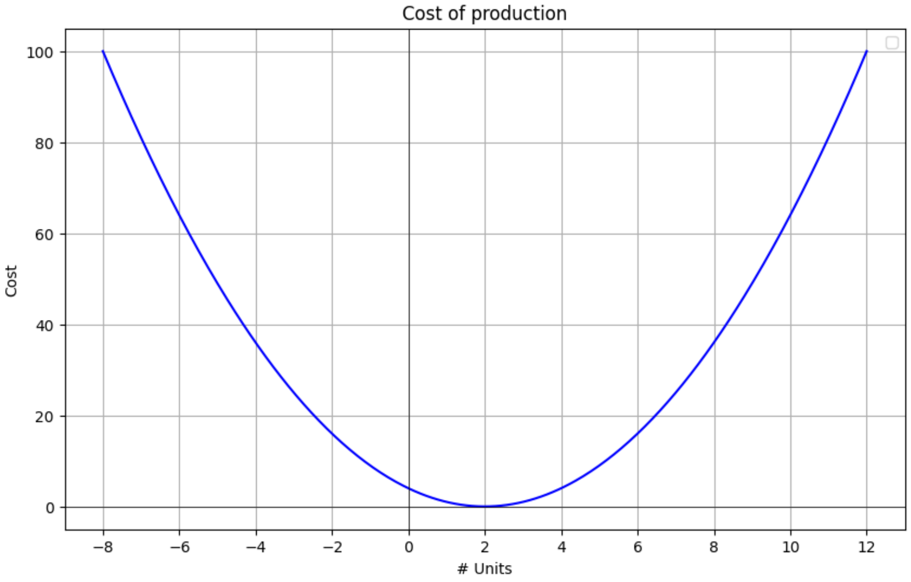
# Tutorial Sections 9-10

1. A company manufactures and sells a particular product. The company's analysts estimate that the cost to produce $x$ units of this product can be well approximated by the cost function $C(x) = x^2 - 4x + 4$.
   (a) Starting with $x_0 = 0$, use the sequential gradient descent (SGD) algorithm to find the optimal number of units that minimises production costs, assuming a tolerance of $\epsilon = 0.05$ and a learning rate of $\alpha = 0.25$.
   (b) What will be the number of units if we choose $\epsilon = 0.01$?
   (c) What would be the effect of increasing the value of the learning rate?

**Solution:** Although this is not strictly an ML problem, we can use this context to practice the different steps of the SGD algorithm. Note that this specific problem can actually be solved analytically, as the model function is well defined by $C(x)$. We can see the analytical solution by finding the minimum of the quadratic function, which is a simple parabola with its minimum at $x = 2$, as its graph below shows.



(a) If we ignore the analytical solution to practice the SGD algorithm, the problem translates into solving the following:

$$x^* = \underset{x' \in \mathbb{R}}{\arg\min}\, C(x').$$

(In this case, the feature in we want to minimize is the number of units, $x$). The SGD algorithm involves starting at $x_0 = 0$. For iteration $n = 0$, we have $C(x_0) = 4$. To compute $x_1$, we need the gradient of the cost function (which, in this case, is just the derivative of the cost function with respect to $x$ since this problem involves only this parameter to optimize):

$$C'(x) = \frac{dC}{dx} = 2x - 4$$

The gradient descent step is then:

$$x_1 = x_0 - \alpha \cdot C'(x_0) = 0 - 0.25 \cdot (-4) = 1$$

With this value, the updated cost function is $C(x_1) = 1^2 - 4 \cdot 1 + 4 = 1$, and the cost function improvement is

$$\Delta C = |C(x_1) - C(x_0)| = |1 - 4| = | - 3| = 3,$$

which is greater than the convergence tolerance, $\epsilon = 0.05$. So, let's compute the same variables for $n = 2$:

$$\begin{aligned} n = 2 :& x_2 = 1.5, C(x_2) = 0.25, \Delta C = 0.75 > \epsilon \\ n = 3 :& x_3 = 1.75, C(x_3) = 0.0625, \Delta C = 0.1875 > \epsilon \\ n = 4 :& x_4 = 1.875, C(x_4) = 0.0156, \Delta C = 0.0469 < \epsilon. \end{aligned}$$

Therefore, considering the parameters given for $\epsilon$ and $\alpha$, and the starting point $x_0 = 0$, the optimal solution is $x^* = x_4 = 1.875$.

(b) If $\epsilon = 0.01$, the algorithm won't stop at $n = 4$ because $\Delta C = C(x_4) - C(x_3) > \epsilon$. Therefore, we would have:

$$\begin{aligned} n = 5 :& x_5 = 1.9375, C(x_5) = 0.0039, \Delta C = 0.0117 > \epsilon \\ n = 6 :& x_6 = 1.9688, C(x_6) = 0.0010, \Delta C = 0.0029 < \epsilon, \end{aligned}$$

which results in the optimal solution $x^* = x_6 = 1.9688$, which is closer to the actual value for the minimum.

(c) Increasing the value of $\alpha$ could have the effect of converging to the optimal solution faster since it will increase the steps between subsequent $x$'s. However, if the learning rate is high enough, the algorithm might overshoot the minimum, potentially moving further away from it with each iteration. This scenario would result in oscillations around the minimum or, in extreme cases, divergence.

2. A small startup has launched a new website and is monitoring the time users spend on their site. Their goal is to understand the relationship between the number of features on their homepage and the average time a user spends on the site. Due to limited traffic and resources, they only have data for three different homepage designs, each with a varying number of features.

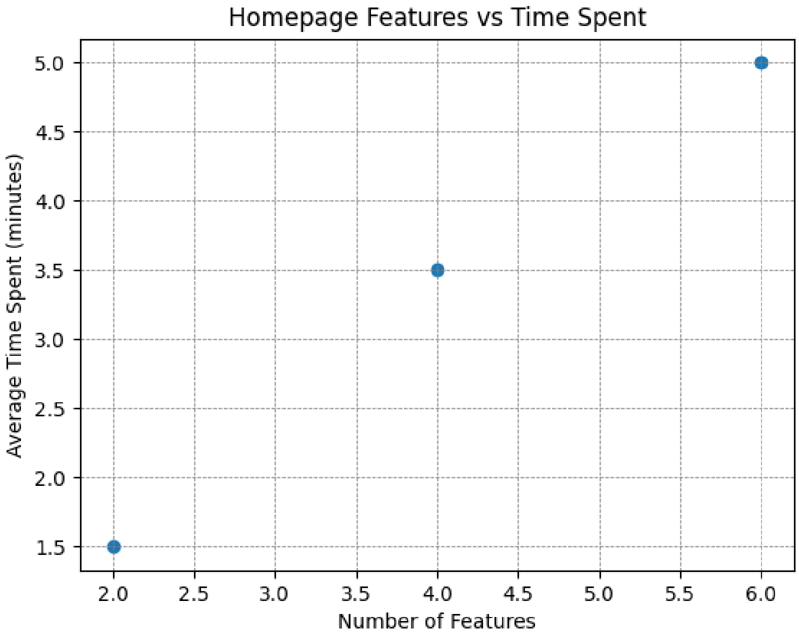
| Homepage | # of features | avg. time spent (min) |
|---|---|---|
| A | 4 | 4 |
| B | 6 | 5 |
| C | 2 | 1.5 |

(a) Plot the the average time spent on the site as function of number of features. Is there a relationship between these two variables? If so, how would you model this relationship?

(b) Use the sequential gradient descent (SGD) approach to fit your regression model to the data. In your regression, you want to minimise the square loss function between the data and your model. Assume a tolerance of $\epsilon = 0.2$ and a learning rate of $\alpha = 0.01$, and start at $w_0 = [0, 0]^T$.

**Solution:** This is a practical problem involving the same example from the notes.

(a) The plot is shown below. The data points suggest there is a linear relationship between the number of features and the average time spent on the website (the greater the number of features, the longer the time spent).



We can model the relationship between the number of features, $x$, and the average time spent, $y$, as linear: $y = ax + b$. Here, the two parameters that can be fit are $a$ and $b$, which we can write in vector notation as $w = [w_1, w_2]^T = [a, b]^T$ (i.e., $w_1 = a, w_2 = b$). This way, we can write the model as a linear regression problem:

$$f(x, w) = w_1 x^1 + w_2 x^2 = w^T \cdot X,$$

where $X = [x^1, x^2]^T = [x, 1]^T$. (Note, remember that, in our notation, the *dimension* of the vector $X$ is represented by the superscript; so, $x^2$ is the second dimension of vector $X$ and not $x \cdot x$, or $x$ squared, in the equation above.)

(b) The SGD can be used to find the parameters $w$ that minimise the sum-of-squares error (SSE) between the model and the data, $w^*$, given as:

$$w^* = \underset{w' \in \mathcal{W}}{\arg\min}\, F(w'),$$

with

$$F(w) = \sum_{i=1}^{3} \left(w^T X_i - y_i\right)^2,$$

where $X_i = [x_i, 1]^T$ for each data point $i$ (note that, although $w$ and $X$ are 2-dimensional vectors, the multiplication $w^T X$ results in a number).
Considering the choice of the starting parameters, we have

$$n = 0 : w_0 = [0, 0]^T, \text{ which leads to } F(w_0) = 43.25.$$

The partial derivative of $F(w)$ with respect to $w$ is given by:

$$F_w(w) = 2\sum_{i=1}^{N} \left(w^T X_i - y_i\right) X_i$$

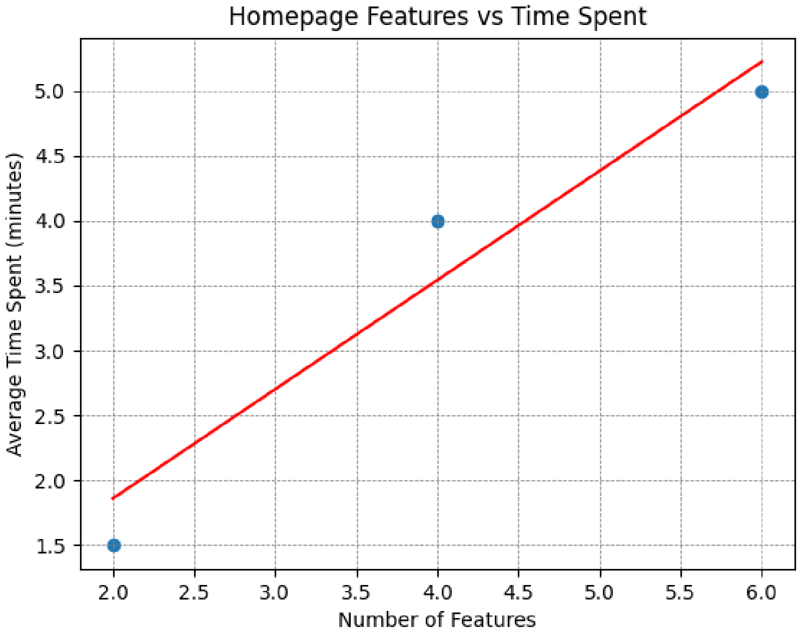
so that $F_w(w_0) = [-98, -21]^T$. This results in the following updates:

$n = 1 : w_1 = [0.98, 0.21]^T, F(w_1) = 1.6539, \Delta F = 41.5961 > \epsilon, F_w(w_1) = [16.8, 3.78]^T$

$n = 2 : w_2 = [0.8120, 0.1722]^T, F(w_2) = 0.4259, \Delta F = 1.2280 > \epsilon, F_w(w_2) = [-2.9232, -0.4788]^T$

$n = 3 : w_3 = [0.8412, 0.1770]^T, F(w_3) = 0.3894, \Delta F = 0.0365 < \epsilon = 0.2$

Therefore, the optimal parameters are $w^\star = w_3 = [0.8412, 0.1770]^T$, which means that the regression line that best fits the data, using the square loss function, will be $f(x) = 0.8412x + 0.1770$. This line is shown in red in the figure below, along with the data points.


Homepage Features vs Time Spent
Average Time Spent (minutes)
Number of Features