# Classification, Regression, Error functions and Optimization

**Neil D. Lawrence**

Department of Computer Science, University of Sheffield, U.K.

11th October 2013

# Outline

# Classification

- We are given data set containing "inputs", $\mathbf{X}$, and "targets", $\mathbf{y}$.
- Each data point consists of an input vector $\mathbf{x}_{i,:}$ and a class label, $y_i$.
- For binary classification assume $y_i$ should be either 1 (yes) or $-1$ (no).
- Input vector can be thought of as features.

# Classification Examples

- Classifying hand written digits from binary images (automatic zip code reading).
- Detecting faces in images (e.g. digital cameras).
- Who a detected face belongs to (e.g. Picasa).
- Classifying type of cancer given gene expression data.
- Categorization of document types (different types of news article on the internet).

# The Perceptron

- Developed in 1957 by Rosenblatt.
- Take a data point at, $\mathbf{x}_i$.
- Predict it belongs to a class, $y_i = 1$ if $\sum_j w_j \mathbf{x}_{i,j} + b > 0$ i.e. $\mathbf{w}^\top \mathbf{x}_i + b > 0$. Otherwise assume $y_i = -1$.

# Perceptron-like Algorithm

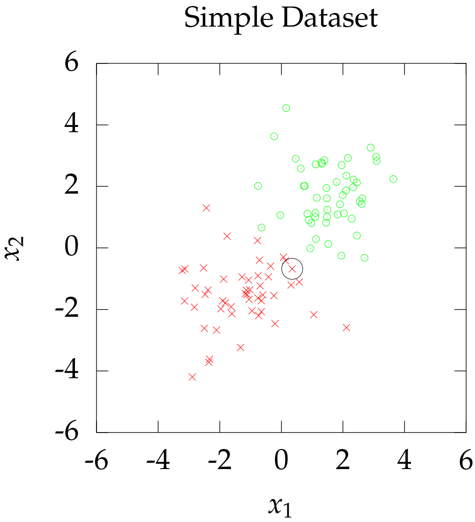
1. Select a random data point $i$.
2. Ensure $i$ is correctly classified by setting $\mathbf{w} = y_i \mathbf{x}_i$.
   - i.e. $\text{sign}\left(\mathbf{w}^\top \mathbf{x}_{i,:}\right) = \text{sign}\left(y_i \mathbf{x}_{i,:}^\top \mathbf{x}_{i,:}\right) = \text{sign}\left(y_i\right) = y_i$

# Perceptron Iteration

1. Select a misclassified point, $i$.
2. Set $\mathbf{w} \leftarrow \mathbf{w} + \eta y_i \mathbf{x}_{i,:}$.
   - If $\eta$ is large enough this will guarantee this point becomes correctly classified.
3. Repeat until there are no misclassified points.

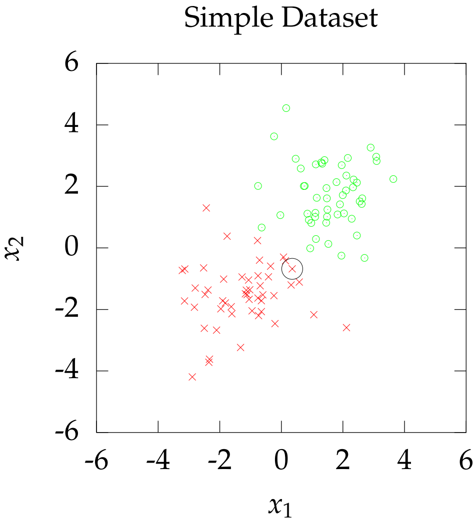
# Perceptron Algorithm

- Iteration 1 data no 29



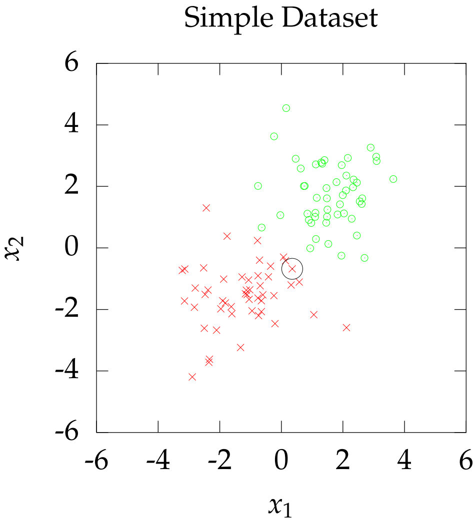
# Perceptron Algorithm

- Iteration 1 data no 29
- $w_1 = 0, w_2 = 0$



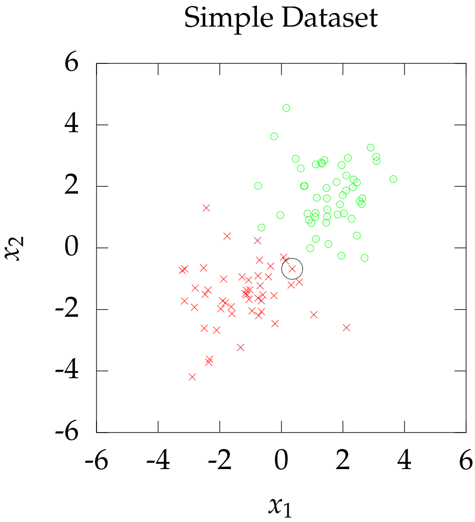
# Perceptron Algorithm

- Iteration 1 data no 29
- $w_1 = 0, w_2 = 0$
- First Iteration



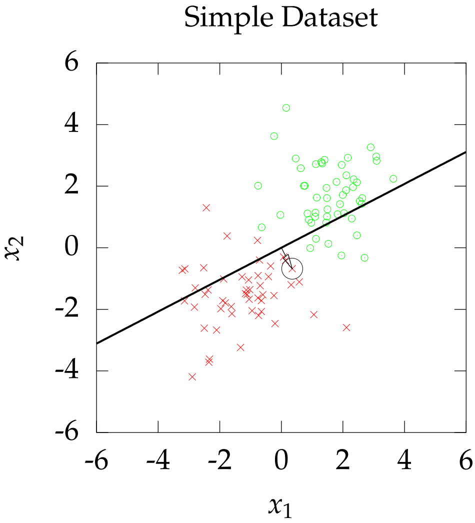
# Perceptron Algorithm

- Iteration 1 data no 29
- $w_1 = 0, w_2 = 0$
- First Iteration
- Set weight vector to data point.



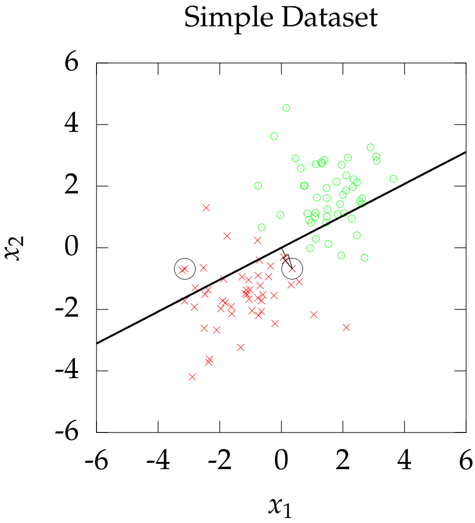
# Perceptron Algorithm

- Iteration 1 data no 29
- $w_1 = 0, w_2 = 0$
- First Iteration
- Set weight vector to data point.
- $\mathbf{w} = y_{29}\mathbf{x}_{29,:}$



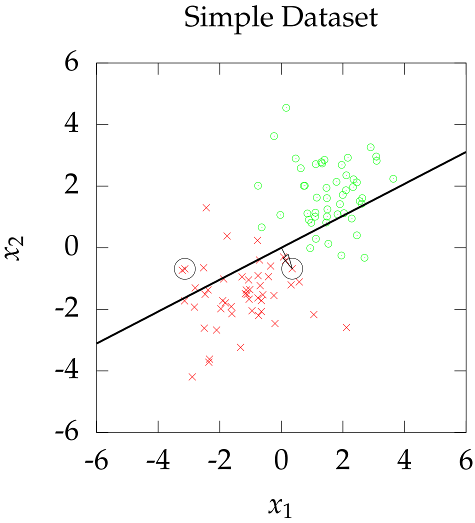
# Perceptron Algorithm

- Iteration 1 data no 29
- $w_1 = 0, w_2 = 0$
- First Iteration
- Set weight vector to data point.
- $\mathbf{w} = y_{29}\mathbf{x}_{29,:}$
- Select new incorrectly classified data point.



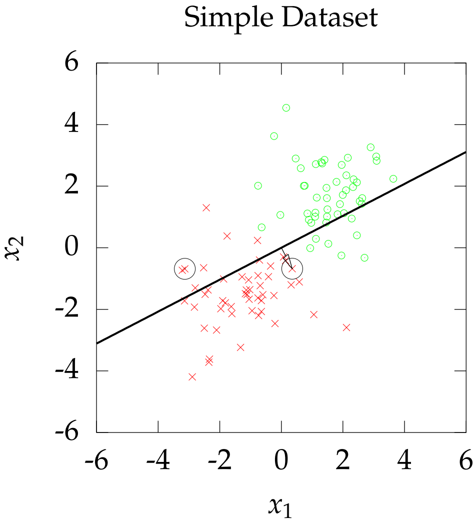
# Perceptron Algorithm

- Iteration 2 data no 16



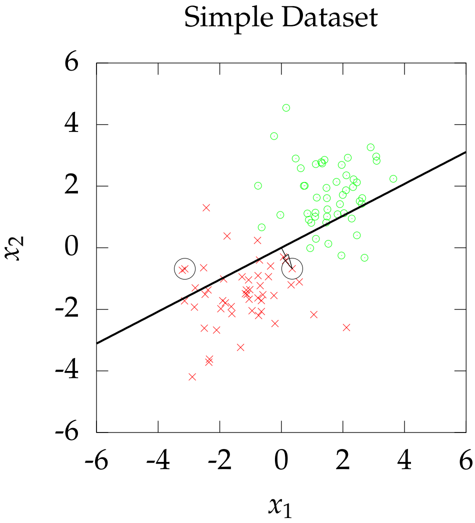
# Perceptron Algorithm

- Iteration 2 data no 16
- $w_1 = 0.3519$,
  $w_2 = -0.6787$



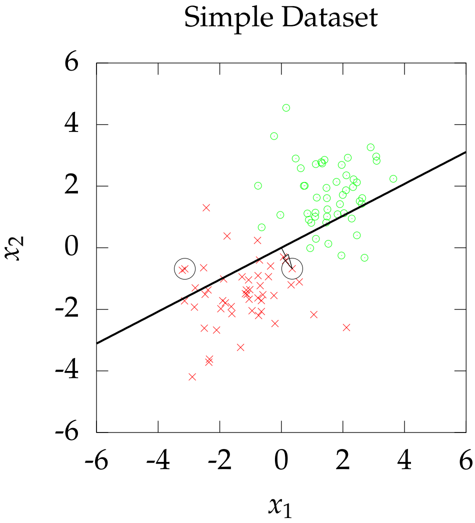
# Perceptron Algorithm

- Iteration 2 data no 16
- $w_1 = 0.3519$,
  $w_2 = -0.6787$
- Incorrect classification



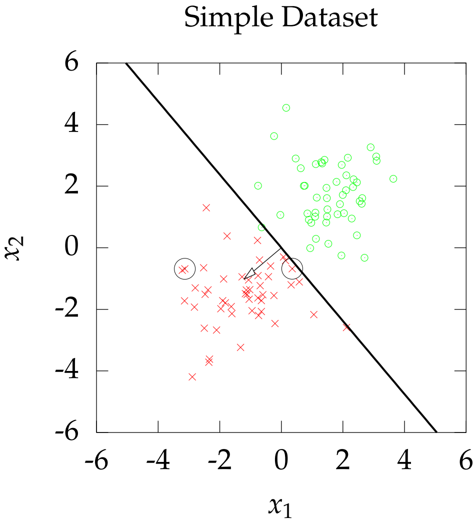
# Perceptron Algorithm

- Iteration 2 data no 16
- $w_1 = 0.3519$,
  $w_2 = -0.6787$
- Incorrect classification
- Adjust weight vector with new data point.



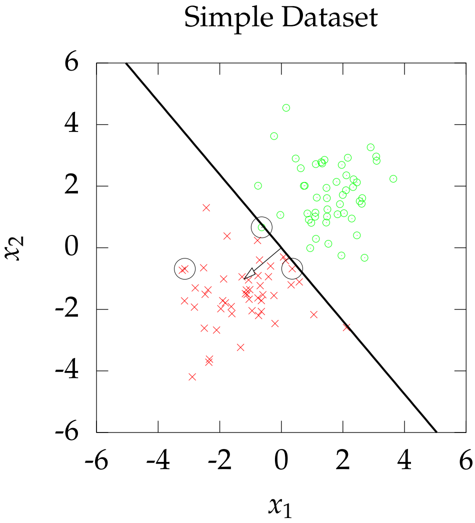
# Perceptron Algorithm

- Iteration 2 data no 16
- $w_1 = 0.3519$,
  $w_2 = -0.6787$
- Incorrect classification
- Adjust weight vector with new data point.
- $\mathbf{w} \leftarrow \mathbf{w} + \eta y_{16} \mathbf{x}_{16,:}$



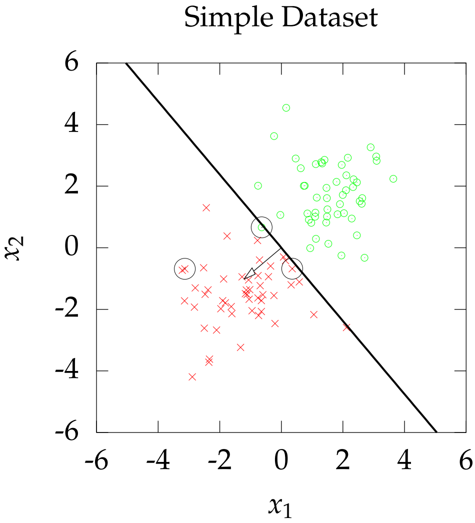
# Perceptron Algorithm

- Iteration 2 data no 16
- $w_1 = 0.3519$,
  $w_2 = -0.6787$
- Incorrect classification
- Adjust weight vector with new data point.
- $\mathbf{w} \leftarrow \mathbf{w} + \eta y_{16} \mathbf{x}_{16,:}$
- Select new incorrectly classified data point.



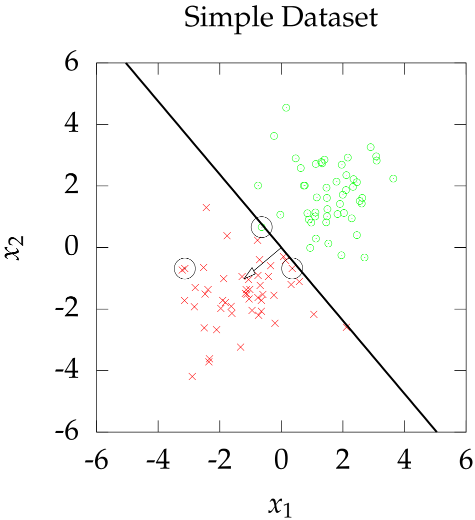
# Perceptron Algorithm

- Iteration 3 data no 58



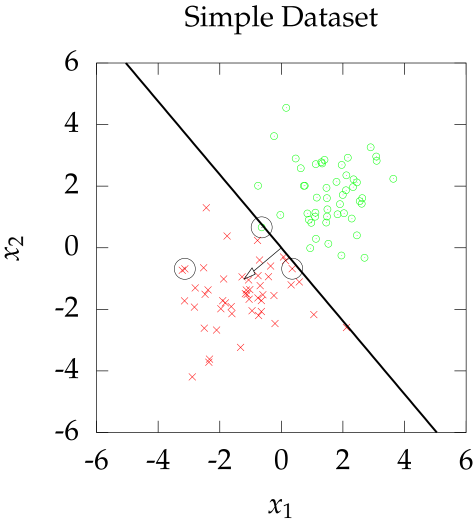
# Perceptron Algorithm

- Iteration 3 data no 58
- $w_1 = -1.2143$,
  $w_2 = -1.0217$



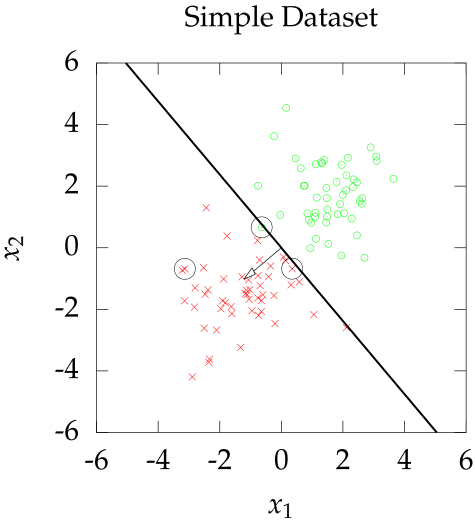
# Perceptron Algorithm

- Iteration 3 data no 58
- $w_1 = -1.2143$,
  $w_2 = -1.0217$
- Incorrect classification



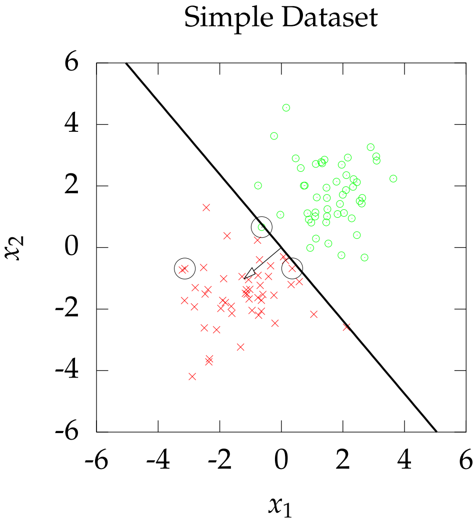
# Perceptron Algorithm

- Iteration 3 data no 58
- $w_1 = -1.2143$, $w_2 = -1.0217$
- Incorrect classification
- Adjust weight vector with new data point.



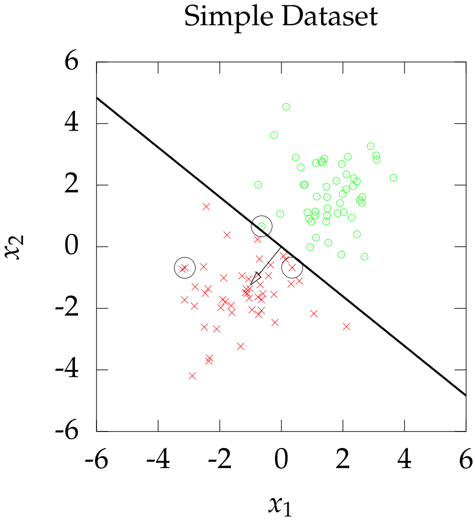
# Perceptron Algorithm

- Iteration 3 data no 58
- $w_1 = -1.2143$,
  $w_2 = -1.0217$
- Incorrect classification
- Adjust weight vector with new data point.
- $\mathbf{w} \leftarrow \mathbf{w} + \eta y_{58} \mathbf{x}_{58,:}$



# Perceptron Algorithm

- Iteration 3 data no 58
- $w_1 = -1.2143$,
  $w_2 = -1.0217$
- Incorrect classification
- Adjust weight vector with new data point.
- $\mathbf{w} \leftarrow \mathbf{w} + \eta y_{58} \mathbf{x}_{58,:}$
- All data correctly classified.



# Regression Examples

- Predict a real value, $y_i$ given some inputs $\mathbf{x}_i$.
- Predict quality of meat given spectral measurements (Tecator data).
- Radiocarbon dating, the C14 calibration curve: predict age given quantity of C14 isotope.
- Predict quality of different Go or Backgammon moves given expert rated training data.

# Linear Regression

## Is there an equivalent learning rule for regression?

- Predict a real value $y$ given $x$.
- We can also construct a learning rule for regression.
  - Define our prediction

$$f(x) = mx + c.$$

  - Define an error

$$\Delta y_i = y_i - f(x_i).$$

# Updating Bias/Intercept

- $c$ represents bias. Add portion of error to bias.

$$c \longrightarrow c + \eta \Delta y_i.$$

$$\Delta y_i = y_i - m x_i - c.$$

1. For +ve error, $c$ and therefore $f(x_i)$ become larger and error magnitude becomes smaller.
2. For -ve error, $c$ and therefore $f(x_i)$ become smaller and error magnitude becomes smaller.

# Updating Slope

- $m$ represents Slope. Add portion of error $\times$ input to slope.

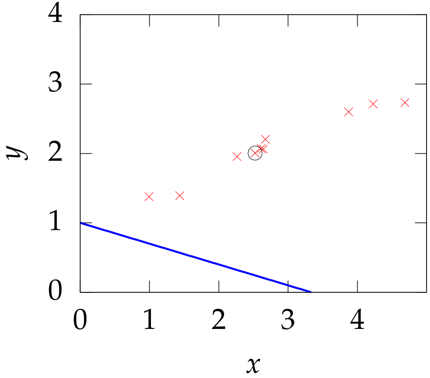
$$m \to m + \eta \Delta y_i x_i.$$

$$\Delta y_i = y_i - m x_i - c.$$

1. For +ve error and +ve input, $m$ becomes larger and $f(x_i)$ becomes larger: error magnitude becomes smaller.
2. For +ve error and -ve input, $m$ becomes smaller and $f(x_i)$ becomes larger: error magnitude becomes smaller.
3. For -ve error and -ve slope, $m$ becomes larger and $f(x_i)$ becomes smaller: error magnitude becomes smaller.
4. For -ve error and +ve input, $m$ becomes smaller and $f(x_i)$ becomes smaller: error magnitude becomes smaller.

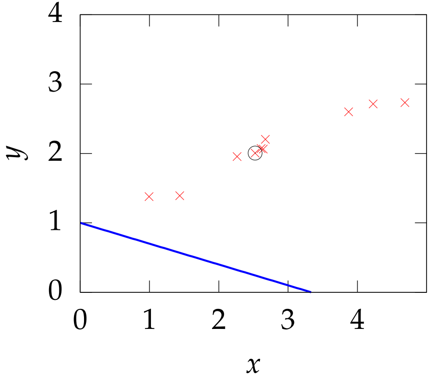
# Linear Regression Example

- Iteration 1 $\hat{m} = -0.3$
  $\hat{c} = 1$



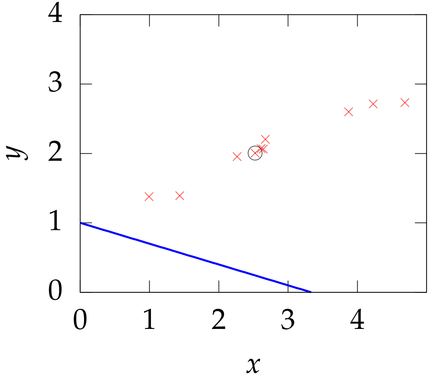
# Linear Regression Example

- Iteration 1 $\hat{m} = -0.3$ $\hat{c} = 1$
  - Present data point 4



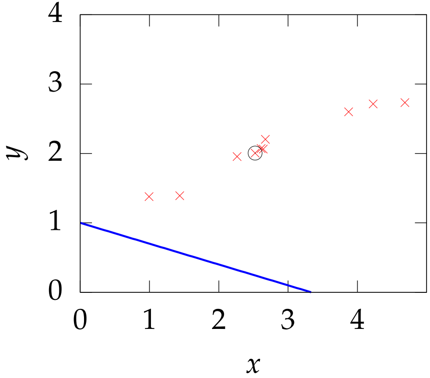
# Linear Regression Example

- Iteration 1 $\hat{m} = -0.3$ $\hat{c} = 1$
  - Present data point 4
  - $\Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c})$



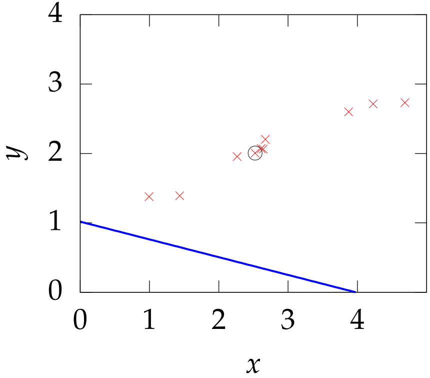
# Linear Regression Example

- Iteration 1 $\hat{m} = -0.3$ $\hat{c} = 1$
  - Present data point 4
  - $\Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_4 \Delta y_4$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_4$



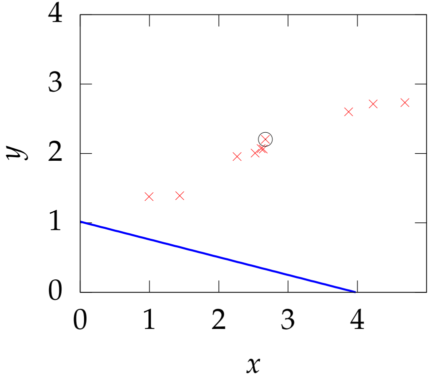
# Linear Regression Example

- Iteration 1 $\hat{m} = -0.3$ $\hat{c} = 1$
  - Present data point 4
  - $\Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_4 \Delta y_4$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_4$
- Updated values
  $\hat{m} = -0.25593$ $\hat{c} = 1.0175$



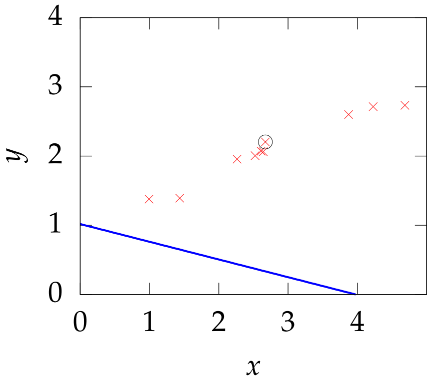
# Linear Regression Example

- Iteration 2 $\hat{m} = -0.25593$ $\hat{c} = 1.0175$



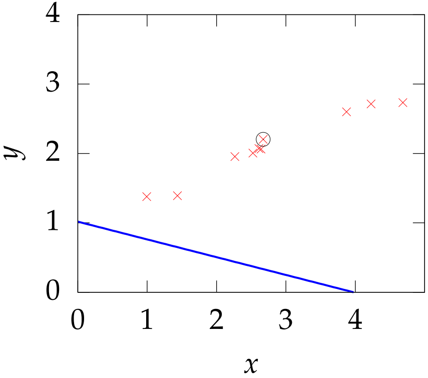
# Linear Regression Example

- Iteration 2 $\hat{m} = -0.25593$ $\hat{c} = 1.0175$
  - Present data point 7



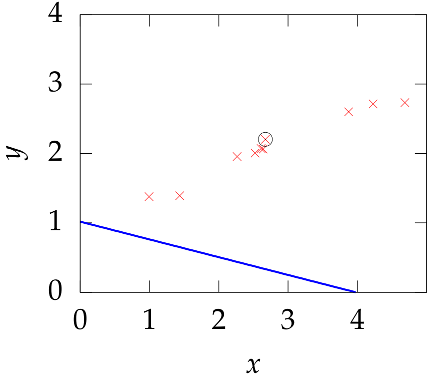
# Linear Regression Example

- Iteration 2 $\hat{m} = -0.25593$ $\hat{c} = 1.0175$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$



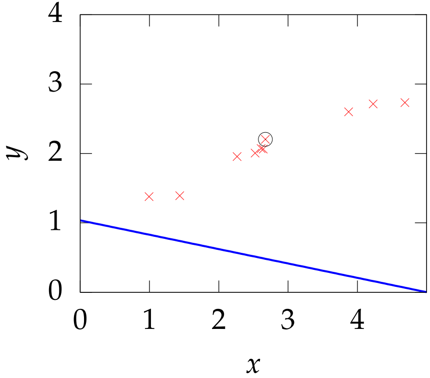
# Linear Regression Example

- Iteration 2 $\hat{m} = -0.25593$ $\hat{c} = 1.0175$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_7$



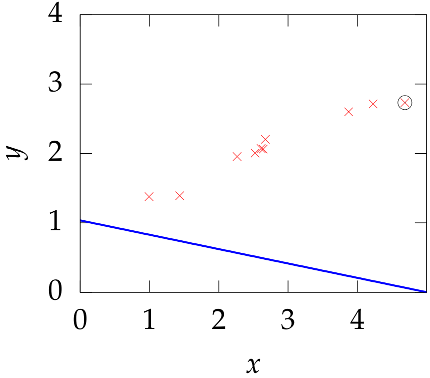
# Linear Regression Example

- Iteration 2 $\hat{m} = -0.25593$ $\hat{c} = 1.0175$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_7$
- Updated values
  $\hat{m} = -0.20693$ $\hat{c} = 1.0358$



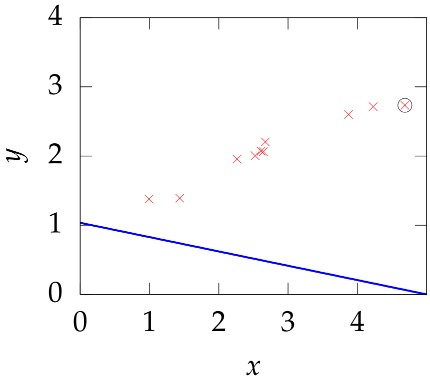
# Linear Regression Example

- Iteration 3 $\hat{m} = -0.20693$ $\hat{c} = 1.0358$



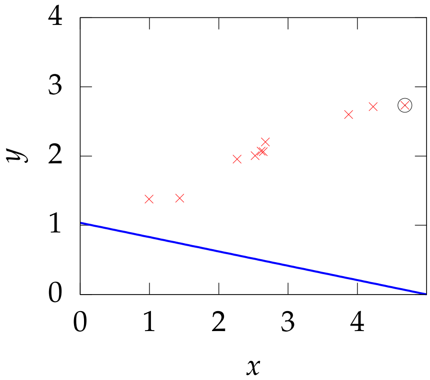
# Linear Regression Example

- Iteration 3 $\hat{m} = -0.20693$ $\hat{c} = 1.0358$
  - Present data point 10



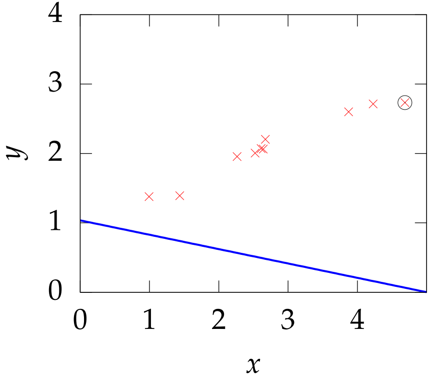
# Linear Regression Example

- Iteration 3 $\hat{m} = -0.20693$ $\hat{c} = 1.0358$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$



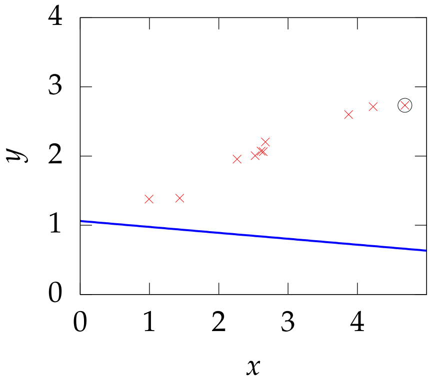
# Linear Regression Example

- Iteration 3 $\hat{m} = -0.20693$ $\hat{c} = 1.0358$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$



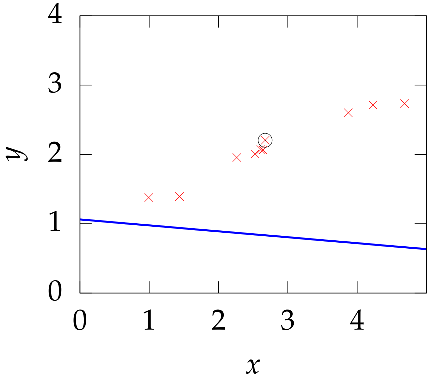
# Linear Regression Example

- Iteration 3 $\hat{m} = -0.20693$ $\hat{c} = 1.0358$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$
- Updated values
  $\hat{m} = -0.085591$ $\hat{c} = 1.0617$



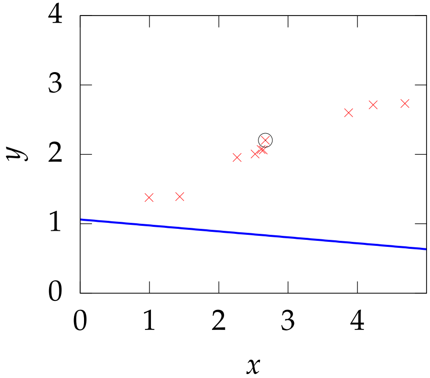
# Linear Regression Example

- Iteration 4
  $\hat{m} = -0.085591$
  $\hat{c} = 1.0617$



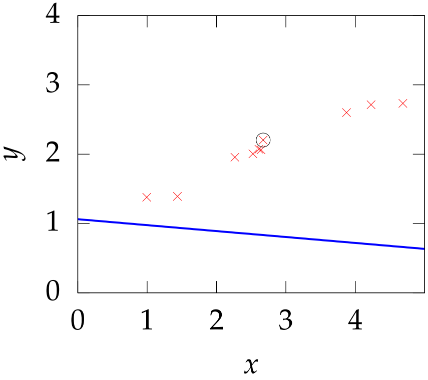
# Linear Regression Example

- Iteration 4
  $\hat{m} = -0.085591$
  $\hat{c} = 1.0617$
  - Present data point 7



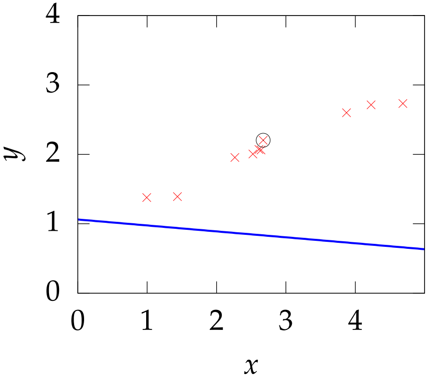
# Linear Regression Example

- Iteration 4
  $\hat{m} = -0.085591$
  $\hat{c} = 1.0617$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$



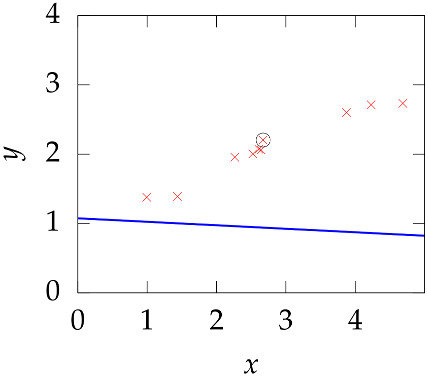
# Linear Regression Example

- Iteration 4
  $\hat{m} = -0.085591$
  $\hat{c} = 1.0617$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_7$



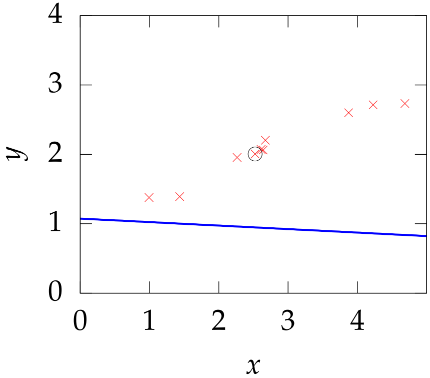
# Linear Regression Example

- Iteration 4
  $\hat{m} = -0.085591$
  $\hat{c} = 1.0617$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_7$
- Updated values
  $\hat{m} = -0.050355$ $\hat{c} = 1.0749$



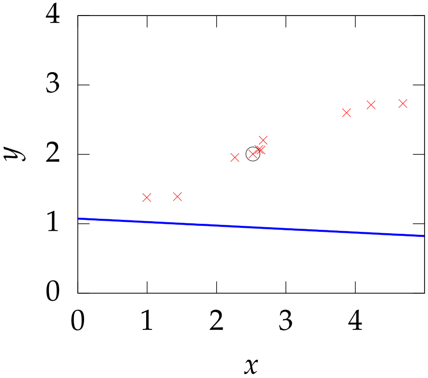
# Linear Regression Example

- Iteration 5
  $\hat{m} = -0.050355$
  $\hat{c} = 1.0749$



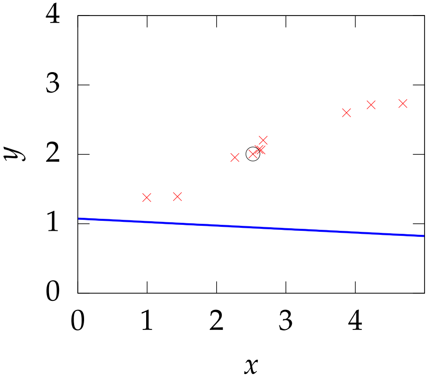
# Linear Regression Example

- Iteration 5
  $\hat{m} = -0.050355$
  $\hat{c} = 1.0749$
  - Present data point 4



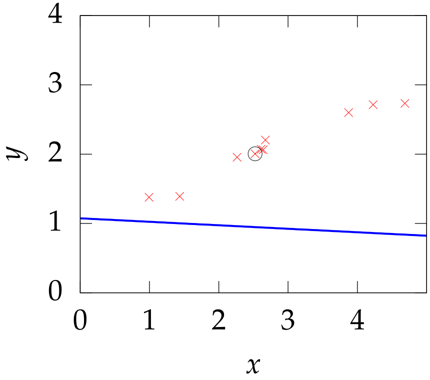
# Linear Regression Example

- Iteration 5
  $\hat{m} = -0.050355$
  $\hat{c} = 1.0749$
  - Present data point 4
  - $\Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c})$



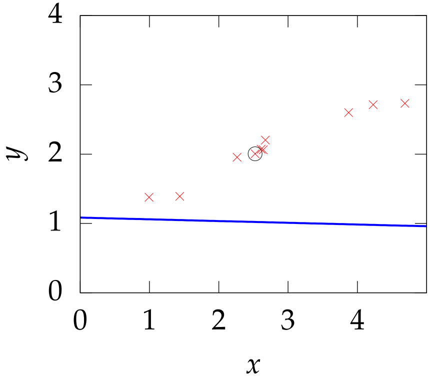
# Linear Regression Example

- Iteration 5
  $\hat{m} = -0.050355$
  $\hat{c} = 1.0749$
  - Present data point 4
  - $\Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_4 \Delta y_4$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_4$



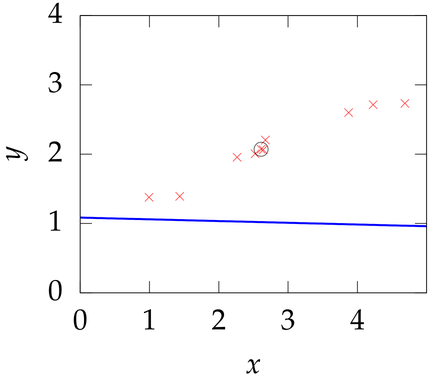
# Linear Regression Example

- Iteration 5
  $\hat{m} = -0.050355$
  $\hat{c} = 1.0749$
  - Present data point 4
  - $\Delta y_4 = (y_4 - \hat{m}x_4 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_4 \Delta y_4$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_4$
- Updated values
  $\hat{m} = -0.024925$ $\hat{c} = 1.0849$



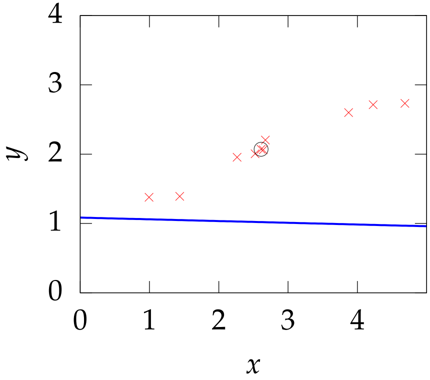
# Linear Regression Example

- Iteration 6
  $\hat{m} = -0.024925$
  $\hat{c} = 1.0849$



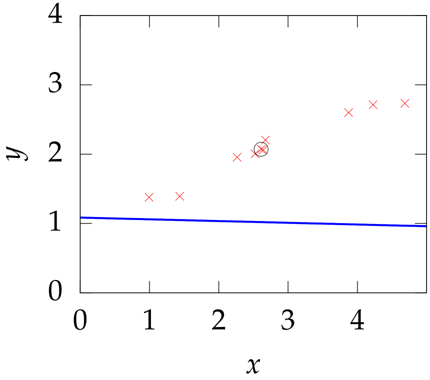
# Linear Regression Example

- Iteration 6
  $\hat{m} = -0.024925$
  $\hat{c} = 1.0849$
  - Present data point 5



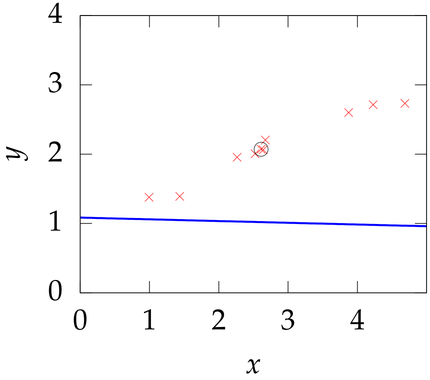
# Linear Regression Example

- Iteration 6
  $\hat{m} = -0.024925$
  $\hat{c} = 1.0849$
  - Present data point 5
  - $\Delta y_5 = (y_5 - \hat{m}x_5 - \hat{c})$



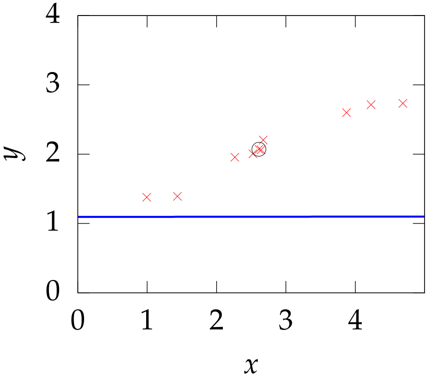
# Linear Regression Example

- Iteration 6
  $\hat{m} = -0.024925$
  $\hat{c} = 1.0849$
  - Present data point 5
  - $\Delta y_5 = (y_5 - \hat{m}x_5 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_5 \Delta y_5$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_5$



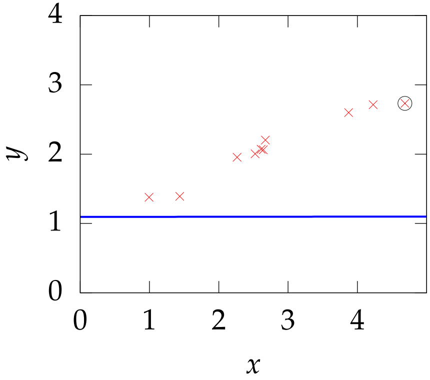
# Linear Regression Example

- Iteration 6
  $\hat{m} = -0.024925$
  $\hat{c} = 1.0849$
  - Present data point 5
  - $\Delta y_5 = (y_5 - \hat{m}x_5 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_5 \Delta y_5$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_5$
- Updated values
  $\hat{m} = 0.00098511$ $\hat{c} = 1.0949$



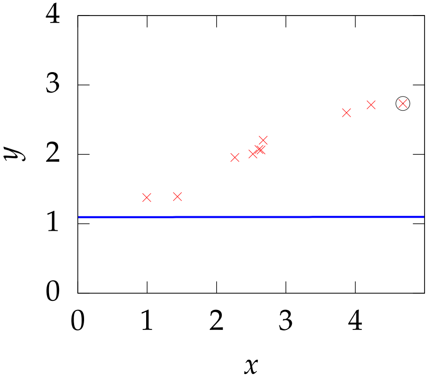
# Linear Regression Example

- Iteration 7
  $\hat{m} = 0.00098511$
  $\hat{c} = 1.0949$



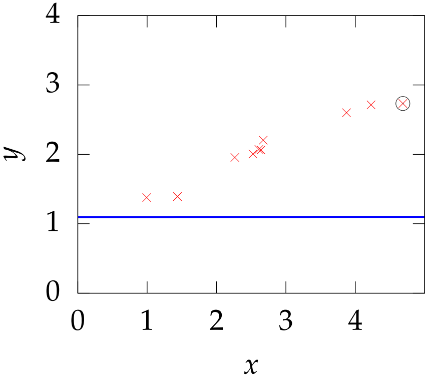
# Linear Regression Example

- Iteration 7
  $\hat{m} = 0.00098511$
  $\hat{c} = 1.0949$
  - Present data point 10



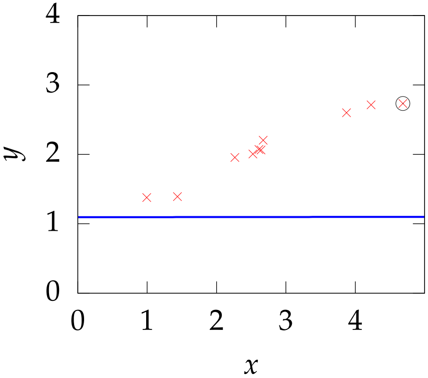
# Linear Regression Example

- Iteration 7
  $\hat{m} = 0.00098511$
  $\hat{c} = 1.0949$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$



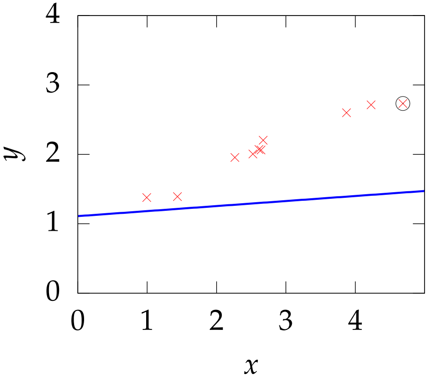
# Linear Regression Example

- Iteration 7
  $\hat{m} = 0.00098511$
  $\hat{c} = 1.0949$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$



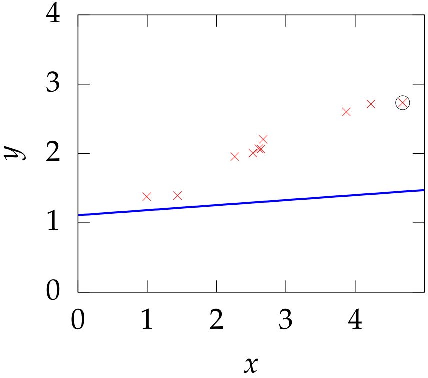
# Linear Regression Example

- Iteration 7
  $\hat{m} = 0.00098511$
  $\hat{c} = 1.0949$
    - Present data point 10
    - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
    - Adjust $\hat{m}$ and $\hat{c}$
      $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
      $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$
- Updated values
  $\hat{m} = 0.072529$ $\hat{c} = 1.1101$



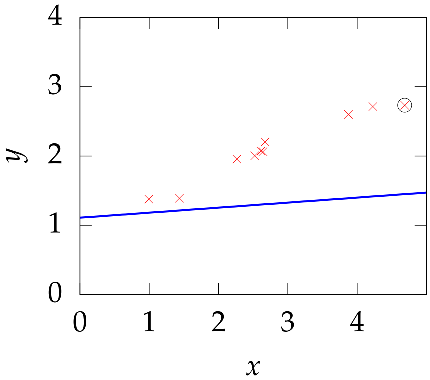
# Linear Regression Example

- Iteration 8 $\hat{m} = 0.072529$ $\hat{c} = 1.1101$



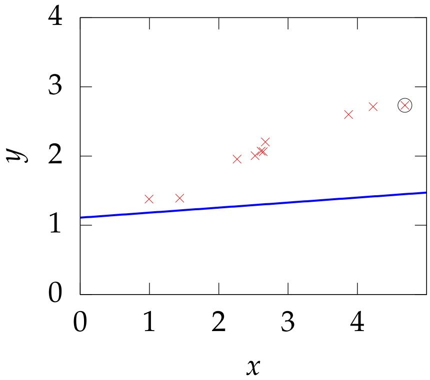
# Linear Regression Example

- Iteration 8 $\hat{m} = 0.072529$ $\hat{c} = 1.1101$
  - Present data point 10



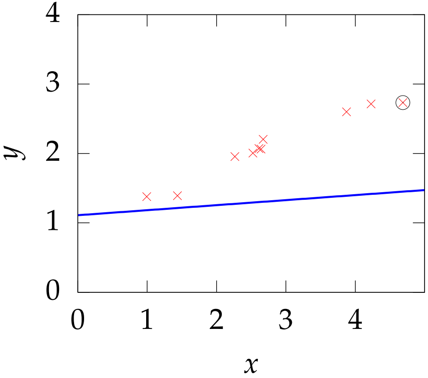
# Linear Regression Example

- Iteration 8 $\hat{m} = 0.072529$ $\hat{c} = 1.1101$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$



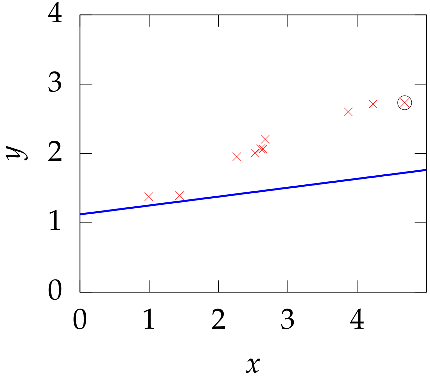
# Linear Regression Example

- Iteration 8 $\hat{m} = 0.072529$ $\hat{c} = 1.1101$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$



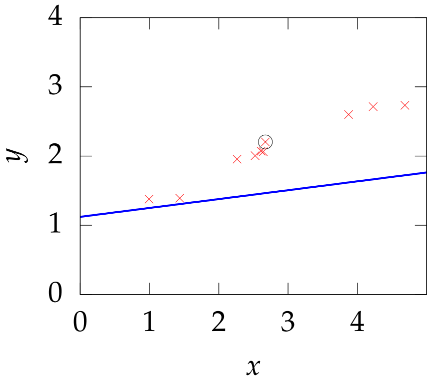
# Linear Regression Example

- Iteration 8 $\hat{m} = 0.072529$ $\hat{c} = 1.1101$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$
- Updated values
  $\hat{m} = 0.1282$ $\hat{c} = 1.122$



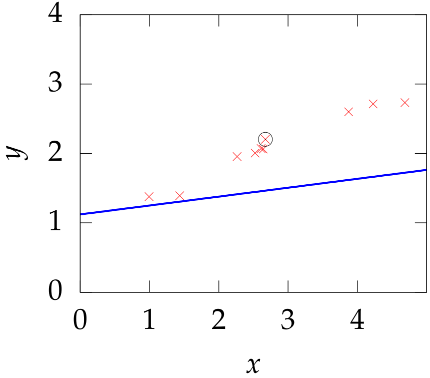
# Linear Regression Example

- Iteration 9 $\hat{m} = 0.1282$ $\hat{c} = 1.122$



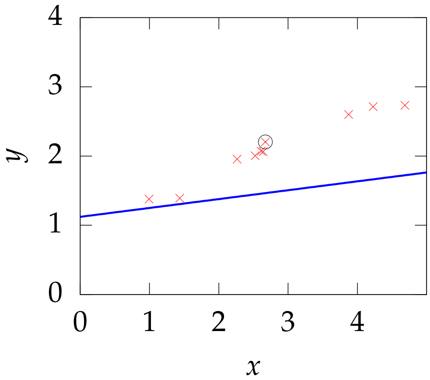
# Linear Regression Example

- Iteration 9 $\hat{m} = 0.1282$ $\hat{c} = 1.122$
  - Present data point 7



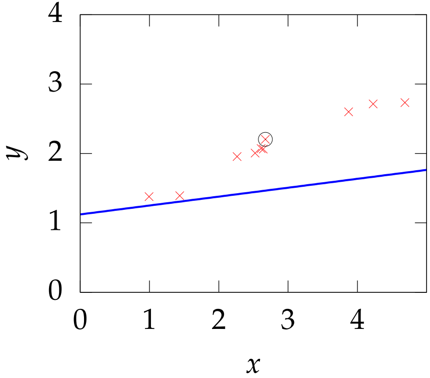
# Linear Regression Example

- Iteration 9 $\hat{m} = 0.1282$ $\hat{c} = 1.122$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$



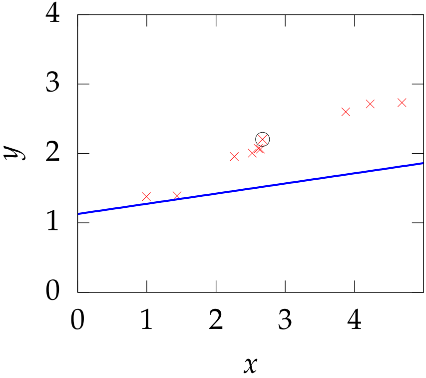
# Linear Regression Example

- Iteration 9 $\hat{m} = 0.1282$ $\hat{c} = 1.122$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_7$



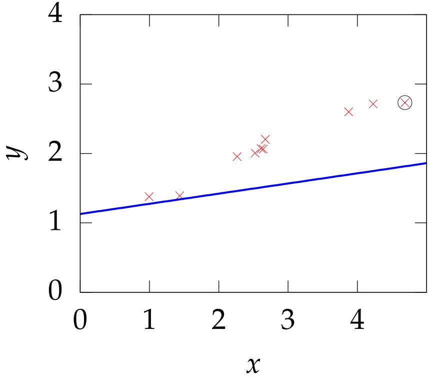
# Linear Regression Example

- Iteration 9 $\hat{m} = 0.1282$ $\hat{c} = 1.122$
  - Present data point 7
  - $\Delta y_7 = (y_7 - \hat{m}x_7 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_7 \Delta y_7$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_7$
- Updated values
  $\hat{m} = 0.14634$ $\hat{c} = 1.1288$



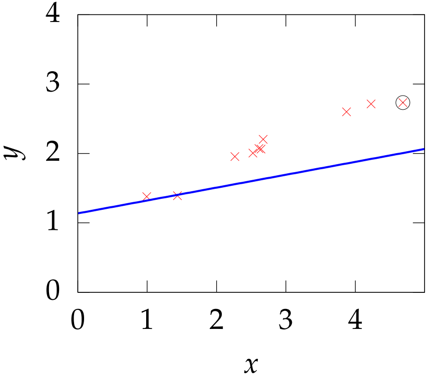
# Linear Regression Example

- Iteration 10 $\hat{m} = 0.14634$ $\hat{c} = 1.1288$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$



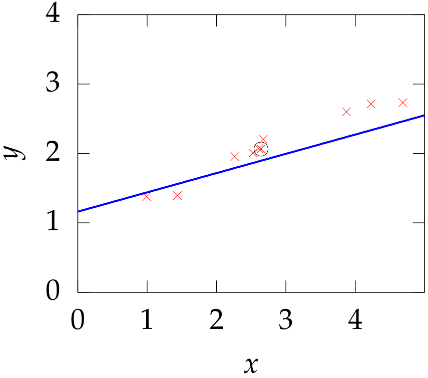
# Linear Regression Example

- Iteration 10 $\hat{m} = 0.14634$ $\hat{c} = 1.1288$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$
- Updated values
  $\hat{m} = 0.18547$ $\hat{c} = 1.1372$



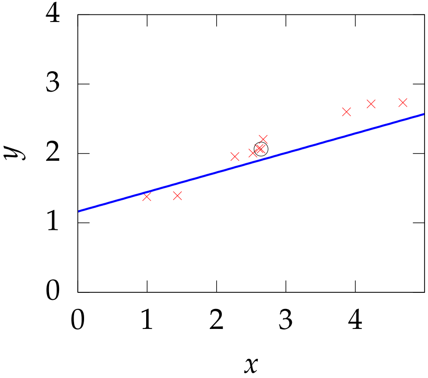
# Linear Regression Example

- Iteration 20 $\hat{m} = 0.27764$ $\hat{c} = 1.1621$
  - Present data point 6
  - $\Delta y_6 = (y_6 - \hat{m}x_6 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_6 \Delta y_6$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_6$



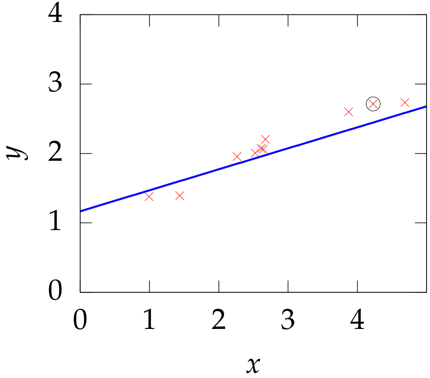
# Linear Regression Example

- Iteration 20 $\hat{m} = 0.27764$ $\hat{c} = 1.1621$
  - Present data point 6
  - $\Delta y_6 = (y_6 - \hat{m}x_6 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_6 \Delta y_6$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_6$
- Updated values
  $\hat{m} = 0.28135$ $\hat{c} = 1.1635$



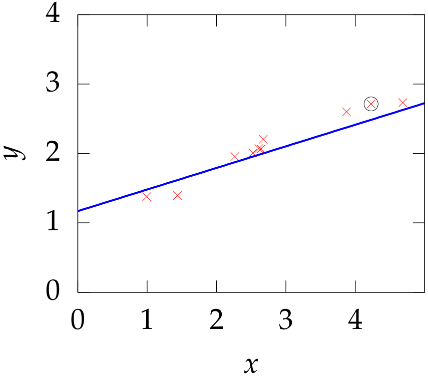
# Linear Regression Example

- Iteration 30 $\hat{m} = 0.30249$ $\hat{c} = 1.1673$
  - Present data point 9
  - $\Delta y_9 = (y_9 - \hat{m}x_9 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_9 \Delta y_9$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_9$



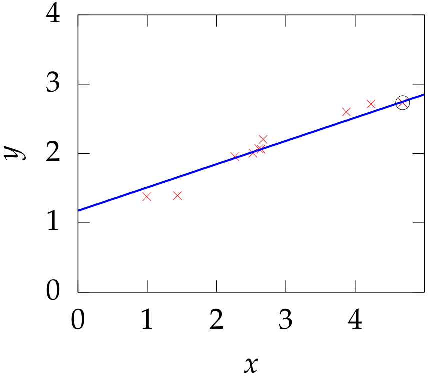
# Linear Regression Example

- Iteration 30 $\hat{m} = 0.30249$ $\hat{c} = 1.1673$
  - Present data point 9
  - $\Delta y_9 = (y_9 - \hat{m}x_9 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_9 \Delta y_9$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_9$
- Updated values
  $\hat{m} = 0.31119$ $\hat{c} = 1.1693$



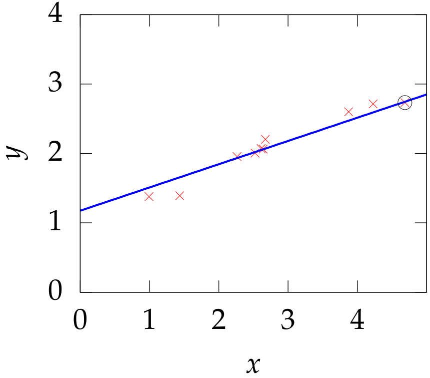
# Linear Regression Example

- Iteration 40 $\hat{m} = 0.33551$ $\hat{c} = 1.1754$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$



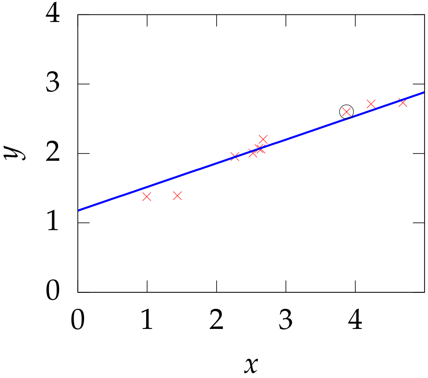
# Linear Regression Example

- Iteration 40 $\hat{m} = 0.33551$ $\hat{c} = 1.1754$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$
- Updated values
  $\hat{m} = 0.33503$ $\hat{c} = 1.1753$



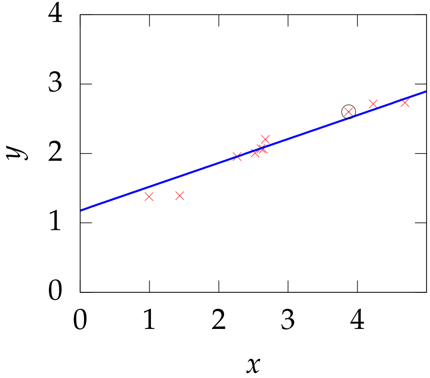
# Linear Regression Example

- Iteration 50 $\hat{m} = 0.34126$ $\hat{c} = 1.1763$
  - Present data point 8
  - $\Delta y_8 = (y_8 - \hat{m}x_8 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_8 \Delta y_8$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_8$



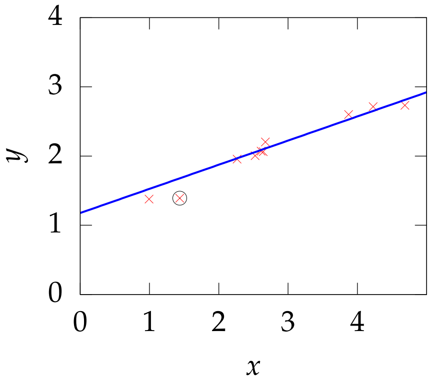
# Linear Regression Example

- Iteration 50 $\hat{m} = 0.34126$ $\hat{c} = 1.1763$
  - Present data point 8
  - $\Delta y_8 = (y_8 - \hat{m}x_8 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_8 \Delta y_8$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_8$
- Updated values
  $\hat{m} = 0.3439$ $\hat{c} = 1.177$



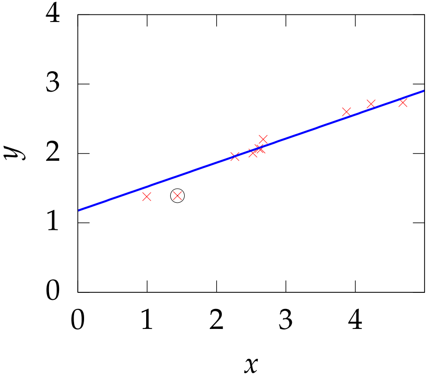
# Linear Regression Example

- Iteration 60 $\hat{m} = 0.34877$ $\hat{c} = 1.1775$
  - Present data point 2
  - $\Delta y_2 = (y_2 - \hat{m}x_2 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_2 \Delta y_2$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_2$



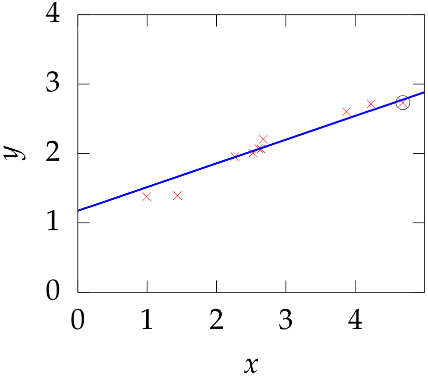
# Linear Regression Example

- Iteration 60 $\hat{m} = 0.34877$ $\hat{c} = 1.1775$
  - Present data point 2
  - $\Delta y_2 = (y_2 - \hat{m}x_2 - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_2 \Delta y_2$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_2$
- Updated values
  $\hat{m} = 0.34621$ $\hat{c} = 1.1757$



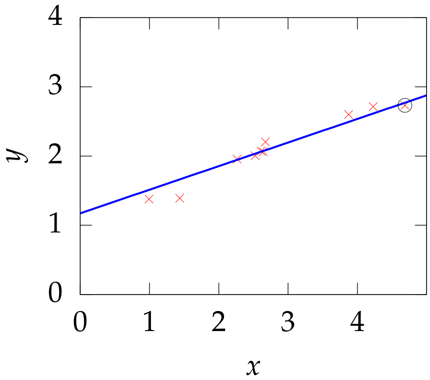
# Linear Regression Example

- Iteration 70 $\hat{m} = 0.34207$ $\hat{c} = 1.1734$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m}x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$



# Linear Regression Example

- Iteration 70 $\hat{m} = 0.34207$ $\hat{c} = 1.1734$
  - Present data point 10
  - $\Delta y_{10} = (y_{10} - \hat{m} x_{10} - \hat{c})$
  - Adjust $\hat{m}$ and $\hat{c}$
    $\hat{m} \leftarrow \hat{m} + \eta x_{10} \Delta y_{10}$
    $\hat{c} \leftarrow \hat{c} + \eta \Delta y_{10}$
- Updated values
  $\hat{m} = 0.34088$ $\hat{c} = 1.1732$



- Problem with Linear Regression—$\mathbf{x}$ may not be linearly related to $\mathbf{y}$.
- Potential solution: create a feature space: define $\phi(\mathbf{x})$ where $\phi(\cdot)$ is a nonlinear function of $\mathbf{x}$.
- Model for target is a linear combination of these nonlinear functions

$$f(\mathbf{x}) = \sum_{j=1}^{K} w_j \phi_j(\mathbf{x}) \tag{1}$$

# Quadratic Basis

- Basis functions can be global. E.g. quadratic basis:

$$[1, x, x^2]$$

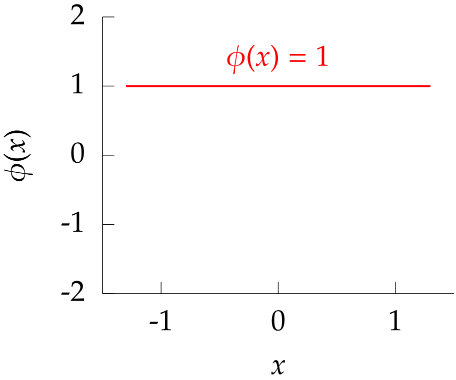



Figure: A quadratic basis.

# Quadratic Basis

- Basis functions can be global. E.g. quadratic basis:

$$[1, x, x^2]$$

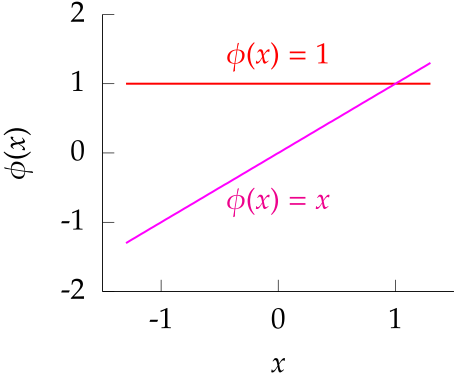



Figure: A quadratic basis.

# Quadratic Basis

- Basis functions can be global. E.g. quadratic basis:

$$[1, x, x^2]$$

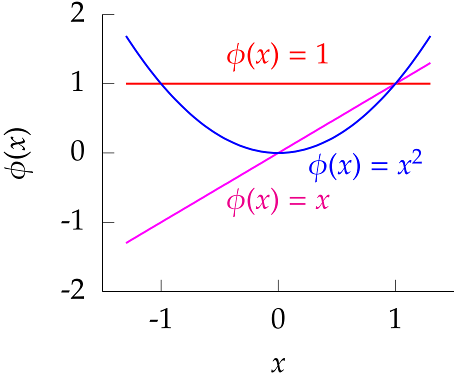



Figure: A quadratic basis.

# Functions Derived from Quadratic Basis

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

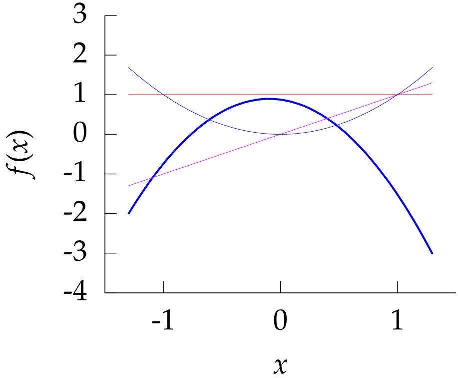



Figure: Function from quadratic basis with weights $w_1 = 0.87466$, $w_2 = -0.38835$, $w_3 = -2.0058$ .

# Functions Derived from Quadratic Basis

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

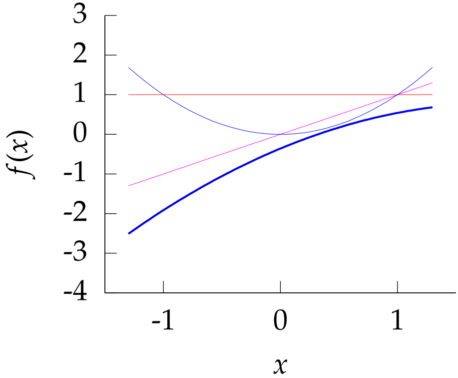



Figure: Function from quadratic basis with weights $w_1 = -0.35908$, $w_2 = 1.2274$, $w_3 = -0.32825$ .

# Functions Derived from Quadratic Basis

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

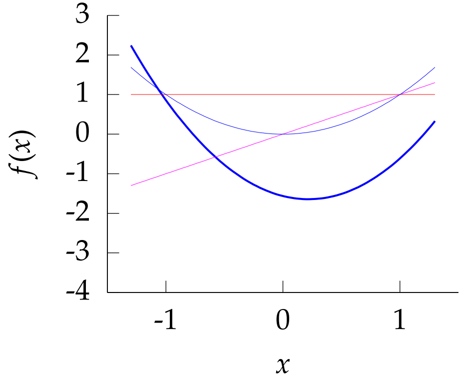



Figure: Function from quadratic basis with weights $w_1 = -1.5638$, $w_2 = -0.73577$, $w_3 = 1.6861$ .

# Radial Basis Functions

- Or they can be local. E.g. radial (or Gaussian) basis $\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{\ell^2}\right)$

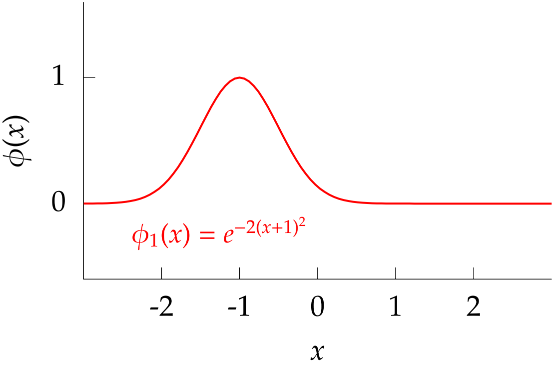



Figure: Radial basis functions.

# Radial Basis Functions

- Or they can be local. E.g. radial (or Gaussian) basis $\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{\ell^2}\right)$

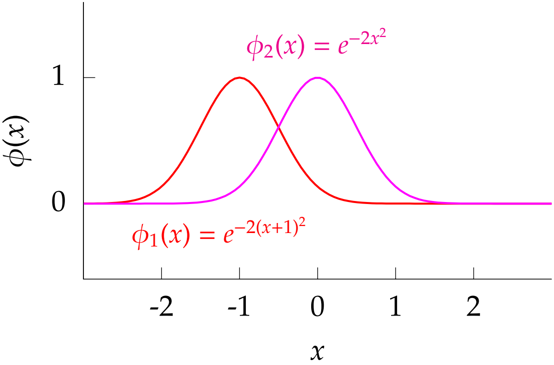



Figure: Radial basis functions.

# Radial Basis Functions

- Or they can be local. E.g. radial (or Gaussian) basis $\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{\ell^2}\right)$

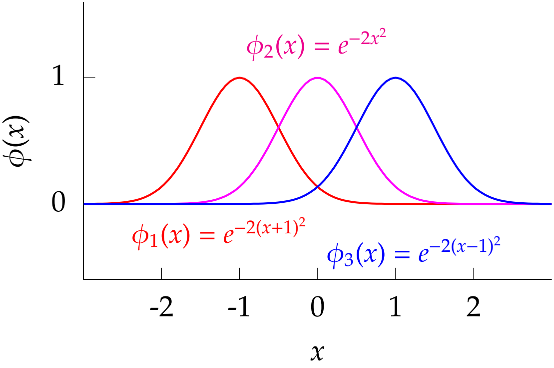



Figure: Radial basis functions.

# Functions Derived from Radial Basis

$$f(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2}$$

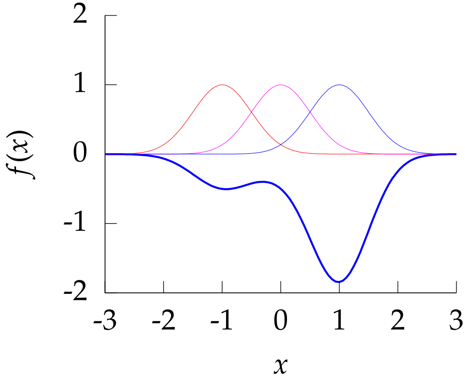



Figure: Function from radial basis with weights $w_1 = -0.47518$, $w_2 = -0.18924$, $w_3 = -1.8183$ .

# Functions Derived from Radial Basis

$$f(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2}$$

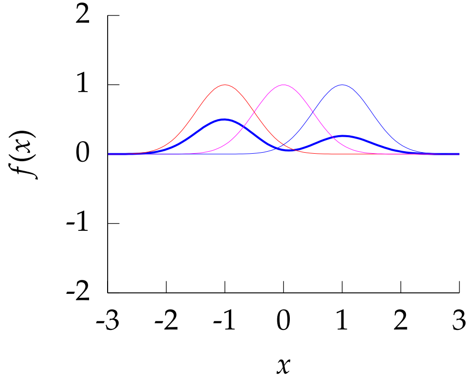



Figure: Function from radial basis with weights $w_1 = 0.50596$, $w_2 = -0.046315$, $w_3 = 0.26813$ .

# Functions Derived from Radial Basis

$$f(x) = w_1 e^{-2(x+1)^2} + w_2 e^{-2x^2} + w_3 e^{-2(x-1)^2}$$

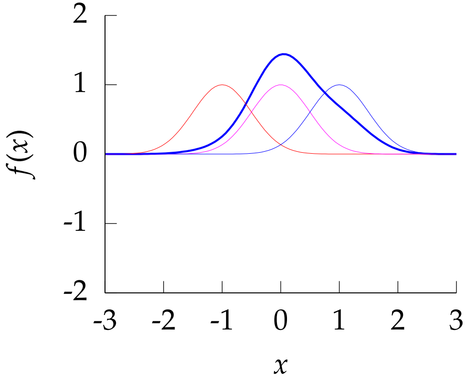
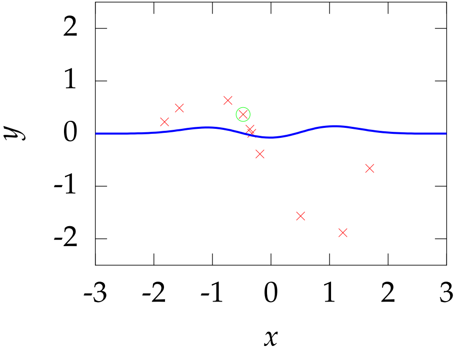



Figure: Function from radial basis with weights $w_1 = 0.07179$, $w_2 = 1.3591$, $w_3 = 0.50604$ .

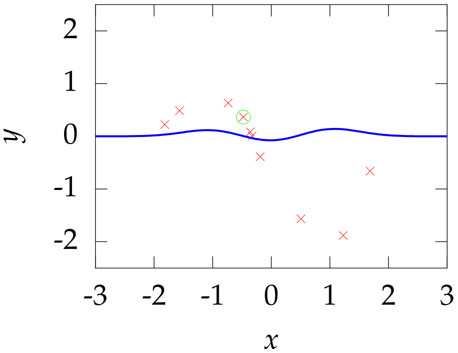
# Nonlinear Regression Example

- Iteration 1
  - $w_1 = 0.13018$,
    $w_2 = -0.11355$,
    $w_3 = 0.15448$
  - Present data point 4



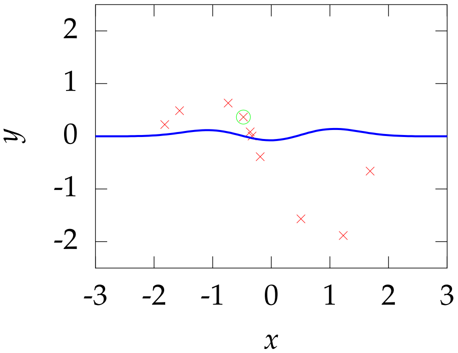
# Nonlinear Regression Example

- Iteration 1
  - $w_1 = 0.13018$,
    $w_2 = -0.11355$,
    $w_3 = 0.15448$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$



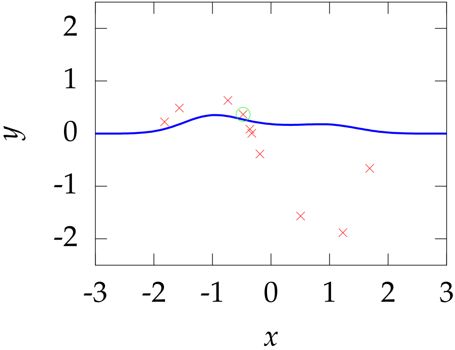
# Nonlinear Regression Example

- Iteration 1
  - $w_1 = 0.13018$,
    $w_2 = -0.11355$,
    $w_3 = 0.15448$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$



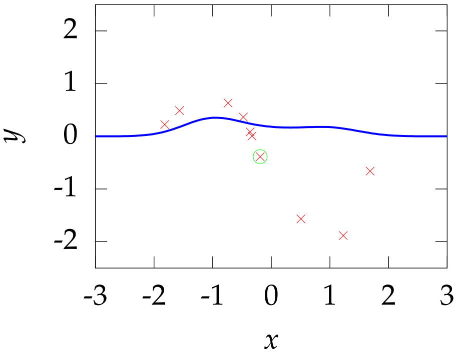
# Nonlinear Regression Example

- Iteration 1
  - $w_1 = 0.13018$,
    $w_2 = -0.11355$,
    $w_3 = 0.15448$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_4 \Delta y_4$



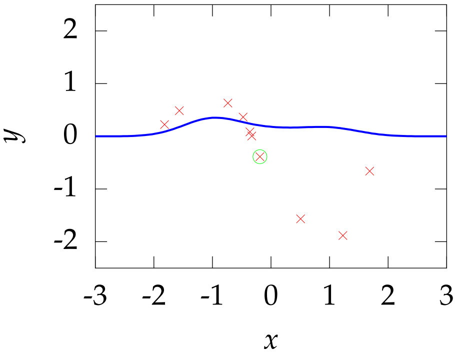
# Nonlinear Regression Example

- Iteration 2
  - $w_1 = 0.33696$,
    $w_2 = 0.11481$,
    $w_3 = 0.1591$
  - Present data point 7



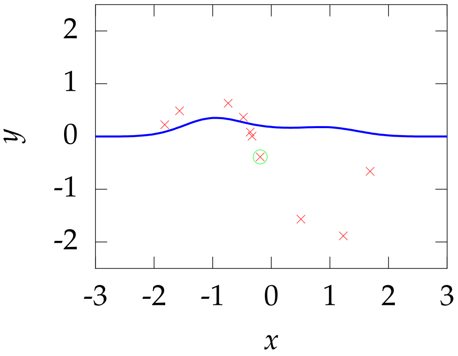
# Nonlinear Regression Example

- Iteration 2
  - $w_1 = 0.33696$,
    $w_2 = 0.11481$,
    $w_3 = 0.1591$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$



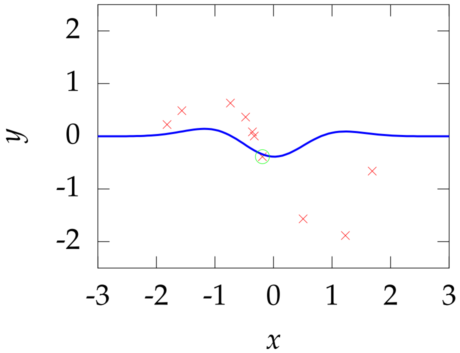
# Nonlinear Regression Example

- Iteration 2
  - $w_1 = 0.33696$,
    $w_2 = 0.11481$,
    $w_3 = 0.1591$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$



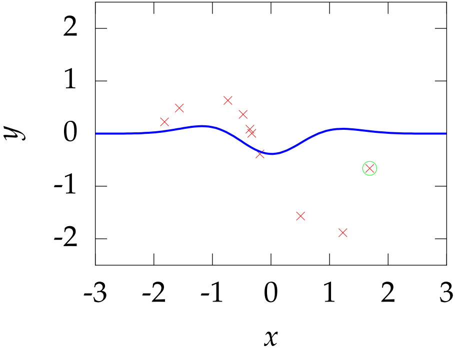
# Nonlinear Regression Example

- Iteration 2
  - $w_1 = 0.33696,$
    $w_2 = 0.11481,$
    $w_3 = 0.1591$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_7 \Delta y_7$



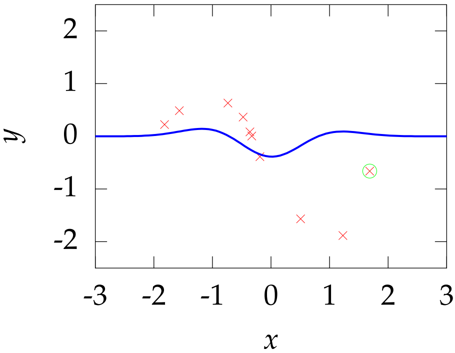
# Nonlinear Regression Example

- Iteration 3
  - $w_1 = 0.18076$,
    $w_2 = -0.4266$,
    $w_3 = 0.12473$
  - Present data point 10



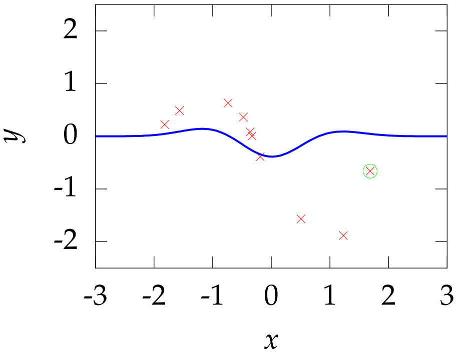
# Nonlinear Regression Example

- Iteration 3
  - $w_1 = 0.18076$,
    $w_2 = -0.4266$,
    $w_3 = 0.12473$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$



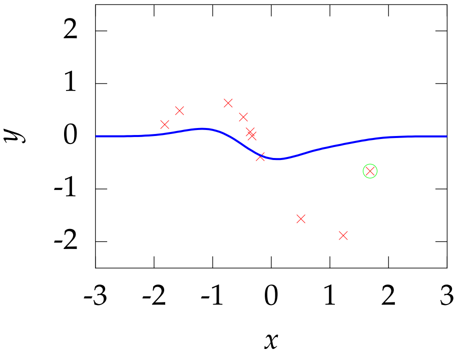
# Nonlinear Regression Example

- Iteration 3
  - $w_1 = 0.18076$,
    $w_2 = -0.4266$,
    $w_3 = 0.12473$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$



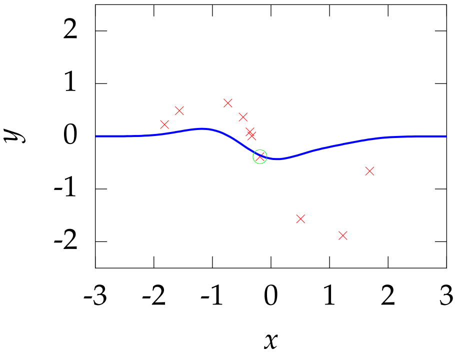
# Nonlinear Regression Example

- Iteration 3
  - $w_1 = 0.18076,$
    $w_2 = -0.4266,$
    $w_3 = 0.12473$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_{10} \Delta y_{10}$



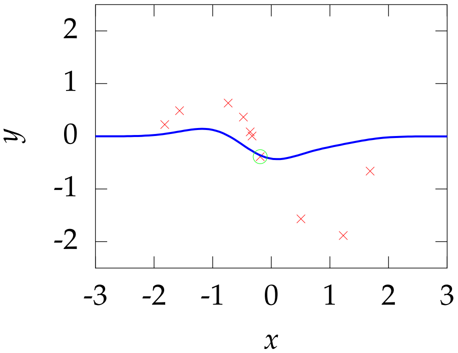
# Nonlinear Regression Example

- Iteration 4
  - $w_1 = 0.18076$,
    $w_2 = -0.42893$,
    $w_3 = -0.14306$
  - Present data point 7



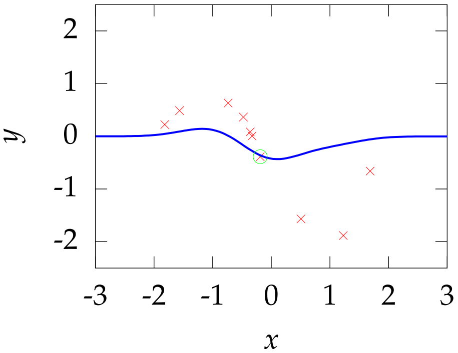
# Nonlinear Regression Example

- Iteration 4
  - $w_1 = 0.18076$,
    $w_2 = -0.42893$,
    $w_3 = -0.14306$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$



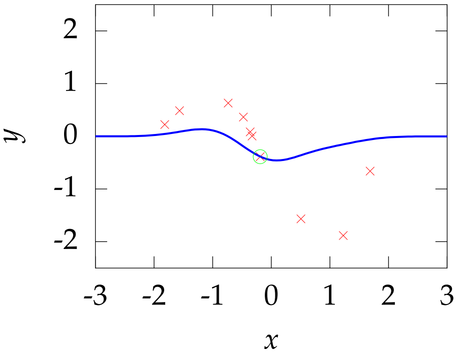
# Nonlinear Regression Example

- Iteration 4
  - $w_1 = 0.18076$,
    $w_2 = -0.42893$,
    $w_3 = -0.14306$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$



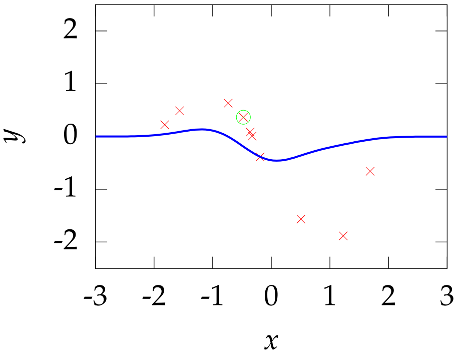
# Nonlinear Regression Example

- Iteration 4
  - $w_1 = 0.18076$,
    $w_2 = -0.42893$,
    $w_3 = -0.14306$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_7 \Delta y_7$



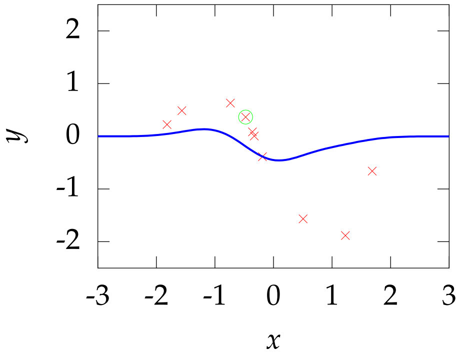
# Nonlinear Regression Example

- Iteration 5
  - $w_1 = 0.17372$,
    $w_2 = -0.45335$,
    $w_3 = -0.14461$
  - Present data point 4



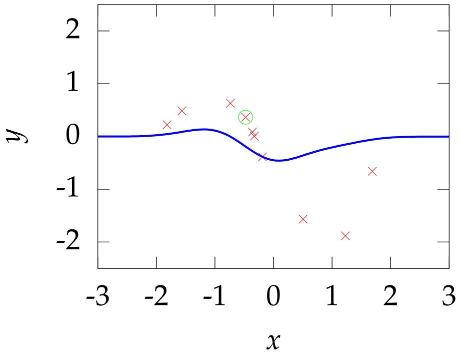
# Nonlinear Regression Example

- Iteration 5
  - $w_1 = 0.17372,$
    $w_2 = -0.45335,$
    $w_3 = -0.14461$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$



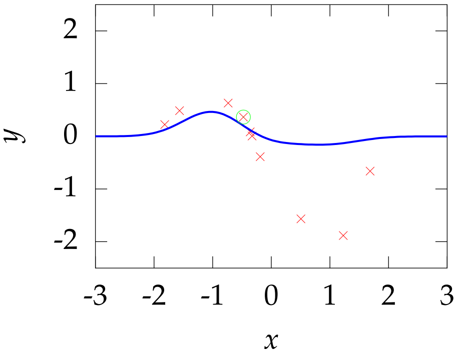
# Nonlinear Regression Example

- Iteration 5
  - $w_1 = 0.17372$,
    $w_2 = -0.45335$,
    $w_3 = -0.14461$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$



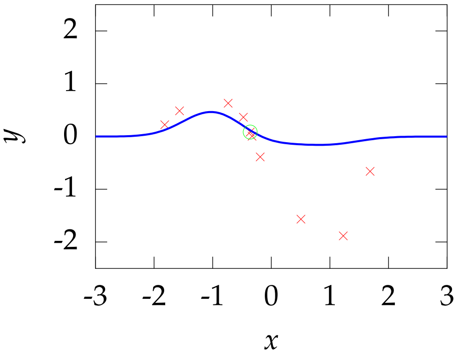
# Nonlinear Regression Example

- Iteration 5
  - $w_1 = 0.17372,$
    $w_2 = -0.45335,$
    $w_3 = -0.14461$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_4 \Delta y_4$



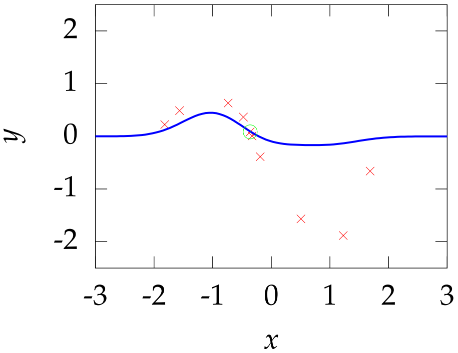
# Nonlinear Regression Example

- Iteration 6
  - $w_1 = 0.47971,$
    $w_2 = -0.11541,$
    $w_3 = -0.13778$
  - Present data point 5
  - $\Delta y_5 = y_5 - \phi_5^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_5 \Delta y_5$



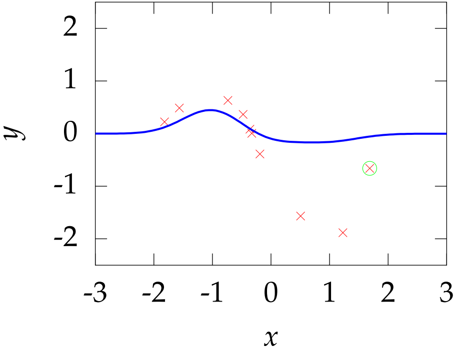
# Nonlinear Regression Example

- Iteration 6
  - $w_1 = 0.47971$,
    $w_2 = -0.11541$,
    $w_3 = -0.13778$
  - Present data point 5
  - $\Delta y_5 = y_5 - \phi_5^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_5 \Delta y_5$



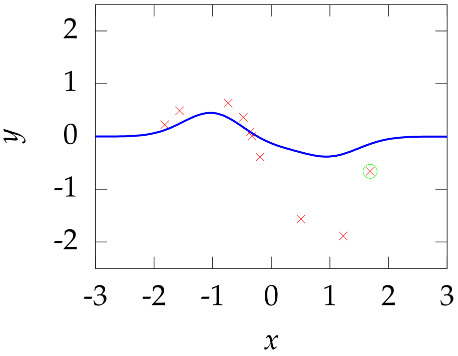
# Nonlinear Regression Example

- Iteration 7
  - $w_1 = 0.46599$,
    $w_2 = -0.13952$,
    $w_3 = -0.13855$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_{10} \Delta y_{10}$



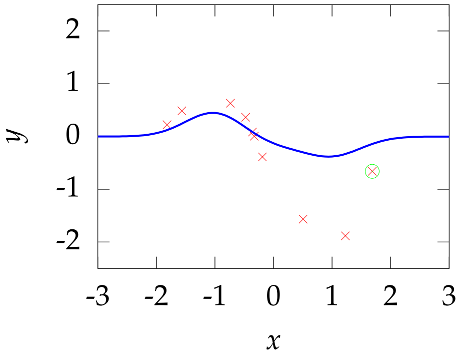
# Nonlinear Regression Example

- Iteration 7
  - $w_1 = 0.46599$,
    $w_2 = -0.13952$,
    $w_3 = -0.13855$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_{10} \Delta y_{10}$



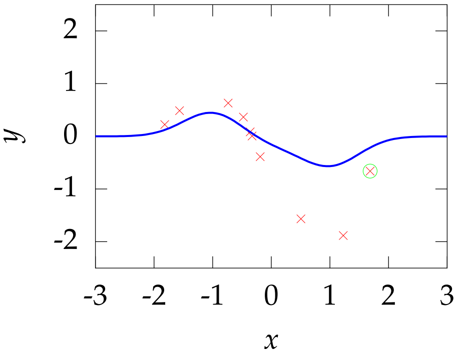
# Nonlinear Regression Example

- Iteration 8
  - $w_1 = 0.46599$,
    $w_2 = -0.14144$,
    $w_3 = -0.35924$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_{10} \Delta y_{10}$



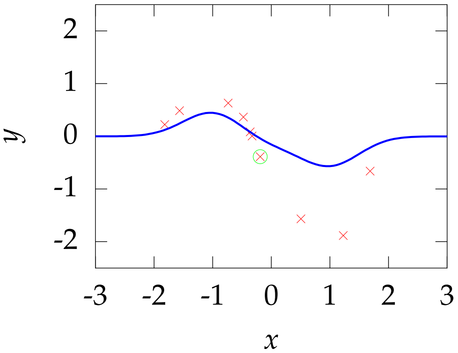
# Nonlinear Regression Example

- Iteration 8
  - $w_1 = 0.46599$,
    $w_2 = -0.14144$,
    $w_3 = -0.35924$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_{10} \Delta y_{10}$



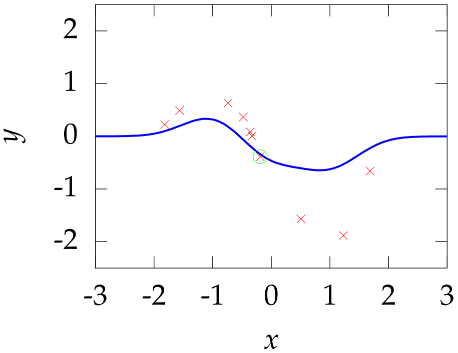
# Nonlinear Regression Example

- Iteration 9
  - $w_1 = 0.46599$,
    $w_2 = -0.14307$,
    $w_3 = -0.54679$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_7 \Delta y_7$



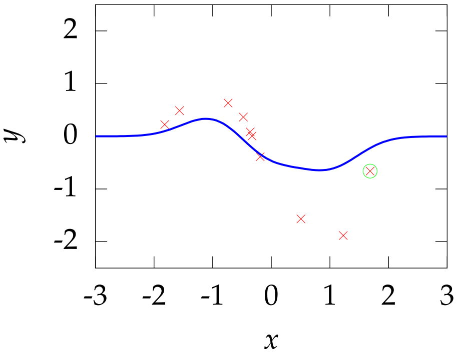
# Nonlinear Regression Example

- Iteration 9
  - $w_1 = 0.46599$,
    $w_2 = -0.14307$,
    $w_3 = -0.54679$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_7 \Delta y_7$



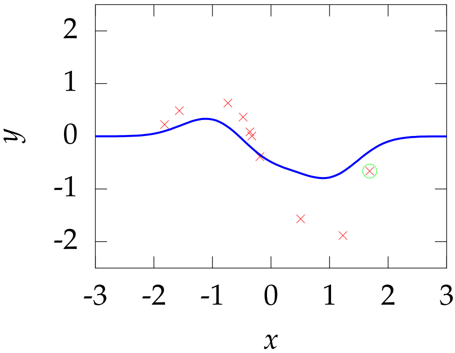
# Nonlinear Regression Example

- Iteration 10
  - $w_1 = 0.38071$,
    $w_2 = -0.43867$,
    $w_3 = -0.56556$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta\boldsymbol{\phi}_{10}\Delta y_{10}$



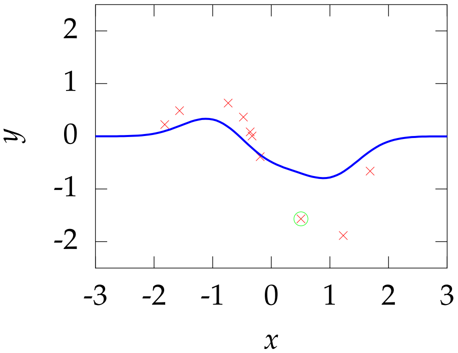
# Nonlinear Regression Example

- Iteration 10
  - $w_1 = 0.38071$,
    $w_2 = -0.43867$,
    $w_3 = -0.56556$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_{10} \Delta y_{10}$



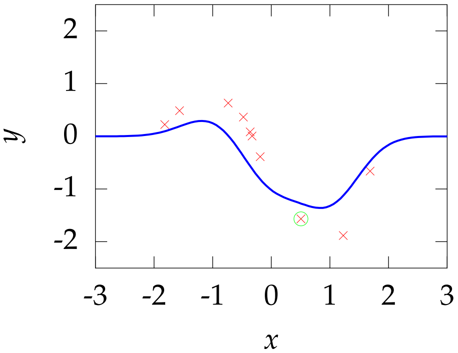
# Nonlinear Regression Example

- Iteration 11
  - $w_1 = 0.38071$,
    $w_2 = -0.44002$,
    $w_3 = -0.7208$
  - Present data point 8
  - $\Delta y_8 = y_8 - \phi_8^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_8 \Delta y_8$



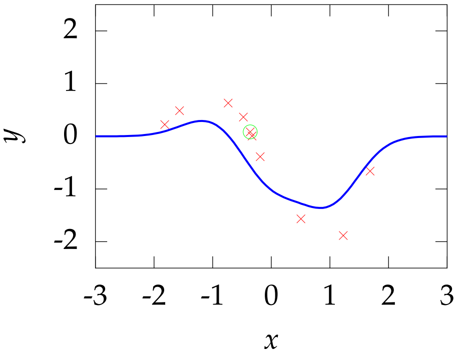
# Nonlinear Regression Example

- Iteration 11
  - $w_1 = 0.38071$,
    $w_2 = -0.44002$,
    $w_3 = -0.7208$
  - Present data point 8
  - $\Delta y_8 = y_8 - \phi_8^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_8 \Delta y_8$



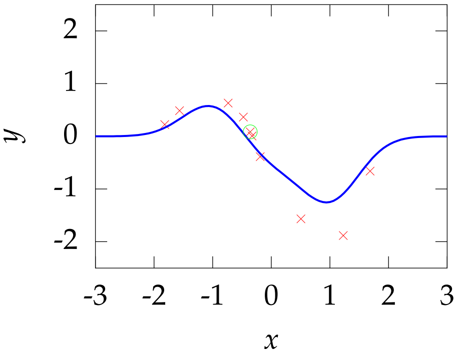
# Nonlinear Regression Example

- Iteration 12
  - $w_1 = 0.37237$,
    $w_2 = -0.90666$,
    $w_3 = -1.1987$
  - Present data point 5
  - $\Delta y_5 = y_5 - \phi_5^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_5 \Delta y_5$



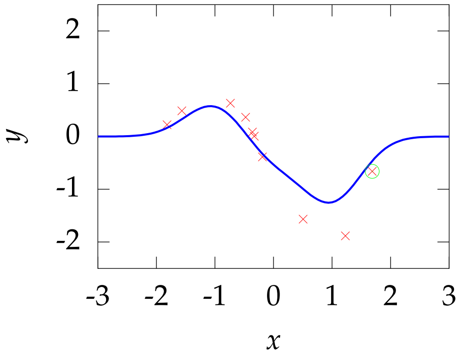
# Nonlinear Regression Example

- Iteration 12
  - $w_1 = 0.37237$,
    $w_2 = -0.90666$,
    $w_3 = -1.1987$
  - Present data point 5
  - $\Delta y_5 = y_5 - \phi_5^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_5 \Delta y_5$



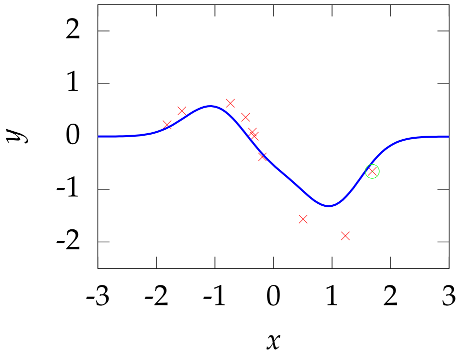
# Nonlinear Regression Example

- Iteration 13
  - $w_1 = 0.62833$,
    $w_2 = -0.45691$,
    $w_3 = -1.1842$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top}\mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta\boldsymbol{\phi}_{10}\Delta y_{10}$



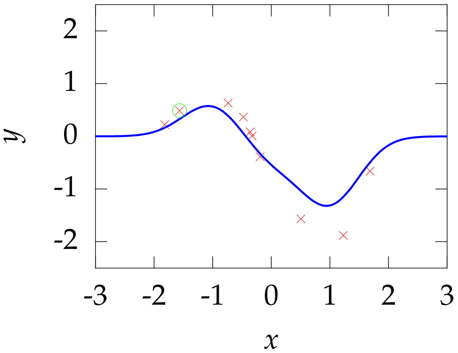
# Nonlinear Regression Example

- Iteration 13
  - $w_1 = 0.62833$,
    $w_2 = -0.45691$,
    $w_3 = -1.1842$
  - Present data point 10
  - $\Delta y_{10} = y_{10} - \phi_{10}^{\top} \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_{10} \Delta y_{10}$



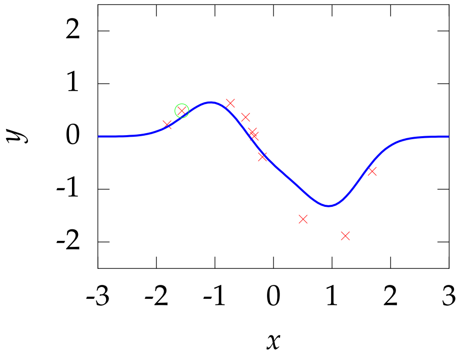
# Nonlinear Regression Example

- Iteration 14
  - $w_1 = 0.62833$,
    $w_2 = -0.4575$,
    $w_3 = -1.252$
  - Present data point 2
  - $\Delta y_2 = y_2 - \phi_2^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_2 \Delta y_2$



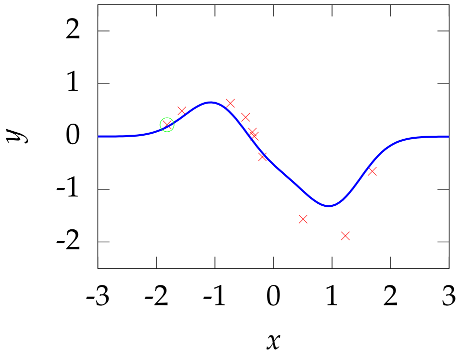
# Nonlinear Regression Example

- Iteration 14
  - $w_1 = 0.62833,$
    $w_2 = -0.4575,$
    $w_3 = -1.252$
  - Present data point 2
  - $\Delta y_2 = y_2 - \phi_2^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_2 \Delta y_2$



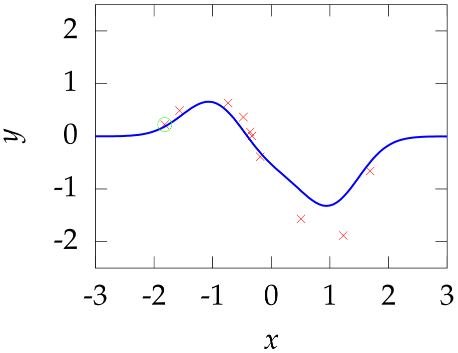
# Nonlinear Regression Example

- Iteration 15
  - $w_1 = 0.7016$,
    $w_2 = -0.45646$,
    $w_3 = -1.252$
  - Present data point 1
  - $\Delta y_1 = y_1 - \phi_1^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_1 \Delta y_1$



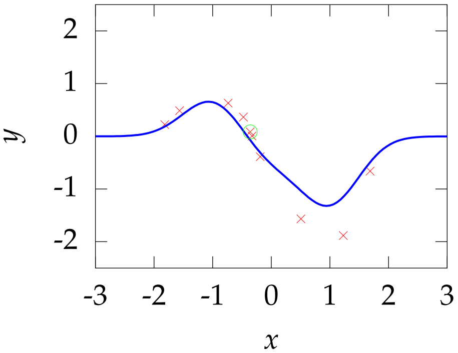
# Nonlinear Regression Example

- Iteration 15
  - $w_1 = 0.7016$,
    $w_2 = -0.45646$,
    $w_3 = -1.252$
  - Present data point 1
  - $\Delta y_1 = y_1 - \phi_1^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_1 \Delta y_1$



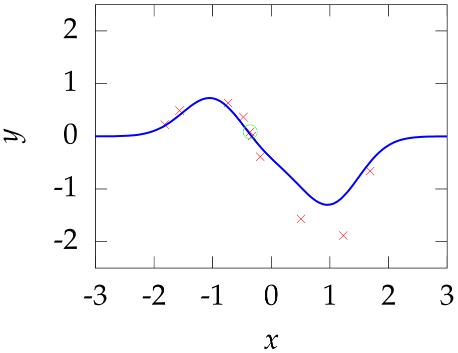
# Nonlinear Regression Example

- Iteration 16
  - $w_1 = 0.7109,$
    $w_2 = -0.45641,$
    $w_3 = -1.252$
  - Present data point 5
  - $\Delta y_5 = y_5 - \phi_5^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_5 \Delta y_5$



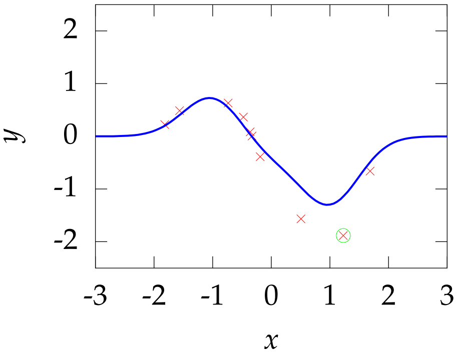
# Nonlinear Regression Example

- Iteration 16
  - $w_1 = 0.7109$,
    $w_2 = -0.45641$,
    $w_3 = -1.252$
  - Present data point 5
  - $\Delta y_5 = y_5 - \phi_5^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_5 \Delta y_5$



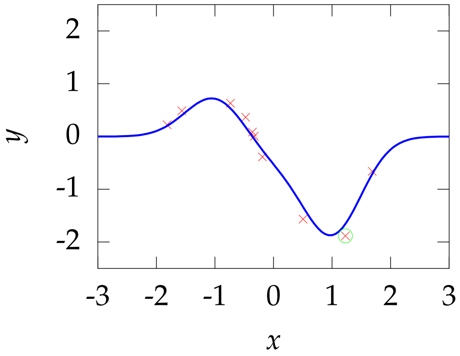
# Nonlinear Regression Example

- Iteration 17
  - $w_1 = 0.77022$,
    $w_2 = -0.35219$,
    $w_3 = -1.2487$
  - Present data point 9
  - $\Delta y_9 = y_9 - \phi_9^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_9 \Delta y_9$



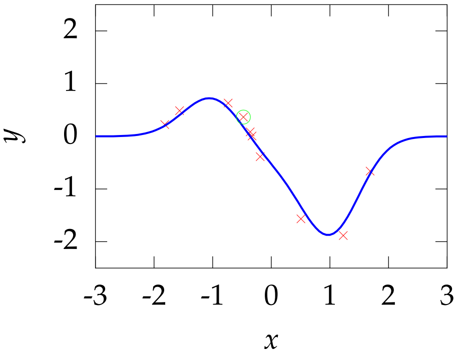
# Nonlinear Regression Example

- Iteration 17
  - $w_1 = 0.77022$,
    $w_2 = -0.35219$,
    $w_3 = -1.2487$
  - Present data point 9
  - $\Delta y_9 = y_9 - \phi_9^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_9 \Delta y_9$



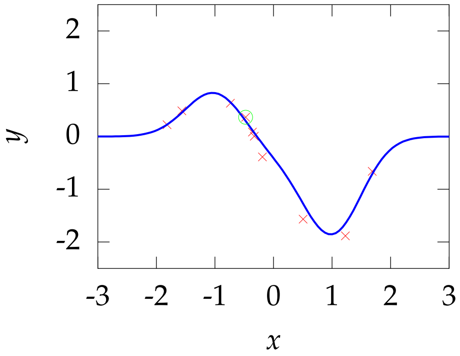
# Nonlinear Regression Example

- Iteration 18
  - $w_1 = 0.77019$,
    $w_2 = -0.3832$,
    $w_3 = -1.8175$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_4 \Delta y_4$



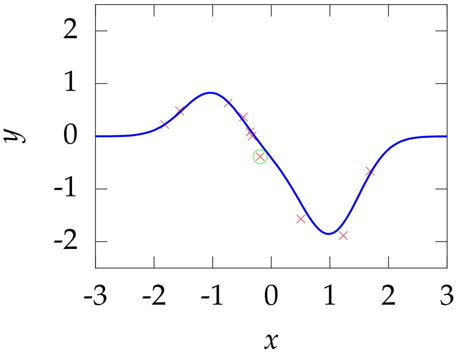
# Nonlinear Regression Example

- Iteration 18
  - $w_1 = 0.77019$,
    $w_2 = -0.3832$,
    $w_3 = -1.8175$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_4 \Delta y_4$



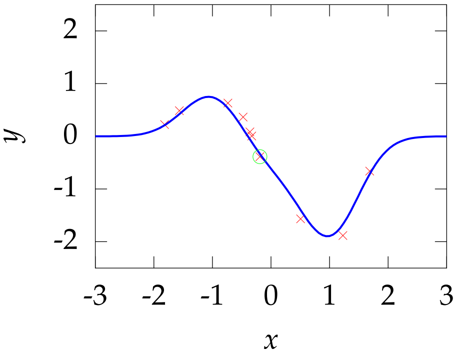
# Nonlinear Regression Example

- Iteration 19
  - $w_1 = 0.86321$,
    $w_2 = -0.28046$,
    $w_3 = -1.8154$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_7 \Delta y_7$



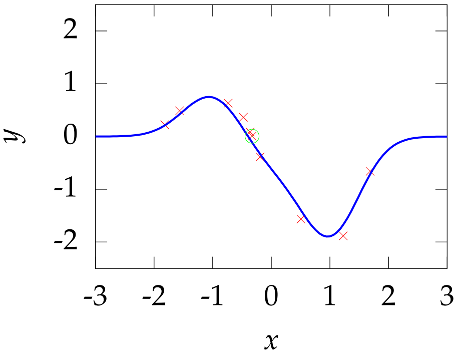
# Nonlinear Regression Example

- Iteration 19
  - $w_1 = 0.86321$,
    $w_2 = -0.28046$,
    $w_3 = -1.8154$
  - Present data point 7
  - $\Delta y_7 = y_7 - \phi_7^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_7 \Delta y_7$



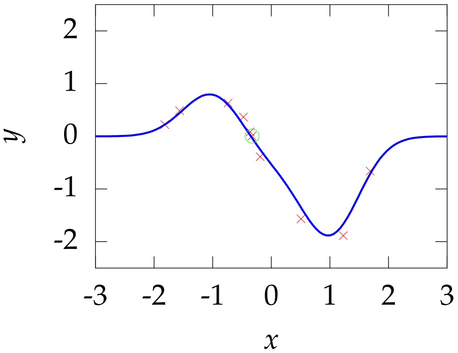
# Nonlinear Regression Example

- Iteration 20
  - $w_1 = 0.80681$,
    $w_2 = -0.47597$,
    $w_3 = -1.8278$
  - Present data point 6
  - $\Delta y_6 = y_6 - \phi_6^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_6 \Delta y_6$



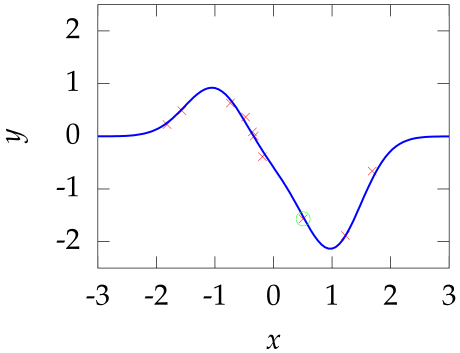
# Nonlinear Regression Example

- Iteration 20
  - $w_1 = 0.80681$, $w_2 = -0.47597$, $w_3 = -1.8278$
  - Present data point 6
  - $\Delta y_6 = y_6 - \phi_6^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_6 \Delta y_6$



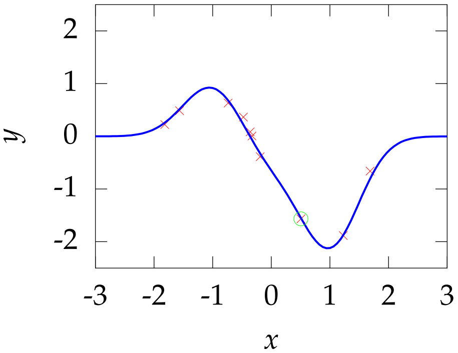
# Nonlinear Regression Example

- Iteration 50
  - $w_1 = 0.9777$,
    $w_2 = -0.4076$,
    $w_3 = -2.038$
  - Present data point 8
  - $\Delta y_8 = y_8 - \phi_8^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_8 \Delta y_8$



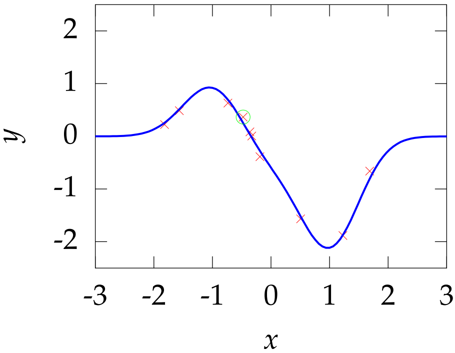
# Nonlinear Regression Example

- Iteration 100
  - $w_1 = 0.98593$,
    $w_2 = -0.49744$,
    $w_3 = -2.046$
  - Present data point 8
  - $\Delta y_8 = y_8 - \phi_8^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_8 \Delta y_8$



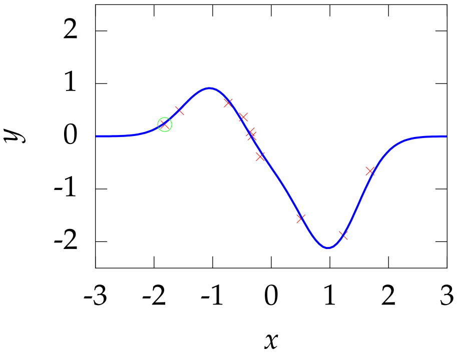
# Nonlinear Regression Example

- Iteration 200
  - $w_1 = 0.95307$,
    $w_2 = -0.48041$,
    $w_3 = -2.0553$
  - Present data point 4
  - $\Delta y_4 = y_4 - \phi_4^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_4 \Delta y_4$



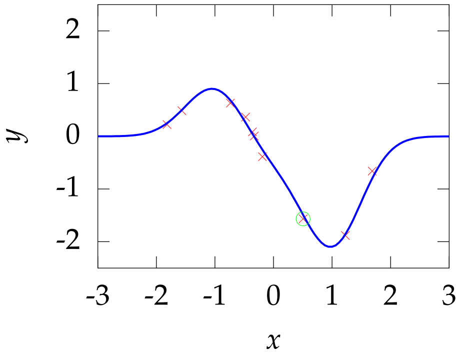
# Nonlinear Regression Example

- Iteration 300
  - $w_1 = 0.97066$,
    $w_2 = -0.44667$,
    $w_3 = -2.0588$
  - Present data point 1
  - $\Delta y_1 = y_1 - \phi_1^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_1 \Delta y_1$



# Nonlinear Regression Example

- Iteration 400
  - $w_1 = 0.95515,$
    $w_2 = -0.40611,$
    $w_3 = -2.0289$
  - Present data point 8
  - $\Delta y_8 = y_8 - \phi_8^\top \mathbf{w}$
  - Adjust $\hat{\mathbf{w}}$
- Updated values
  $\hat{\mathbf{w}} \leftarrow \hat{\mathbf{w}} + \eta \boldsymbol{\phi}_8 \Delta y_8$



# Mathematical Interpretation

- What is the mathematical interpretation?
  - There is a cost function.
  - It expresses mismatch between your prediction and reality.

$$E(\mathbf{w}) = \sum_{i=1}^{n} \left( \sum_{j=1}^{K} w_j \phi_j(x_i) - y_i \right)^2$$

  - This is known as the sum of squares error.

# Mathematical Interpretation

- What is the mathematical interpretation?
  - There is a cost function.
  - It expresses mismatch between your prediction and reality.

$$E(\mathbf{w}) = \sum_{i=1}^{n} \left(\mathbf{w}^\top \boldsymbol{\phi}_i - y_i\right)^2$$

  - This is known as the sum of squares error.
  - Defining $\boldsymbol{\phi}_i = \left[\phi_1(x_i), \ldots, \phi_K(x_i)\right]^\top$.

# Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Gradient of error function:

$$\frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}} = -2\sum_{i=1}^{n} \boldsymbol{\phi}_i \left(y_i - \mathbf{w}^\top \boldsymbol{\phi}_i\right)$$

# Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Gradient of error function:

$$\frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}} = -2\sum_{i=1}^{n} \boldsymbol{\phi}_i \Delta y_i$$

- Where $\Delta y_i = (y_i - \mathbf{w}^\top \boldsymbol{\phi}_i)$.

# Minimization via Gradient Descent

- One way of minimizing is steepest descent.
- Initialize algorithm with $\mathbf{w}$.
- Compute gradient of error function, $\frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}}$.
- Change $\mathbf{w}$ by moving in steepest downhill direction.

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}}$$

# Steepest Descent

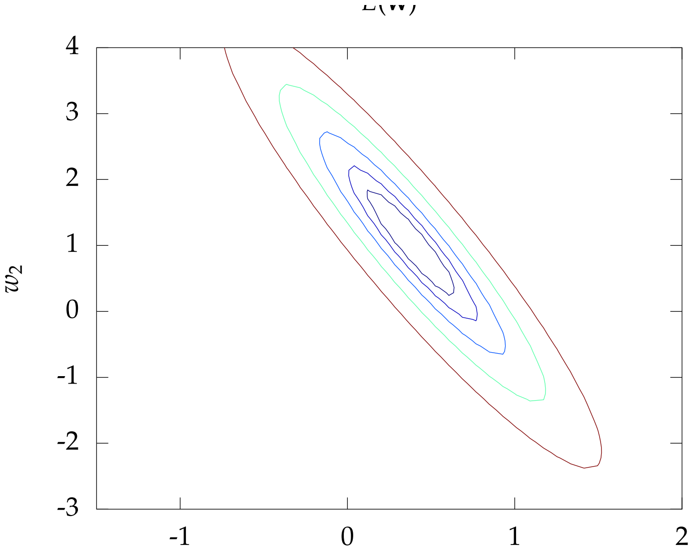



Figure: Steepest descent on a quadratic error surface.

# Stochastic Gradient Descent

## How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}}$$

# Stochastic Gradient Descent

## How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - 2\eta \sum_{i=1}^{n} \boldsymbol{\phi}_i \left( \mathbf{w}^\top \boldsymbol{\phi}_i - y_i \right)$$

# Stochastic Gradient Descent

## How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - \eta' \sum_{i=1}^{n} \boldsymbol{\phi}_i \left( \mathbf{w}^\top \boldsymbol{\phi}_i - y_i \right)$$

# Stochastic Gradient Descent

## How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - \eta' \sum_{i=1}^{n} \boldsymbol{\phi}_i \left( \mathbf{w}^\top \boldsymbol{\phi}_i - y_i \right)$$

# Stochastic Gradient Descent

## How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - \eta' \sum_{i=1}^{n} \boldsymbol{\phi}_i \Delta y_i$$

# Stochastic Gradient Descent

## How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - \eta' \sum_{i=1}^{n} \boldsymbol{\phi}_i \Delta y_i$$

- And the stochastic approximation is

$$\mathbf{w} \leftarrow \mathbf{w} - \eta' \boldsymbol{\phi}_i \Delta y_i$$

# Stochastic Gradient Descent

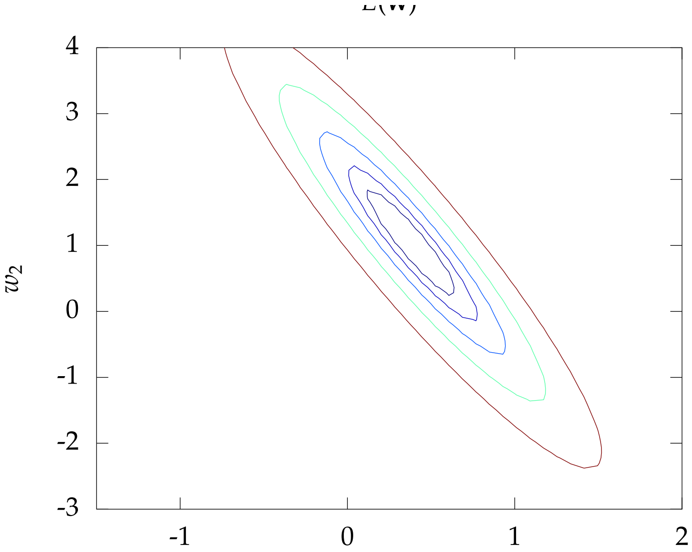



Figure: Stochastic gradient descent on a quadratic error surface.

# Modern View of Error Functions

- Error function has a probabilistic interpretation (maximum likelihood).
- Error function is an actual loss function that you want to minimize (empirical risk minimization).
- For these interpretations probability and optimization theory become important.
- Much of the last 15 years of machine learning research has focused on probabilistic interpretations or clever relaxations of difficult objective functions.

# Important Concepts Not Covered

- Optimization methods.
  - Second order methods, conjugate gradient, quasi-Newton and Newton.
  - Effective heuristics such as momentum.
- Local vs global solutions.

# Mathematical Interpretation

- What is the mathematical interpretation?
  - There is a cost function.
  - It expresses mismatch between your prediction and reality.

$$E(\mathbf{w}) = \sum_{i=1}^{n} \left( \sum_{j=1}^{K} w_j \phi_j(x_i) - y_i \right)^2$$

  - This is known as the sum of squares error.

# Mathematical Interpretation

- What is the mathematical interpretation?
  - There is a cost function.
  - It expresses mismatch between your prediction and reality.

$$E(\mathbf{w}) = \sum_{i=1}^{n} \left(\mathbf{w}^\top \boldsymbol{\phi}_i - y_i\right)^2$$

  - This is known as the sum of squares error.
  - Defining $\boldsymbol{\phi}_i = \left[\phi_1(x_i), \ldots, \phi_K(x_i)\right]^\top$.

# Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Gradient of error function:

$$\frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}} = -2\sum_{i=1}^{n} \boldsymbol{\phi}_i \left(y_i - \mathbf{w}^\top \boldsymbol{\phi}_i\right)$$

# Learning is Optimization

- Learning is minimization of the cost function.
- At the minima the gradient is zero.
- Gradient of error function:

$$\frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}} = -2\sum_{i=1}^{n} \boldsymbol{\phi}_i \Delta y_i$$

- Where $\Delta y_i = (y_i - \mathbf{w}^\top \boldsymbol{\phi}_i)$.

# Minimization via Gradient Descent

- One way of minimizing is steepest descent.
- Initialize algorithm with $\mathbf{w}$.
- Compute gradient of error function, $\frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}}$.
- Change $\mathbf{w}$ by moving in steepest downhill direction.

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}}$$

# Steepest Descent

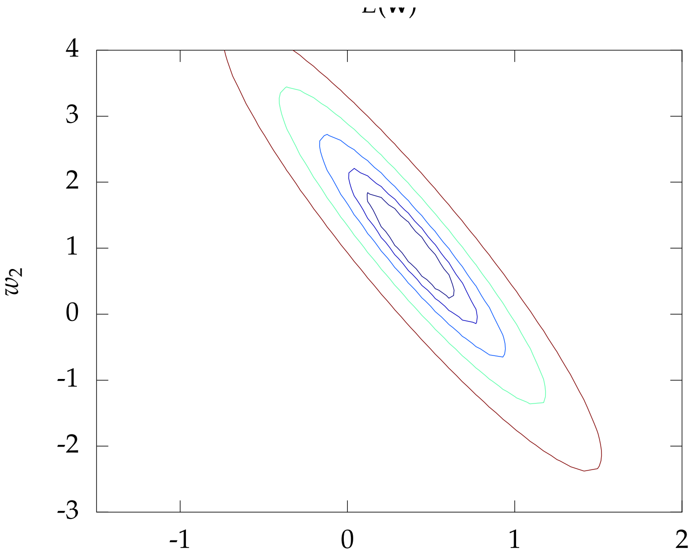



Figure: Steepest descent on a quadratic error surface.

# Stochastic Gradient Descent

## How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}}$$

# Stochastic Gradient Descent

## How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - 2\eta \sum_{i=1}^{n} \boldsymbol{\phi}_i \left( \mathbf{w}^\top \boldsymbol{\phi}_i - y_i \right)$$

# Stochastic Gradient Descent

## How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - \eta' \sum_{i=1}^{n} \boldsymbol{\phi}_i \left(\mathbf{w}^\top \boldsymbol{\phi}_i - y_i\right)$$

# Stochastic Gradient Descent

## How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - \eta' \sum_{i=1}^{n} \boldsymbol{\phi}_i \left( \mathbf{w}^\top \boldsymbol{\phi}_i - y_i \right)$$

# Stochastic Gradient Descent

## How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - \eta' \sum_{i=1}^{n} \boldsymbol{\phi}_i \Delta y_i$$

# Stochastic Gradient Descent

## How does this relate to learning rules we presented?

- For regression, the learning rule can be seen as a variant of gradient descent.
- This variant is known as stochastic gradient descent.
- For regression steepest descent gives

$$\mathbf{w} \leftarrow \mathbf{w} - \eta' \sum_{i=1}^{n} \boldsymbol{\phi}_i \Delta y_i$$

- And the stochastic approximation is

$$\mathbf{w} \leftarrow \mathbf{w} - \eta' \boldsymbol{\phi}_i \Delta y_i$$

# Stochastic Gradient Descent

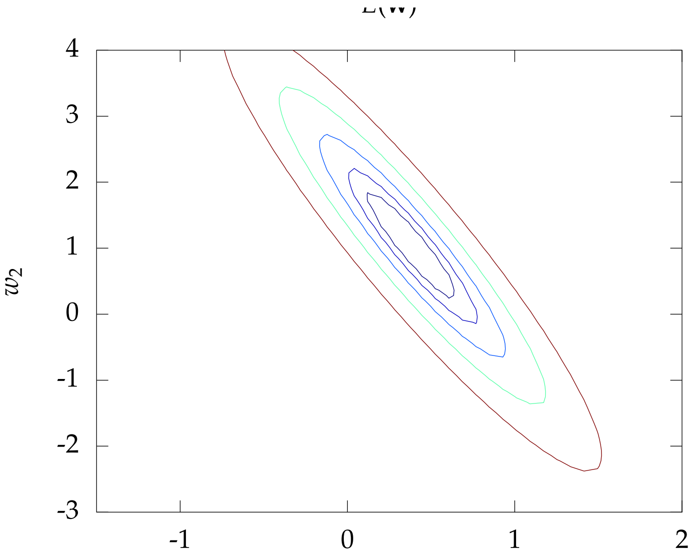



Figure: Stochastic gradient descent on a quadratic error surface.

# Modern View of Error Functions

- Error function has a probabilistic interpretation (maximum likelihood).
- Error function is an actual loss function that you want to minimize (empirical risk minimization).
- For these interpretations probability and optimization theory become important.
- Much of the last 15 years of machine learning research has focused on probabilistic interpretations or clever relaxations of difficult objective functions.

# Important Concepts Not Covered

- Optimization methods.
  - Second order methods, conjugate gradient, quasi-Newton and Newton.
  - Effective heuristics such as momentum.
- Local vs global solutions.

# References I