w_4 x_1^2 + w_3 x_1 x_2 + w_2 x + w_1 y + b = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ b \end{bmatrix}$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \xrightarrow[\phi]{} \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \\ 1 \end{bmatrix}$$

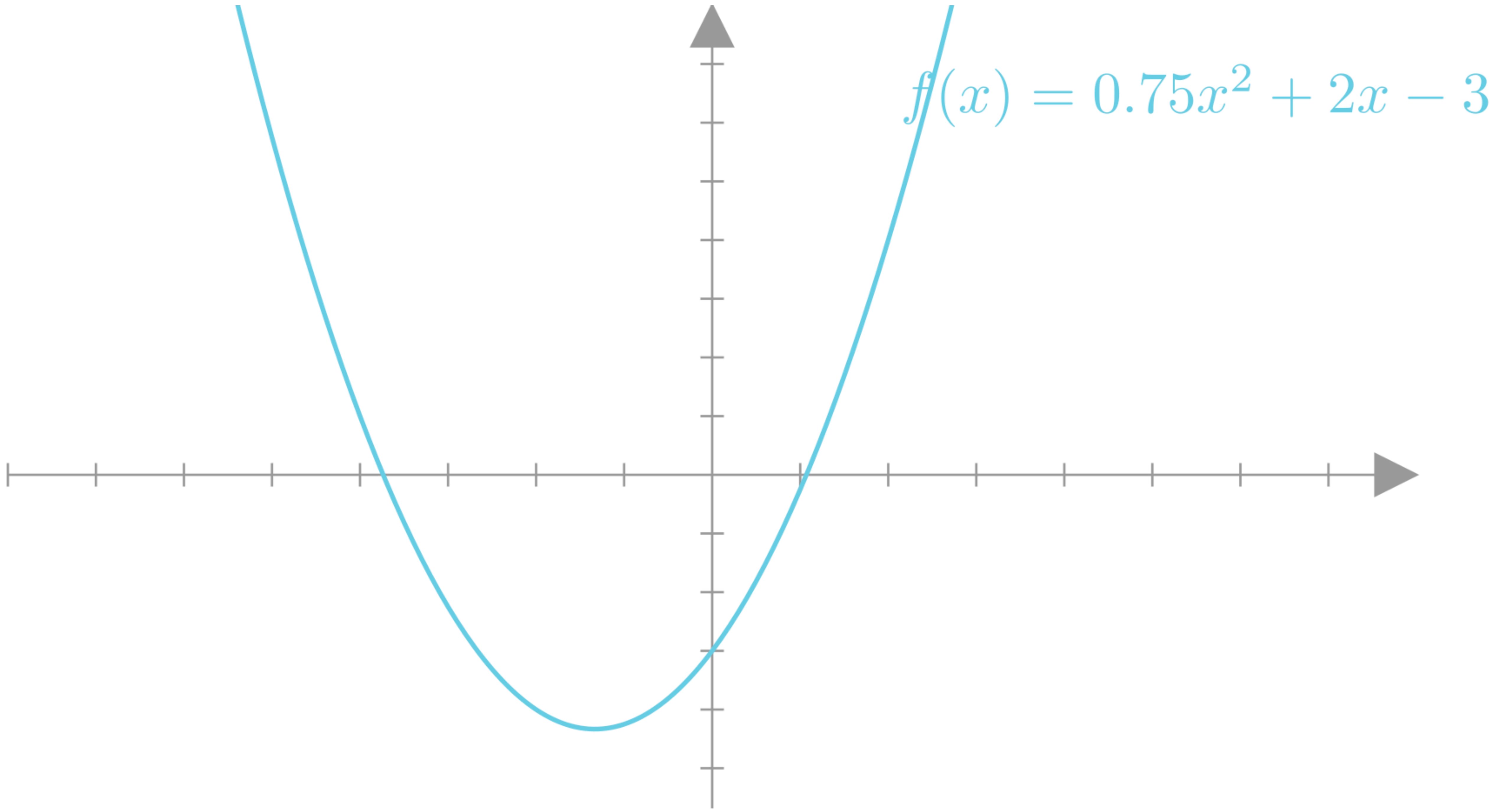
Polynomial functions as vector operations

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \xrightarrow[\phi]{} \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \\ 1 \end{bmatrix}$$

$$\phi(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \\ 1 \end{bmatrix}$$

$$f(\mathbf{x}) = \phi(\mathbf{x})^T \mathbf{w}$$

Quadratic function as a feature transform



$$f(\mathbf{x}) = \phi(\mathbf{x})^T \mathbf{w} = w_2 x_1^2 + w_1 x_1 + b, \quad \phi(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_1^2 \\ 1 \end{bmatrix}$$

Fitting quadratic regression

Prediction function

$$f(\mathbf{x}) = \phi(\mathbf{x})^T \mathbf{w}, \quad \phi(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \\ 1 \end{bmatrix}$$

Negative log-likelihood loss

$$\mathbf{Loss}(\mathbf{w}) = \mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

$$= \frac{1}{2\sigma^2} \sum_{i=1}^{N} \big(y_i - \phi(\mathbf{x}_i)^T \mathbf{w}\big)^2 + N \log \sigma\sqrt{2\pi}$$

Probabilistic model

$$y_i \sim \mathcal{N}\big(\phi(\mathbf{x}_i)^T \mathbf{w}, \sigma^2\big)$$

Optimization problem

$$\mathbf{w}^* = \underset{\mathbf{w}}{\text{argmin}} \mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y})$$

Fitting quadratic regression

Prediction function

$$f(\mathbf{x}) = \phi(\mathbf{x})^T \mathbf{w}, \quad \phi(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \\ 1 \end{bmatrix}$$

Probabilistic model

$$y_i \sim \mathcal{N}\left(\phi(\mathbf{x}_i)^T \mathbf{w}, \sigma^2\right)$$

Negative log-likelihood loss

$$\mathbf{Loss}(\mathbf{w}) = \mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

$$= \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - \phi(\mathbf{x}_i)^T \mathbf{w}\right)^2 + N \log \sigma \sqrt{2\pi}$$

What is the gradient of the log-likelihood with respect to w?

Fitting quadratic regression

Prediction function

$$f(\mathbf{x}) = \phi(\mathbf{x})^T \mathbf{w}, \quad \phi(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \\ 1 \end{bmatrix}$$

Probabilistic model

$$y_i \sim \mathcal{N}\left(\phi(\mathbf{x}_i)^T \mathbf{w}, \sigma^2\right)$$

Negative log-likelihood loss

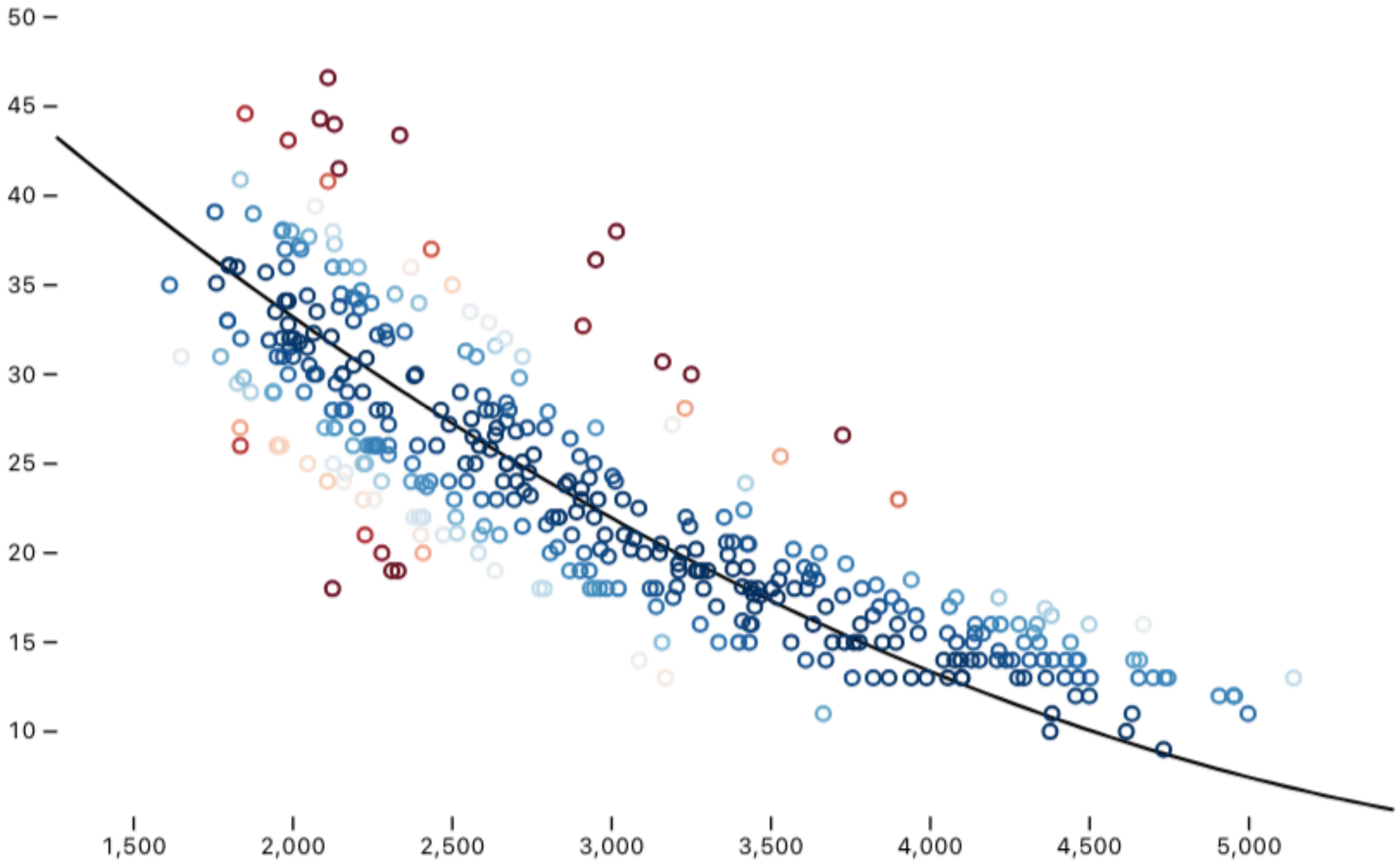
$$\textbf{Loss}(\mathbf{w}) = \textbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \mathbf{w})$$

$$= \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - \phi(\mathbf{x}_i)^T \mathbf{w}\right)^2 + N \log \sigma \sqrt{2\pi}$$

What is the gradient of the log-likelihood with respect to w?

$$\nabla_{\mathbf{w}} \textbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(\phi(\mathbf{x}_i)^T \mathbf{w} - y_i\right) \phi(\mathbf{x}_i)$$

Quadratic regression on real data



Quadratic logistic regression

Prediction function

$$f(\mathbf{x}) = \mathbb{I}(\phi(\mathbf{x})^T\mathbf{w} \geq 0), \quad \phi(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_2 \\ x_1x_2 \\ x_1^2 \\ x_2^2 \\ 1 \end{bmatrix}$$

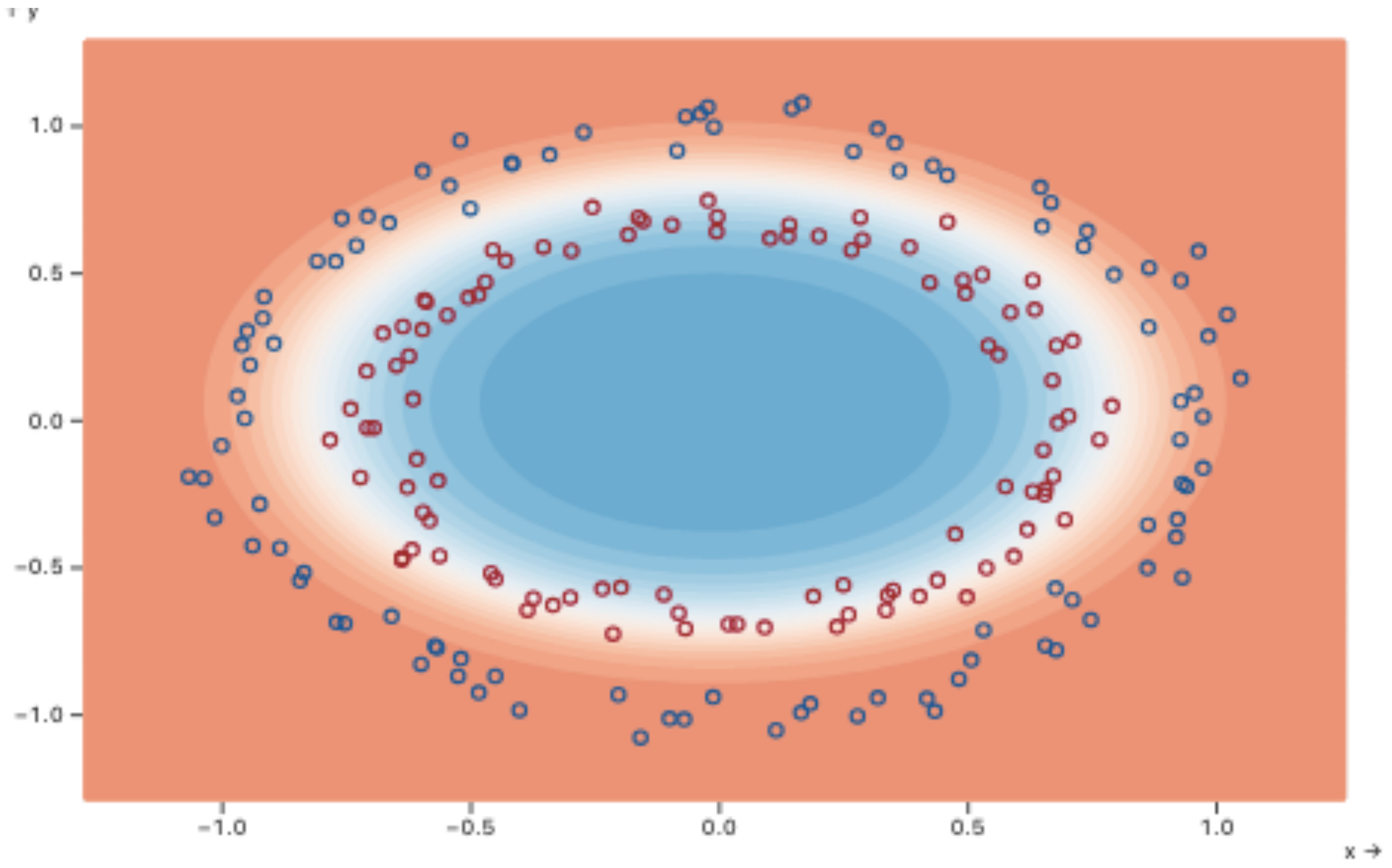
Probabilistic model

$$y_i \sim \mathbf{Bernoulli}\big(\sigma(\phi(\mathbf{x_i})^\mathbf{T}\mathbf{w})\big), \quad p(y_i = 1 \mid \mathbf{x}_i, \mathbf{w}) = \sigma\big(\phi(\mathbf{x}_i)^T\mathbf{w}\big)$$

Negative log-likelihood loss

$$\mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{N} \log \sigma\big((2y_i - 1)\phi(\mathbf{x}_i)^T\mathbf{w}\big)$$

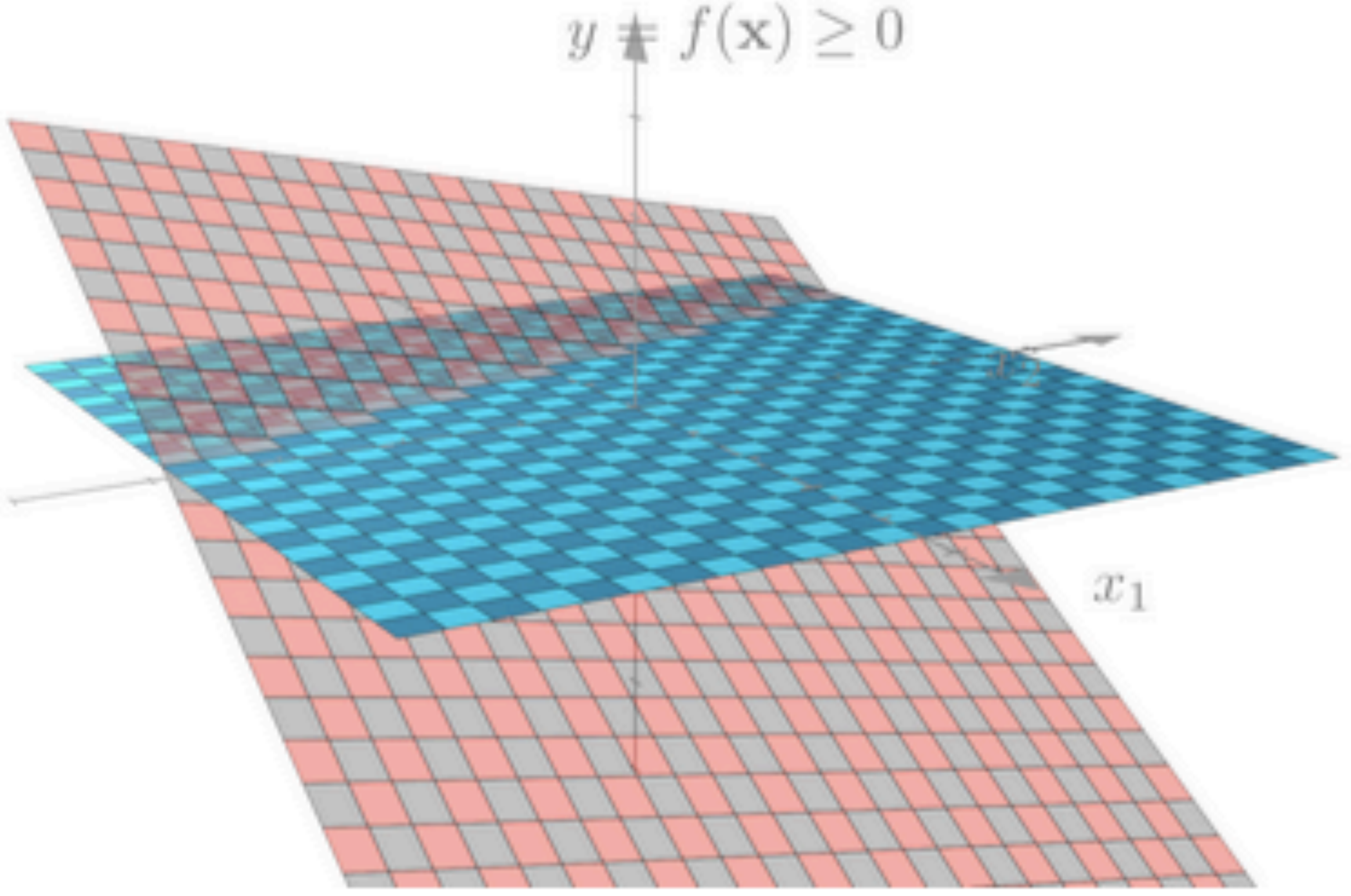
With two inputs



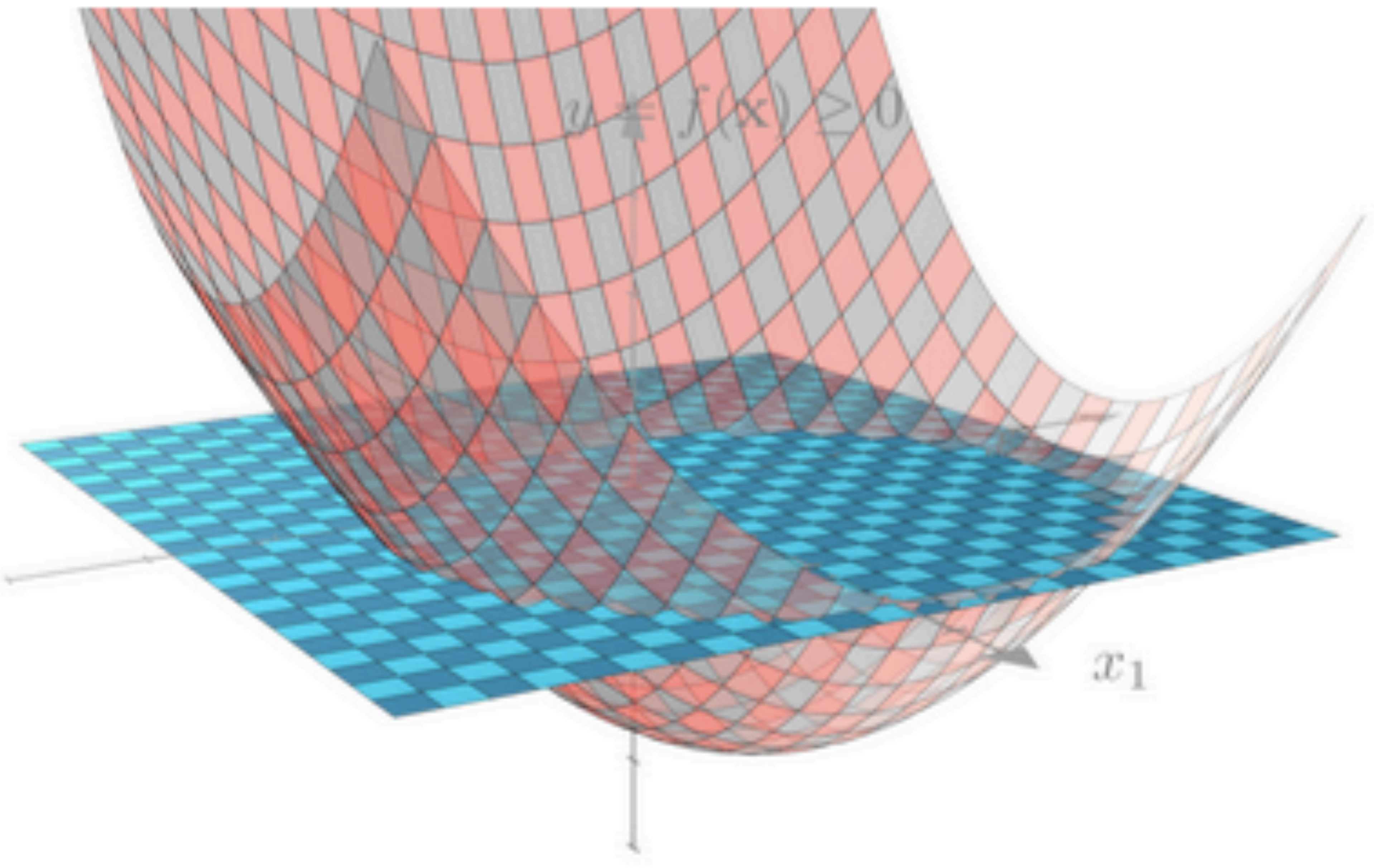
Quadratic decision boundaries

$$f(\mathbf{x}) = \mathbb{I}(\phi(\mathbf{x})^T \mathbf{w} \geq 0), \quad \phi(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \\ 1 \end{bmatrix}$$

Linear decision boundary

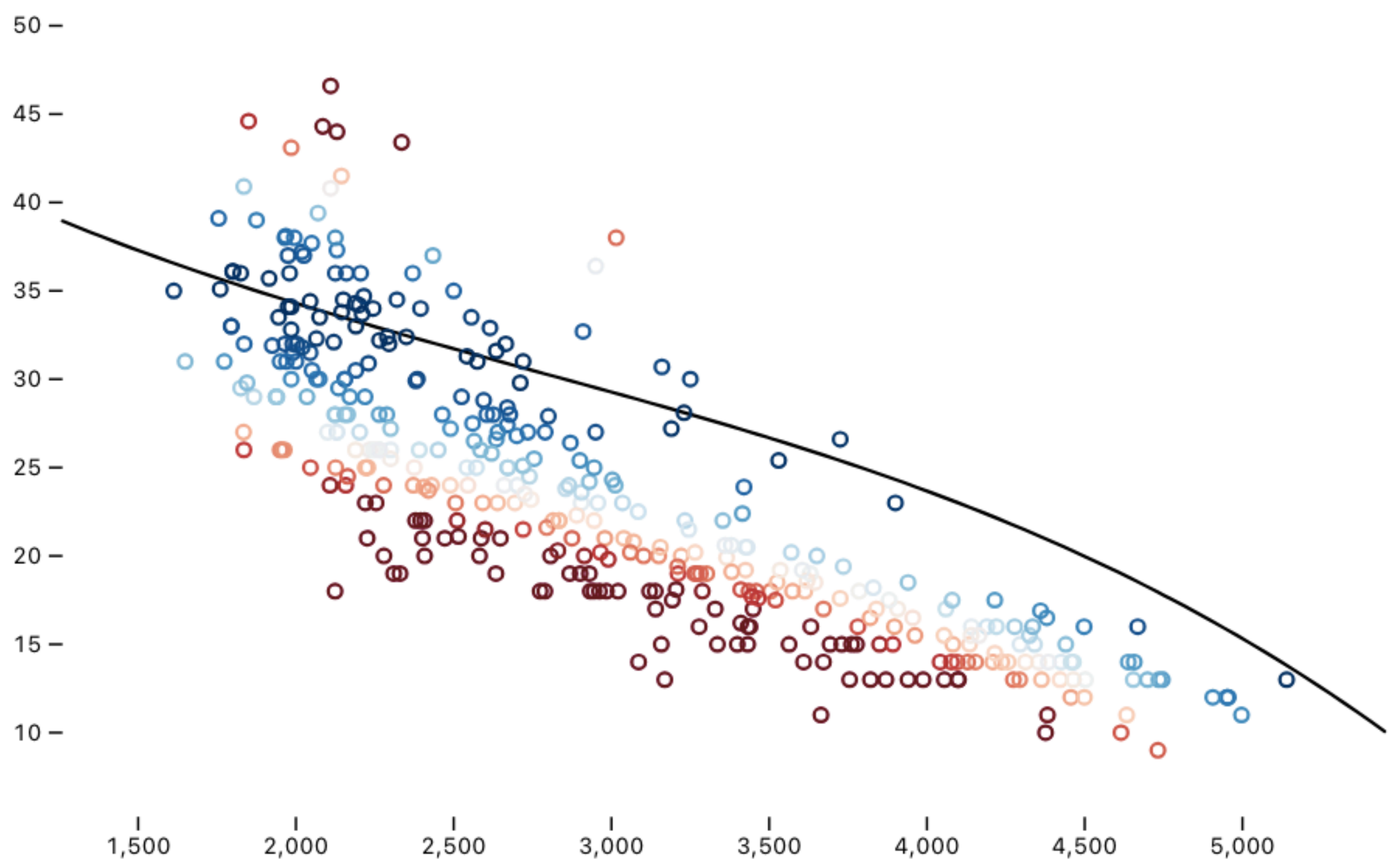


Quadratic decision boundary



Cubic feature transforms

$$f(\mathbf{x}) = \phi(\mathbf{x})^T\mathbf{w} = w_3 x_1^3 + w_2 x_1^2 + w_1 x_1 + b, \quad \phi(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_1^2 \\ x_1^3 \\ 1 \end{bmatrix}$$



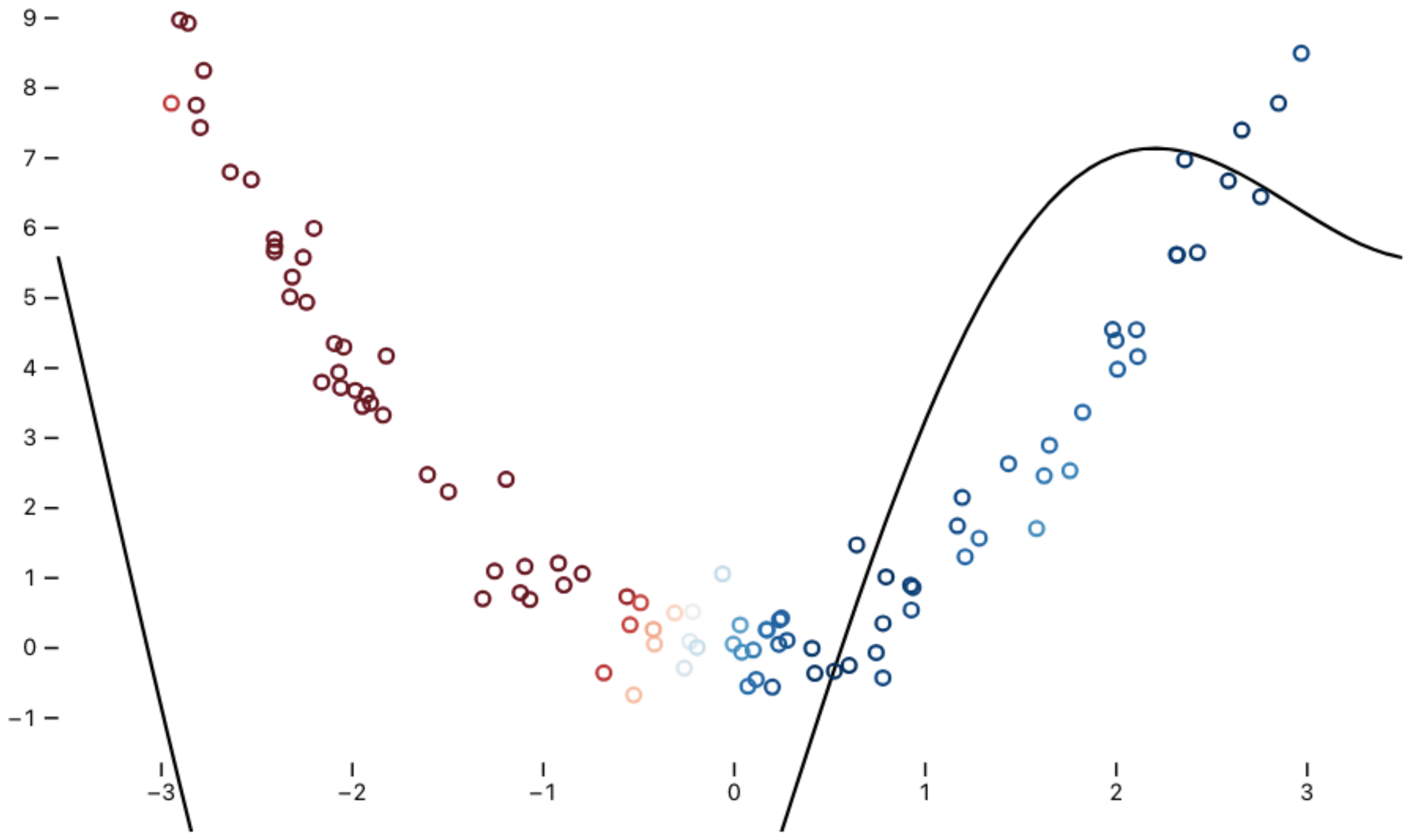
Other feature transforms

$$\phi(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_2 \\ \sin(x_1) \\ \sin(x_2) \\ \cos(x_1) \\ \cos(x_2) \\ 1 \end{bmatrix}$$

$$\phi(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_2 \\ \sigma(x_1) \\ \sigma(x_2) \\ 1 \end{bmatrix}$$

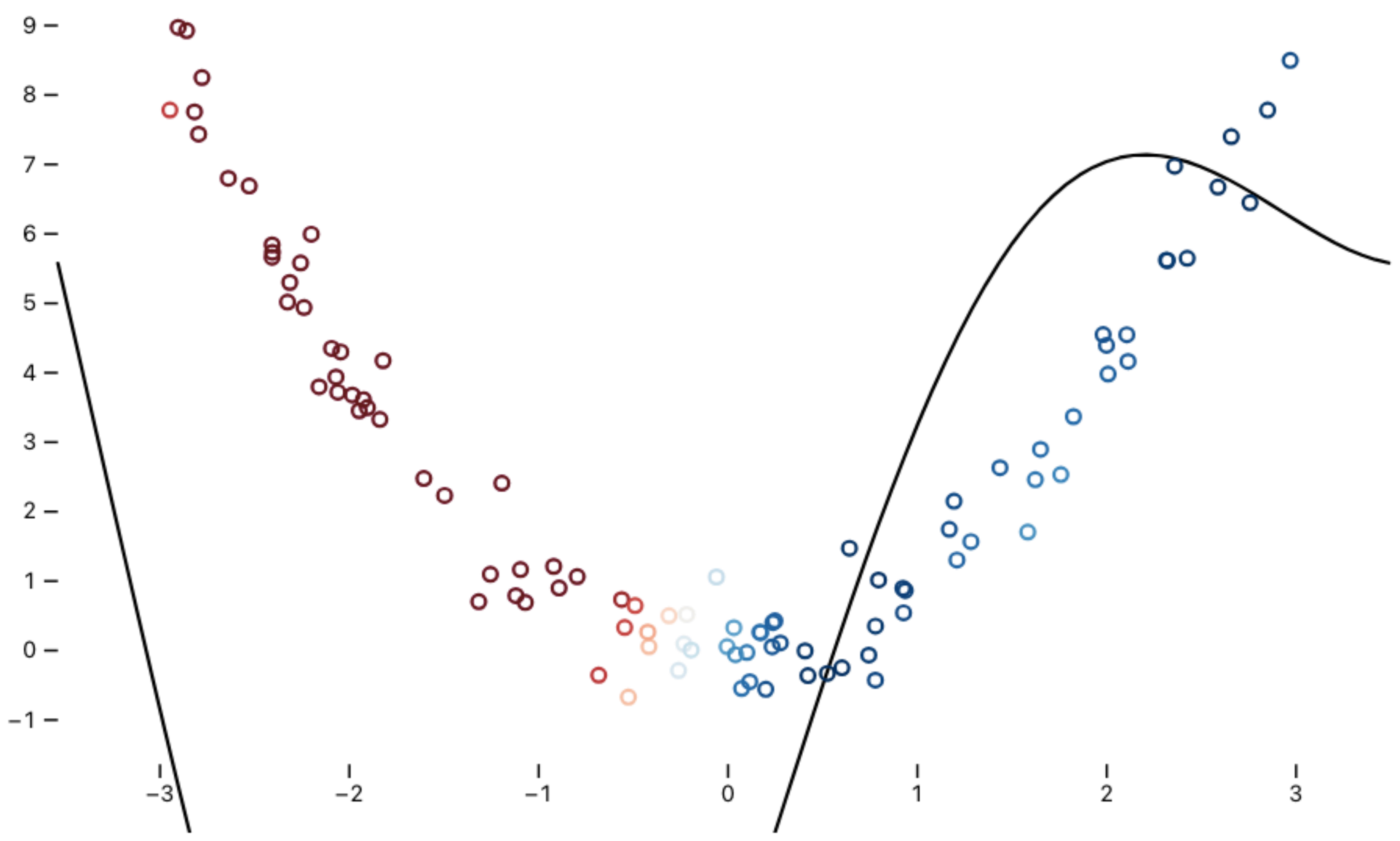
Other feature transforms

$$f(\mathbf{x}) = \phi(\mathbf{x})^T\mathbf{w} = w_3 e^{x_1} + w_2 \sin(x_1) + w_1 x_1^2 + b, \quad \phi(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_1^2 \\ \sin(x_1) \\ e^{x_1} \\ 1 \end{bmatrix}$$



Other feature transforms

$$f(\mathbf{x}) = \phi(\mathbf{x})^T\mathbf{w} = w_3 e^{x_1} + w_2 \sin(x_1) + w_1 x_1^2 + b, \quad \phi(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_1^2 \\ \sin(x_1) \\ e^{x_1} \\ 1 \end{bmatrix}$$



Back to our viz