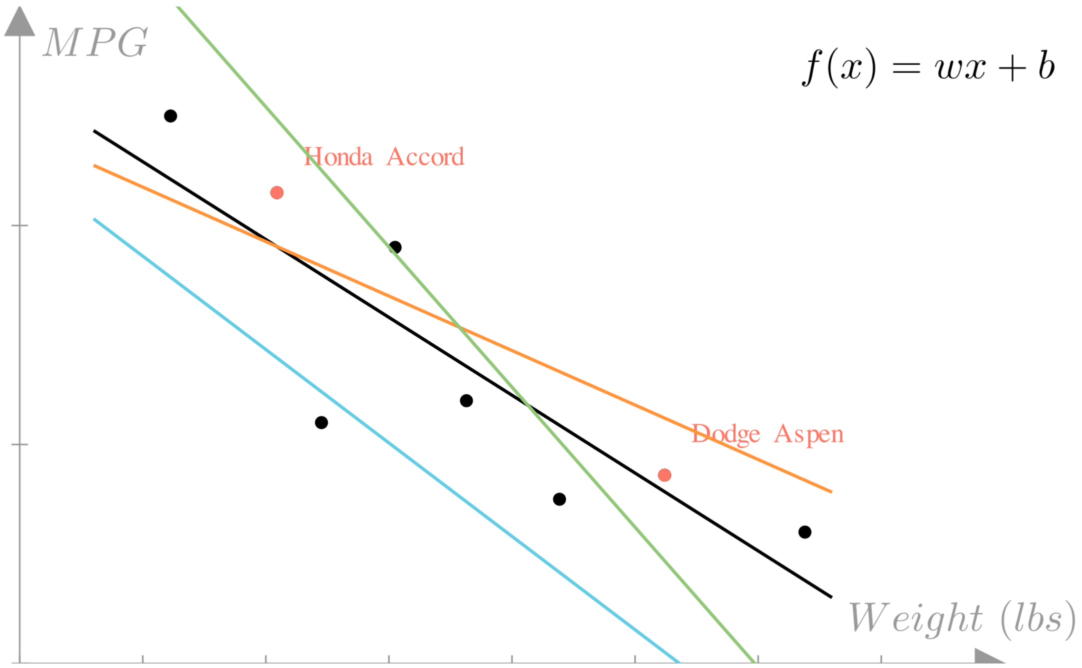
$$f(\mathbf{x}) = \sum_{i=1}^{n} x_i w_i + b$$

We typically refer to $\mathbf{w}$ specifically as the **weight vector** (or weights) and b as the **bias**. To summarize:

$$\textbf{Affine function: } f(\mathbf{x}) = \mathbf{x}^T\mathbf{w} + b, \quad \textbf{Parameters: } \quad (\text{Weights: } \mathbf{w},\ \text{Bias: } b)$$

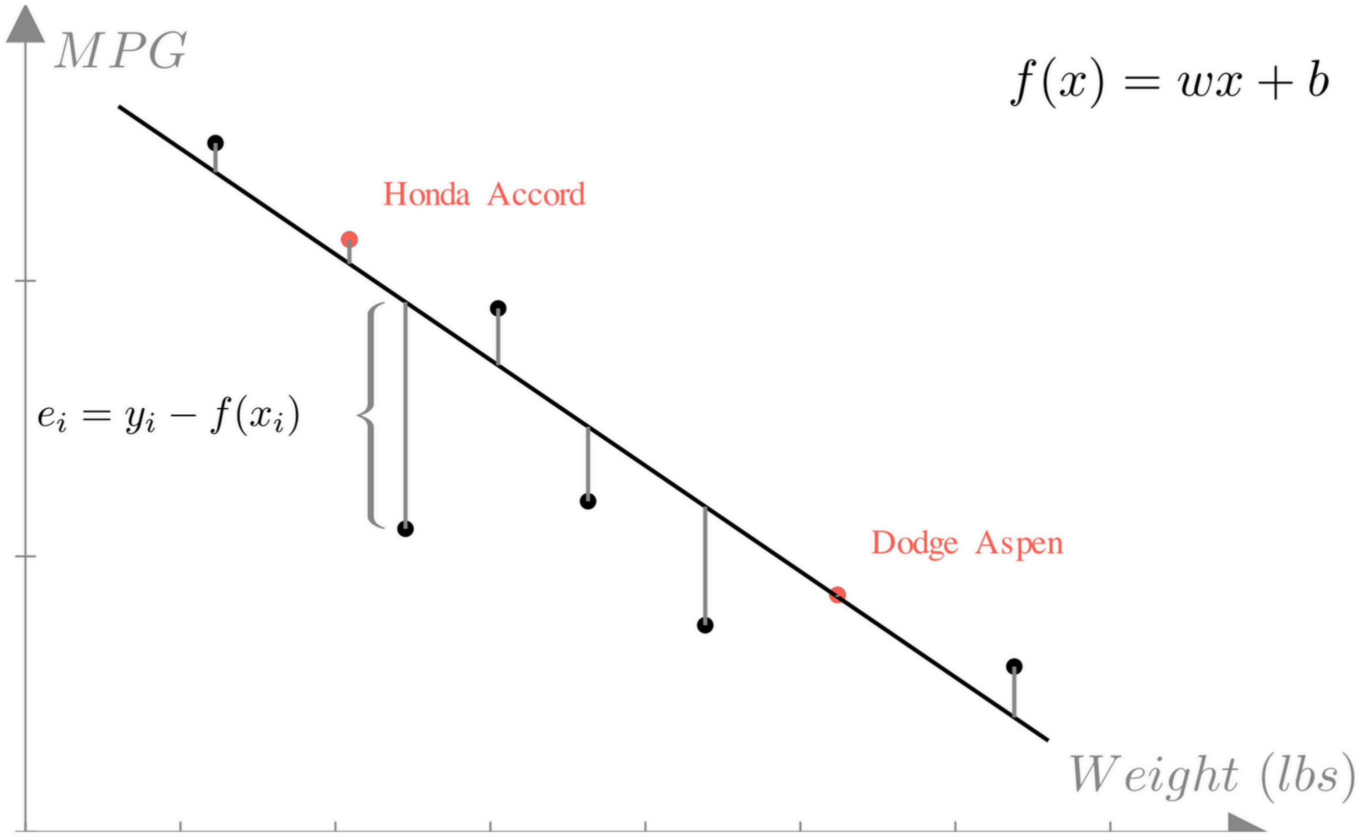
```
class Regression:
    def __init__(self, weights):
        self.weights = weights

    def predict(self, x):
        return np.dot(x, self.weights)
```



The **residual** or **error** of a prediction is the difference between the prediction and the true output:

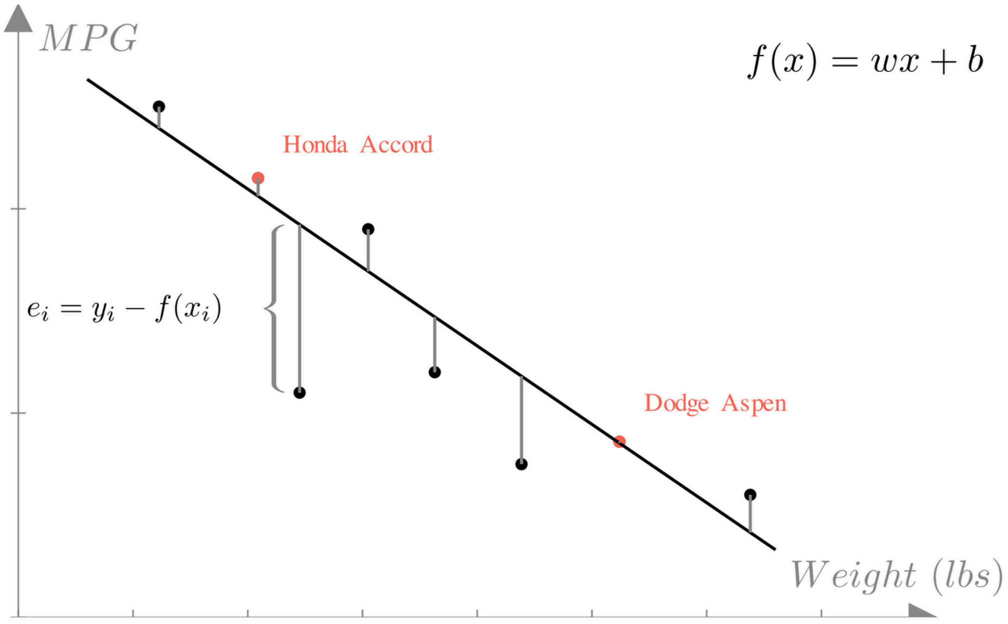
$$e_i = y_i - f(\mathbf{x}_i)$$



$$\mathcal{D} = \{(\mathbf{x}_1, y_1),\ (\mathbf{x}_2, y_2),\ \ldots\ (\mathbf{x}_N, y_N)\}$$

Mean Squared error

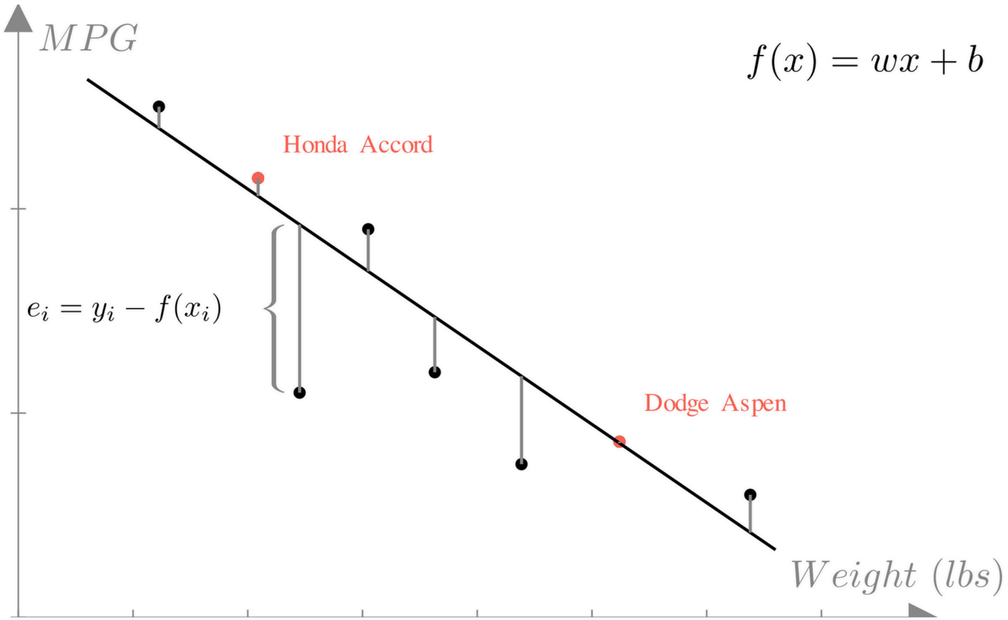
$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f(\mathbf{x}_i) - y_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^T\mathbf{w} - y_i)^2$$



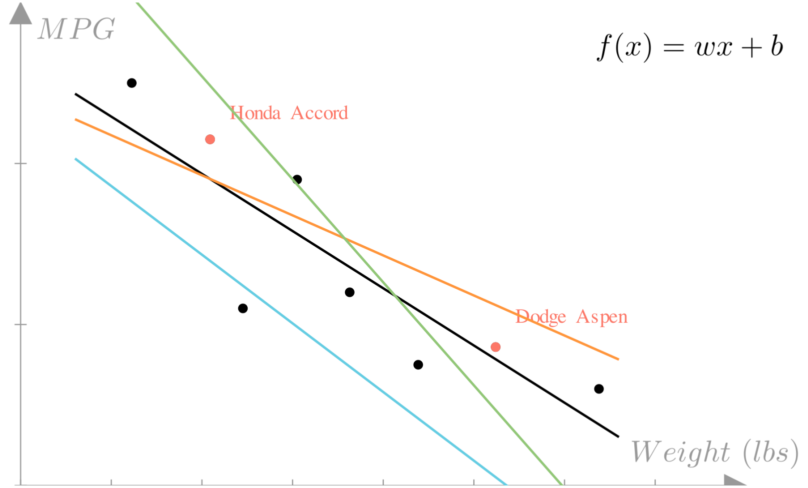
$$\mathcal{D} = \{(\mathbf{x}_1, y_1),\ (\mathbf{x}_2, y_2),\ \ldots\ (\mathbf{x}_N, y_N)\}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \ldots & x_{Nn} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

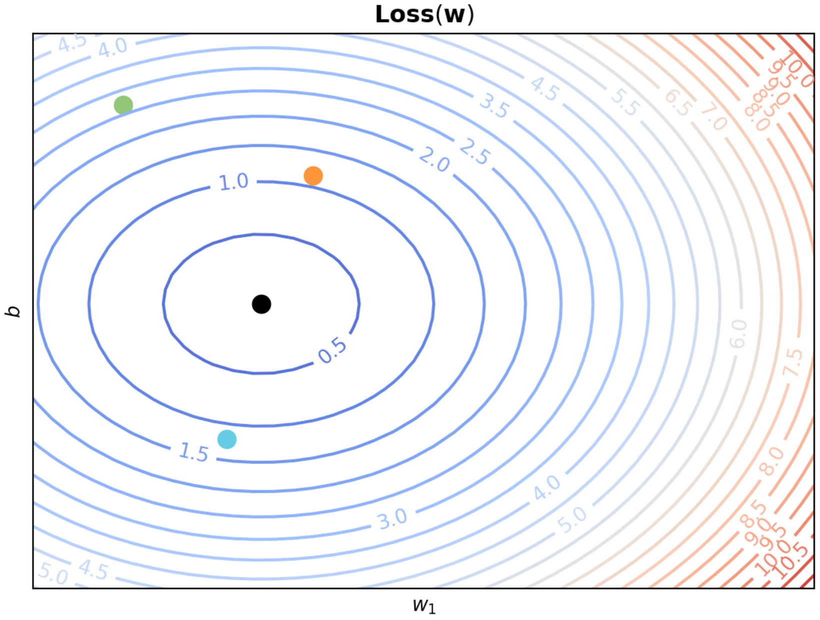
$$MSE(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i^T \mathbf{w} - y_i)^2$$



$$\mathbf{Loss}(\mathbf{w}) = MSE(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^T\mathbf{w} - y_i)^2$$

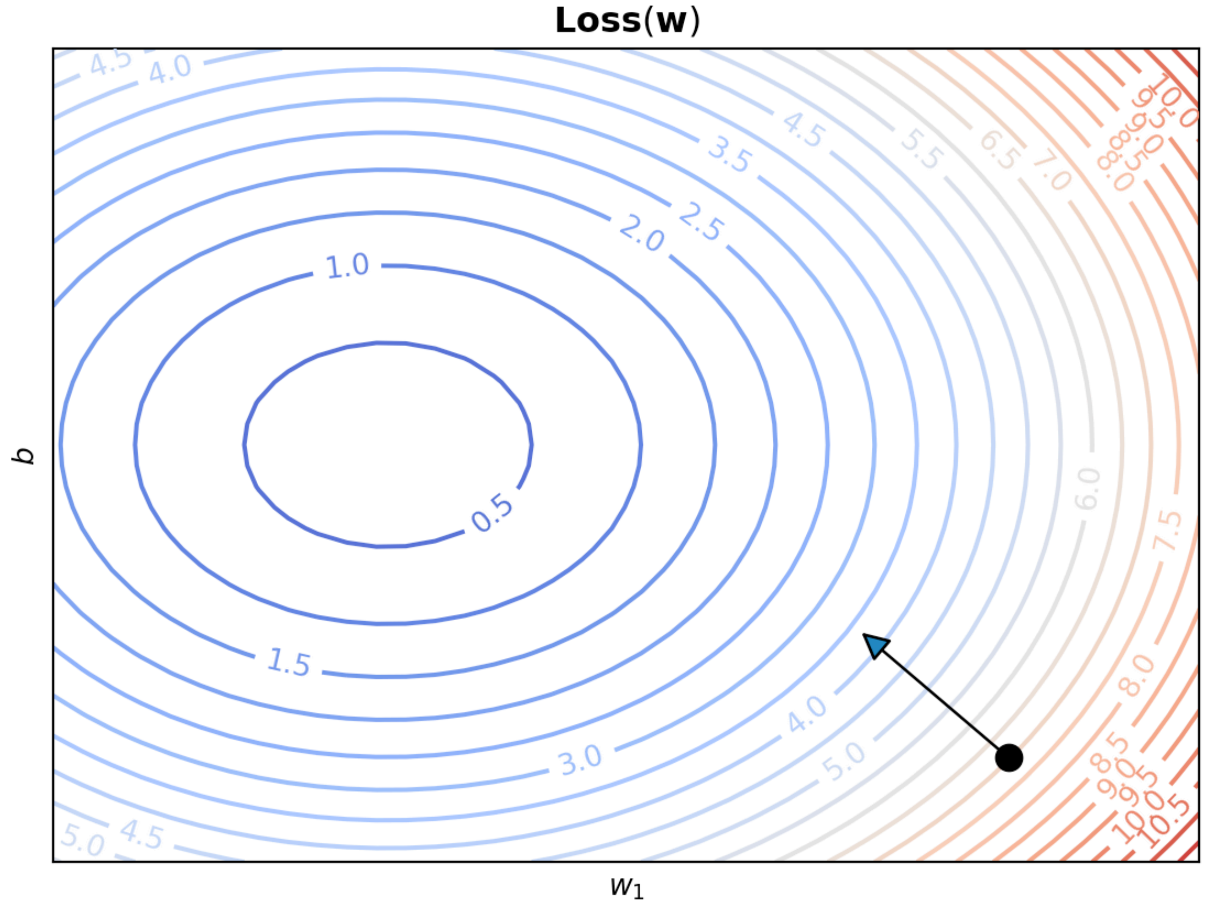


$$\mathbf{w}^* = \underset{\mathbf{w}}{\text{argmin}}\ \mathbf{Loss}(\mathbf{w}) = \underset{\mathbf{w}}{\text{argmin}}\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i^T\mathbf{w} - y_i)^2$$



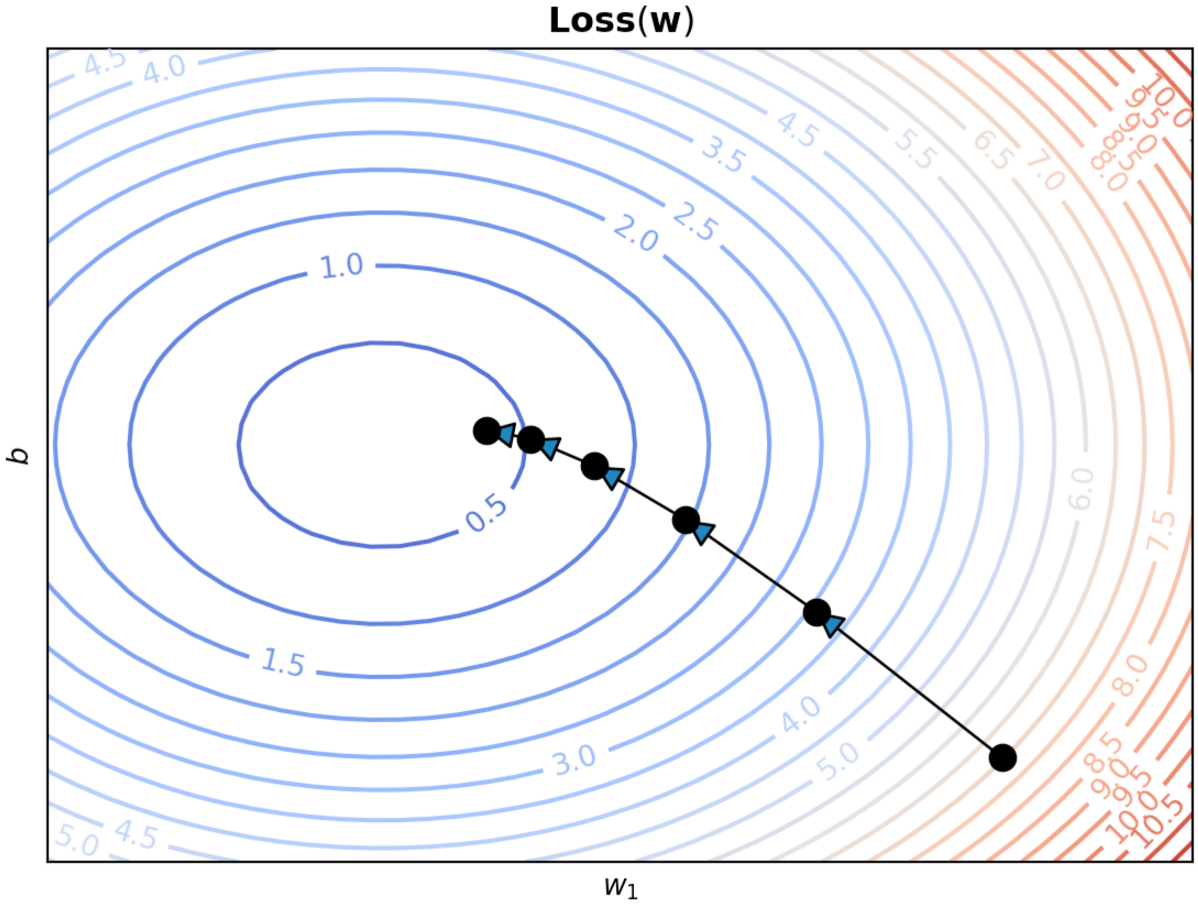
$$\text{Find: } \mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \, f(\mathbf{w})$$

$$\mathbf{w}^{(1)} \leftarrow \mathbf{w}^{(0)} + \mathbf{g}$$



Find: $\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\, f(\mathbf{w})$

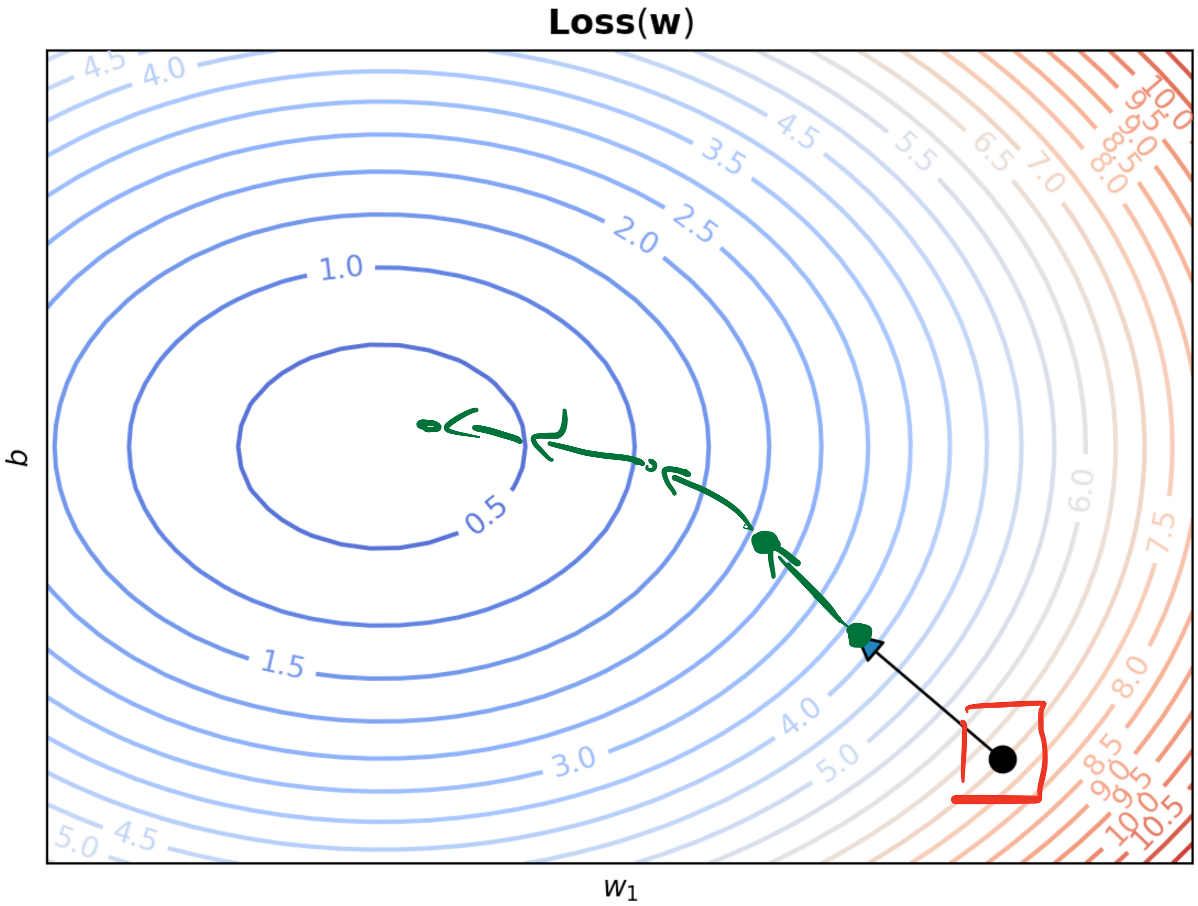
$$\mathbf{w}^{(1)} \leftarrow \mathbf{w}^{(0)} + \mathbf{g}$$



gradient descent

Find: $\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\, f(\mathbf{w})$

$$\mathbf{w}^{(1)} \leftarrow \mathbf{w}^{(0)} - \nabla f(\mathbf{w}^{(0)})$$



The **gradient** of a vector-input function is a vector such that each element is the partial derivative of the function with respect to the corresponding element of the input vector. We'll use the same notation as derivatives for gradients.

$$\frac{df}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \\ \vdots \end{bmatrix}$$ — gradient

$$\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$\frac{\partial}{\partial x_1}\mathbf{x}^T\mathbf{x} = \sum_{i=1}^{3} x_i x_i = \sum_{i=1}^{3} x_i^2 = x_1^2 + \cancel{x_2^2} + \cancel{x_3^2}$$

$$\frac{\partial}{\partial x_1} x^T x = 2x_1$$

The **gradient** of a vector-input function is a vector such that each element is the partial derivative of the function with respect to the corresponding element of the input vector. We'll use the same notation as derivatives for gradients.

$$\frac{df}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \\ \vdots \end{bmatrix}$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad \frac{\partial}{\partial x_1}\mathbf{x}^T\mathbf{x} = 2x_1 \qquad \frac{\partial}{\partial x_2} = 2x_2 \qquad \frac{\partial}{\partial x_3} = 2x_3$$

$$\frac{d}{d\mathbf{x}}\mathbf{x}^T\mathbf{x} = \begin{bmatrix} 2x_1 \\ 2x_2 \\ 2x_3 \end{bmatrix} = 2\mathbf{x}$$

The **gradient** of a vector-input function is a vector such that each element is the partial derivative of the function with respect to the corresponding element of the input vector. We'll use the same notation as derivatives for gradients.

$$\frac{df}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \\ \vdots \end{bmatrix}$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad \frac{\partial}{\partial x_1} f(\mathbf{x}^T\mathbf{x}) = f'(x^Tx)\left(\frac{\partial}{\partial x_1} x^Tx\right) = f'(x^Tx)(2x_1)$$

$$\frac{d}{d\mathbf{x}} f(\mathbf{x}^T\mathbf{x}) = \begin{bmatrix} f'(x^Tx)2x_1 \\ f'(x^Tx)2x_2 \\ f'(x^Tx)2x_3 \end{bmatrix} = f'(x^Tx)2\bar{\underline{x}}$$

The **gradient** of a vector-input function is a vector such that each element is the partial derivative of the function with respect to the corresponding element of the input vector. We'll use the same notation as derivatives for gradients.

$$\frac{df}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \\ \vdots \end{bmatrix}$$

$$\frac{\partial}{\partial x} \sum f(x) = \sum \frac{\partial}{\partial x} f(x)$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad \frac{\partial}{\partial x_1} \sum_{i=1}^{3} f(x_i) = \sum_{i=1}^{3} \frac{\partial}{\partial x_1} f(x_i) = \frac{\partial}{\partial x_1} f(x_1) = f'(x_1)$$

The **gradient** of a vector-input function is a vector such that each element is the partial derivative of the function with respect to the corresponding element of the input vector. We'll use the same notation as derivatives for gradients.

$$\frac{df}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \\ \vdots \end{bmatrix}$$

$$\frac{df}{d\mathbf{x}} = \underbrace{\nabla f(\mathbf{x})}_{\nabla f(x)}$$

$$\frac{df}{d\mathbf{x}} = \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), \quad \frac{df}{d\mathbf{y}} = \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$$

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

Pred.

$$f(x) = x^T w$$

$$= \frac{1}{N}\frac{d}{dw}\left(\sum_{i=1}^{N}(x_i^T w - y_i)^2\right) = \frac{1}{N}\sum_{i=1}^{N}\frac{d}{dw}(x_i^T w - y_i)^2$$

$$= \frac{2}{N}\sum_{i=1}^{N}(x_i^T w - y_i)\frac{d}{dw}(x_i^T w - y_i)$$

$$= \frac{2}{N}\sum_{i=1}^{N}(x_i^T w - y_i)\, x_i^T$$

$$f(x) = x^2$$

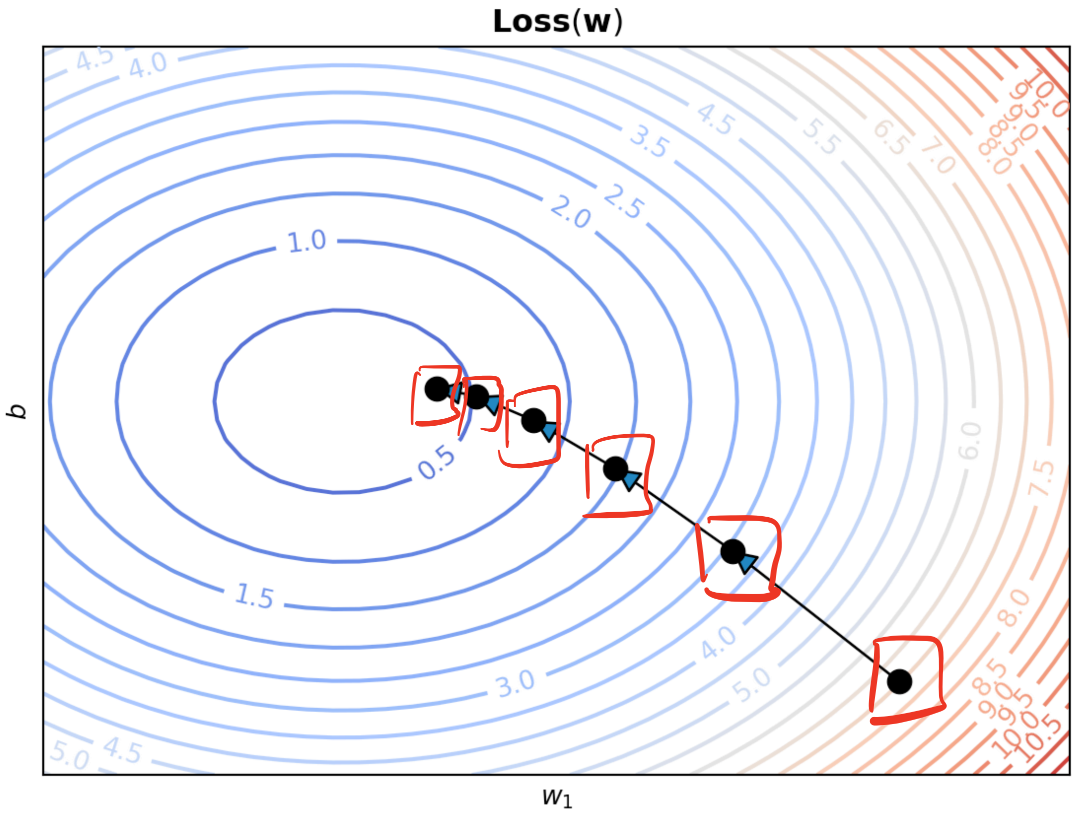
$$f'(x) = 2x$$

$$\frac{d}{dw_1}x_i^T w = \frac{d}{dw_i}\left(\sum_{j=1}^{c} x_{ic} w_c\right) = x_{ic} \qquad \frac{d}{dw}x_i^T w = x_i$$

$$\text{Find: } \mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\ f(\mathbf{w})$$

$$\mathbf{w}^{(1)} \leftarrow \mathbf{w}^{(0)} - \nabla f(\mathbf{w}^{(0)})$$

$\nabla f(w^*) = 0$



Recall that it’s minimum value $\mathbf{w}^*$, a function f *must* have a gradient of $\mathbf{0}$.

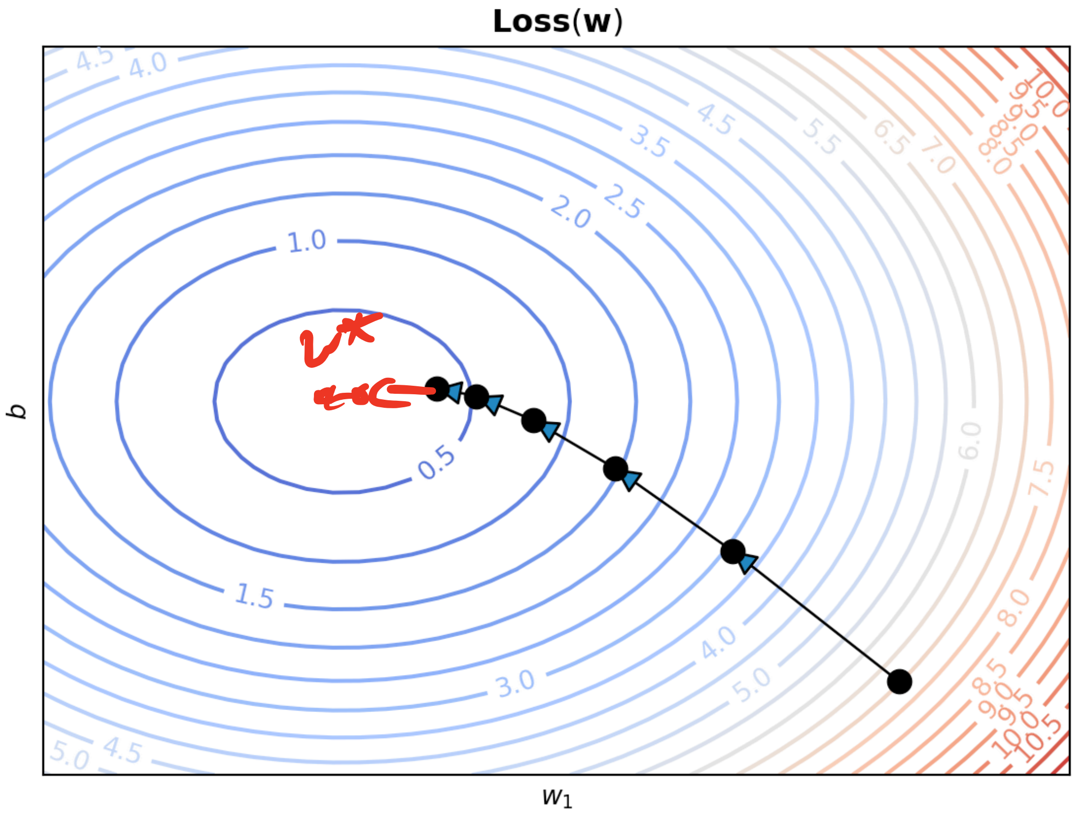
$$\nabla f(\mathbf{w}^*) = \mathbf{0}$$

It follows that:

$$\mathbf{w}^* = \mathbf{w}^* - \nabla f(\mathbf{w}^*)$$

Gradient descent

$$\text{While } \nabla f(\mathbf{w}^{(i)}) \neq \mathbf{0}: \quad \mathbf{w}^{(i+1)} \leftarrow \mathbf{w}^{(i)} - \nabla f(\mathbf{w}^{(i)})$$



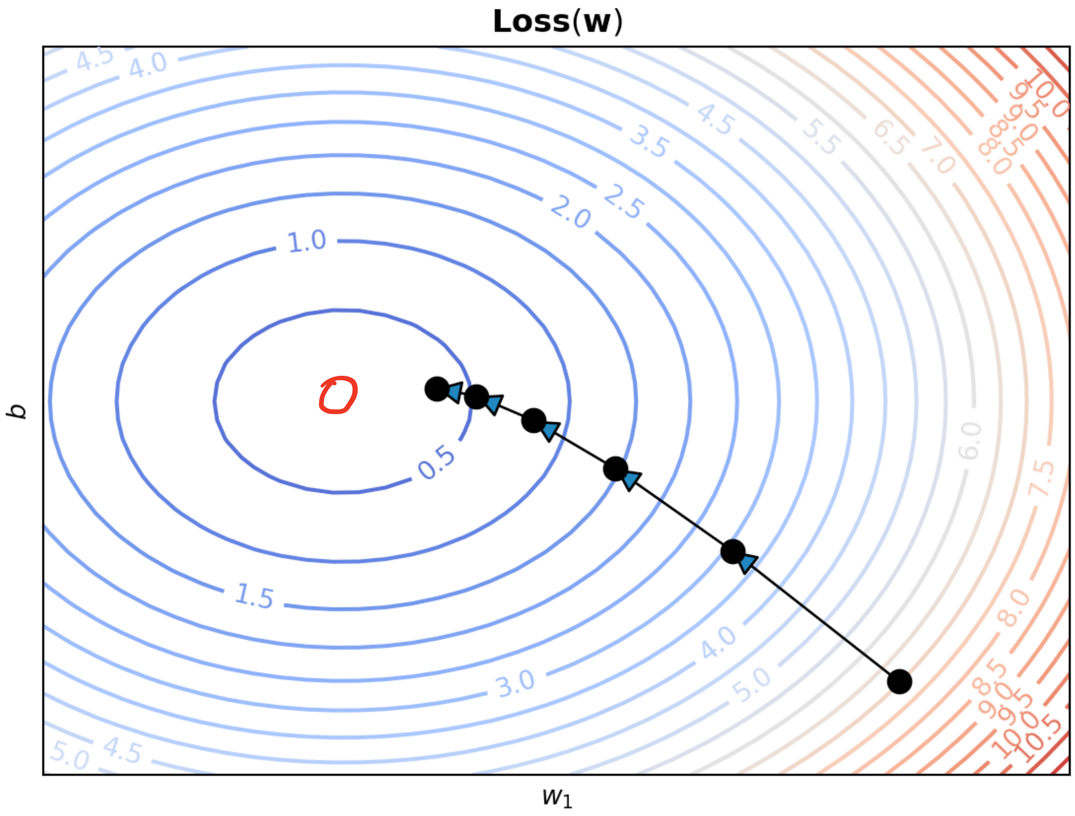
Recall that it's minimum value $\mathbf{w}^*$, a function f *must* have a gradient of $\mathbf{0}$.

$$\nabla f(\mathbf{w}^*) = \mathbf{0}$$

It follows that:

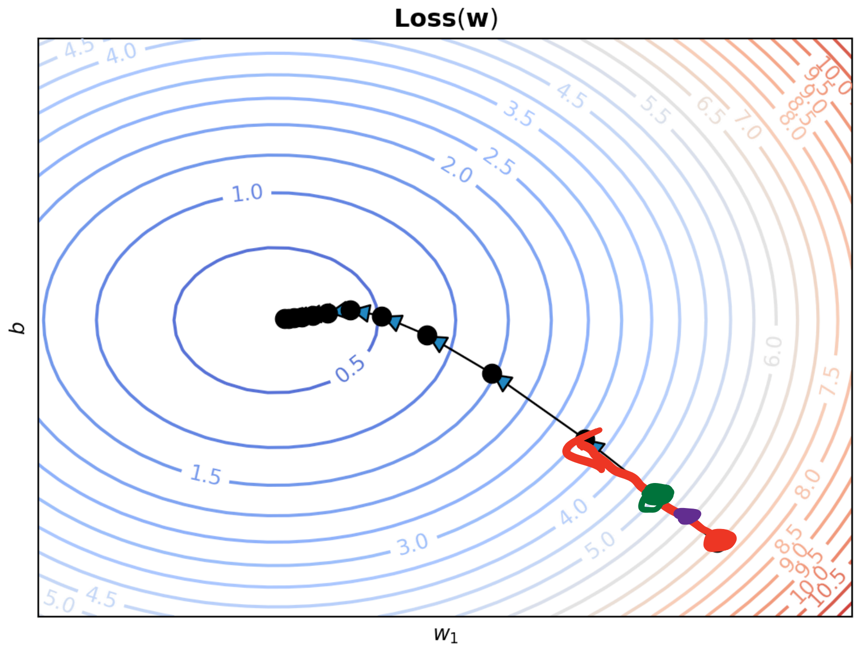
$$\mathbf{w}^* = \mathbf{w}^* - \nabla f(\mathbf{w}^*)$$

$$\text{While } ||\nabla f(\mathbf{w}^{(i)})||_2 > \epsilon : \quad \mathbf{w}^{(i+1)} \leftarrow \mathbf{w}^{(i)} - \nabla f(\mathbf{w}^{(i)})$$



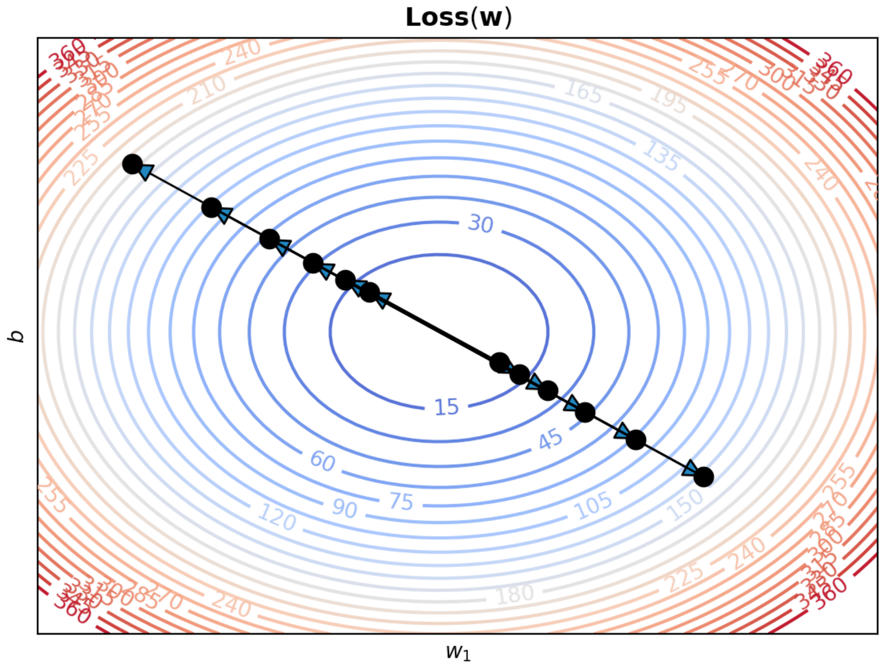
$$w^{(i+1)} \leftarrow w^{(i)} - \nabla f(w)$$

$$w^{(i+1)} \leftarrow w^{(i)} - \underset{\substack{\uparrow \\ \text{stepsize}}}{\alpha} \nabla f(w)$$

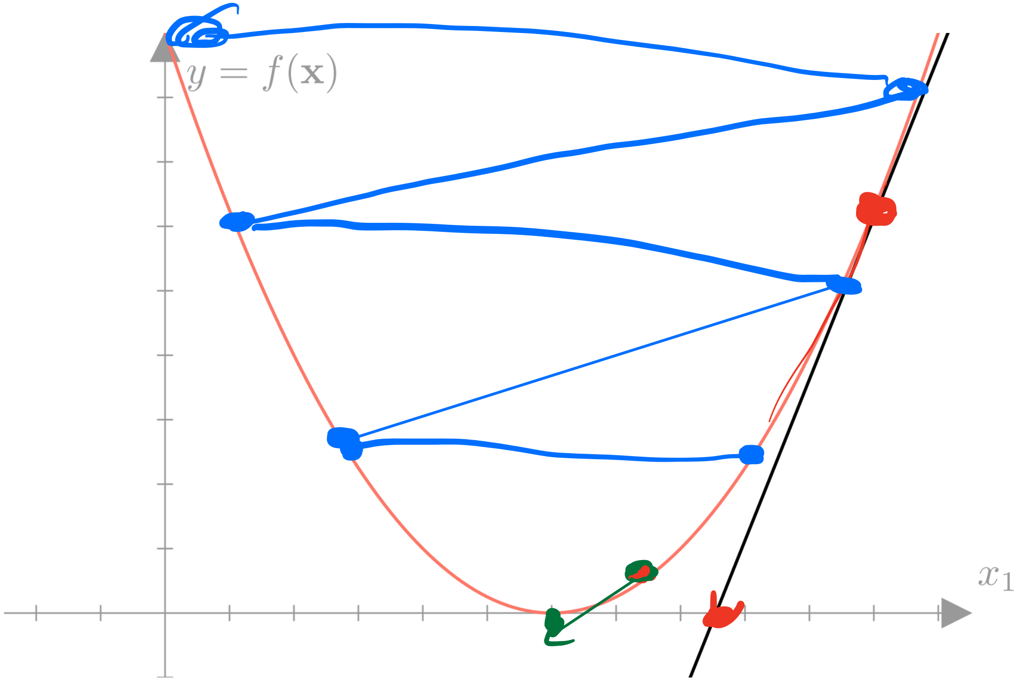


$\alpha = .5$

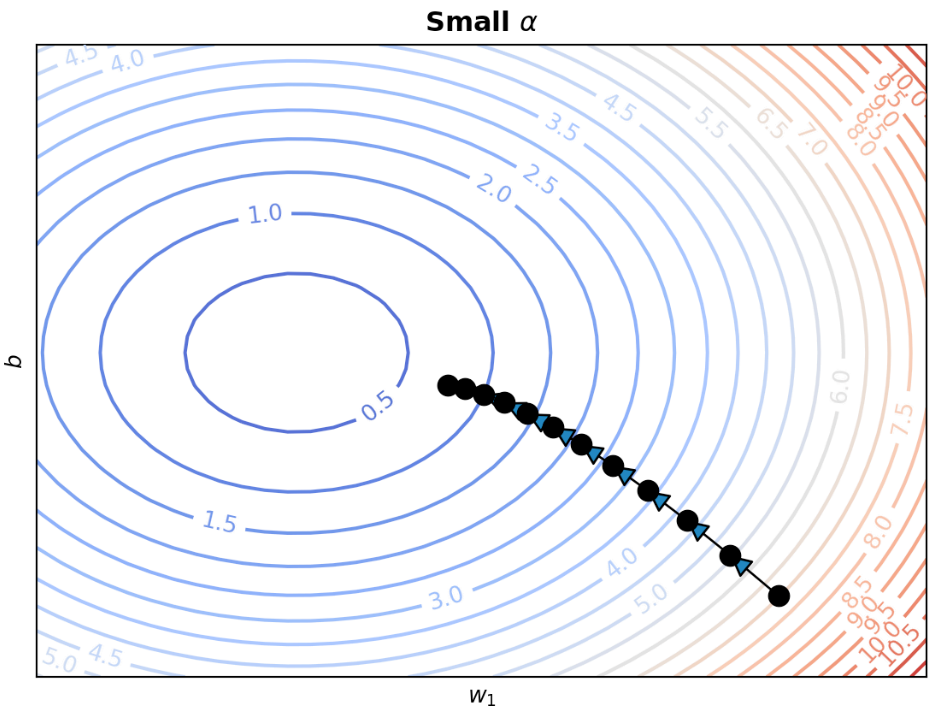
$\alpha = .25$

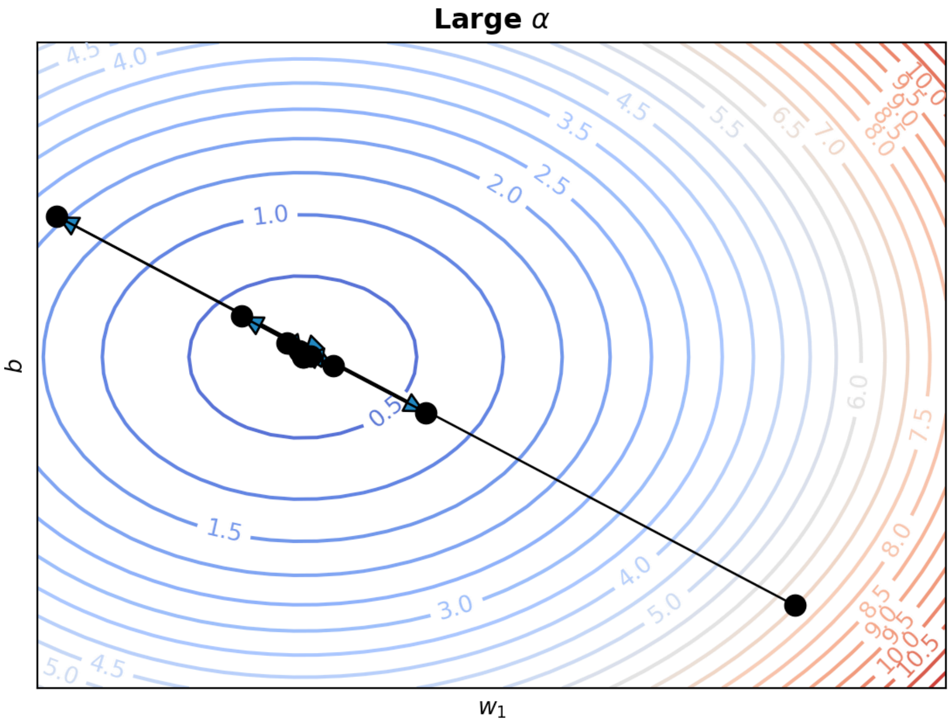


$$\frac{df}{d\mathbf{w}} = \lim_{\gamma \to 0} \max_{\|\epsilon\|_2 < \gamma} \frac{f(\mathbf{w} + \epsilon) - f(\mathbf{w})}{\|\epsilon\|_2}$$



$$\mathbf{w}^{(i+1)} \leftarrow \mathbf{w}^{(i)} - \alpha \nabla f(\mathbf{w}^{(i)})$$





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

$$= \frac{2}{N} \sum_{i=1}^{N} (\mathbf{x}_i^T \mathbf{w} - y_i) \mathbf{x}_i$$

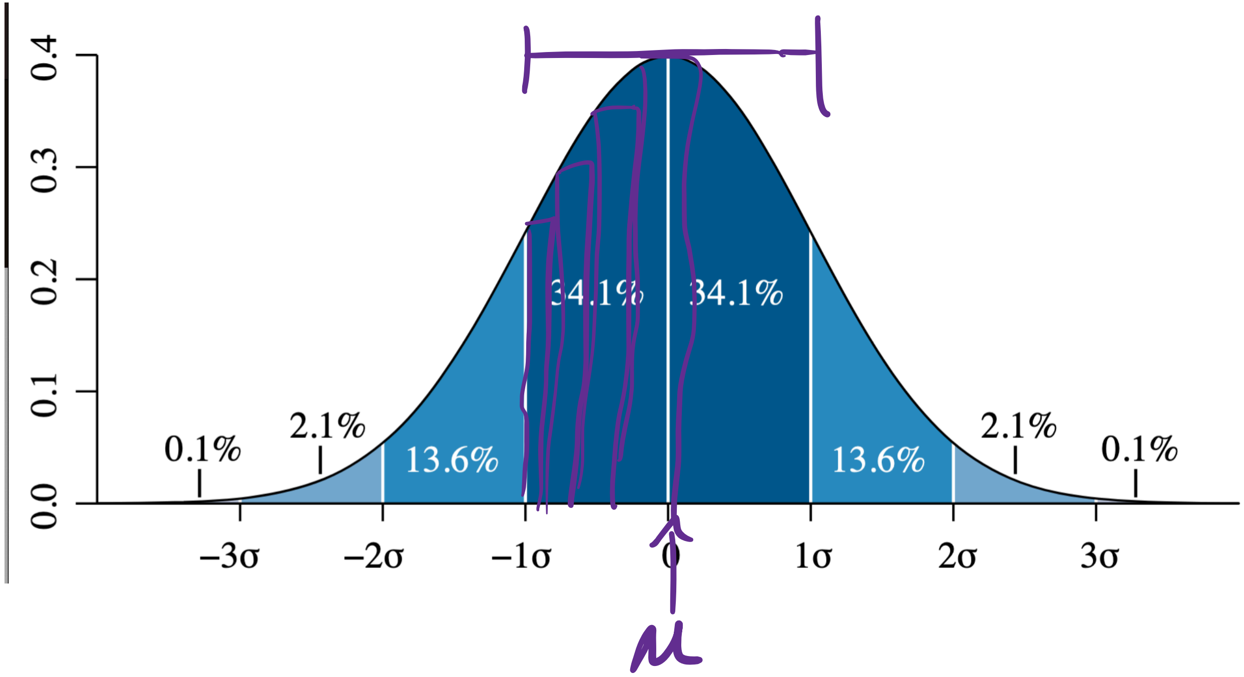
With this gradient our gradient descent update becomes:

$$\mathbf{w}^{(i+1)} \longleftarrow \mathbf{w}^{(i)} - \alpha \left(\frac{2}{N} \right) \sum_{i=1}^{N} (\mathbf{x}_i^T \mathbf{w}^{(i)} - y_i) \mathbf{x}_i$$

$\exp(x) = e^x$

Normal distribution

$$p(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\underbrace{-\frac{1}{2\sigma^2}(y-\mu)^2} \right)$$

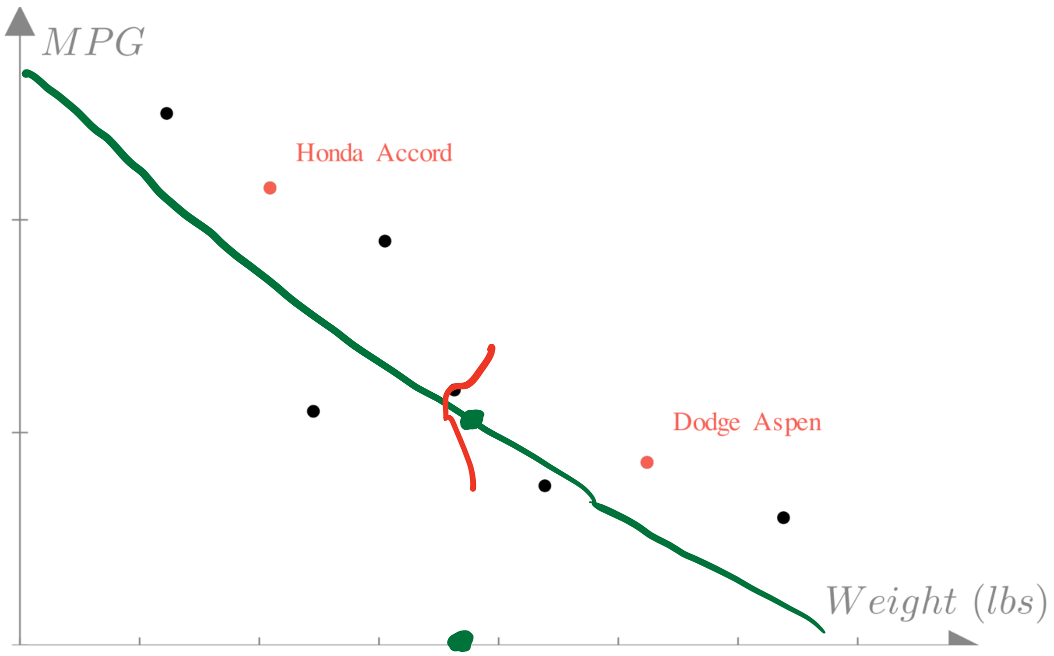


$$y_i \sim \mathcal{N}\left(\mathbf{x}_i^T \mathbf{w},\ \sigma^2\right)$$

$$p(y_i \mid \mathbf{x}_i, \mathbf{w}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T \mathbf{w})^2\right)$$

m

$\mu = x^T w$

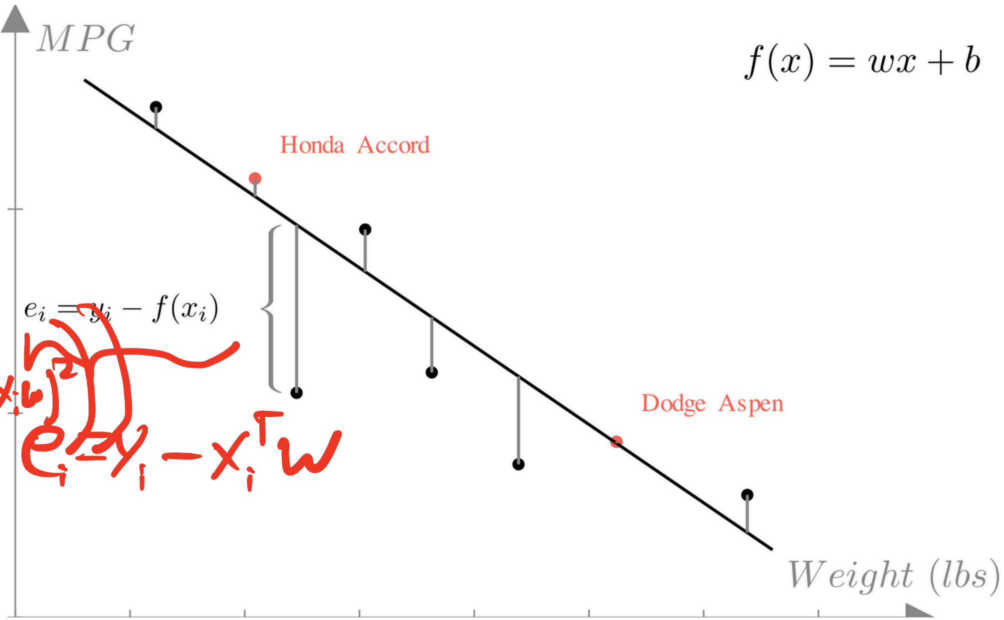


$$y_i \sim \mathcal{N}\left(\mathbf{x}_i^T\mathbf{w},\ \sigma^2\right)$$

$$p(y_i \mid \mathbf{x}_i, \mathbf{w}) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^T\mathbf{w})^2\right)$$

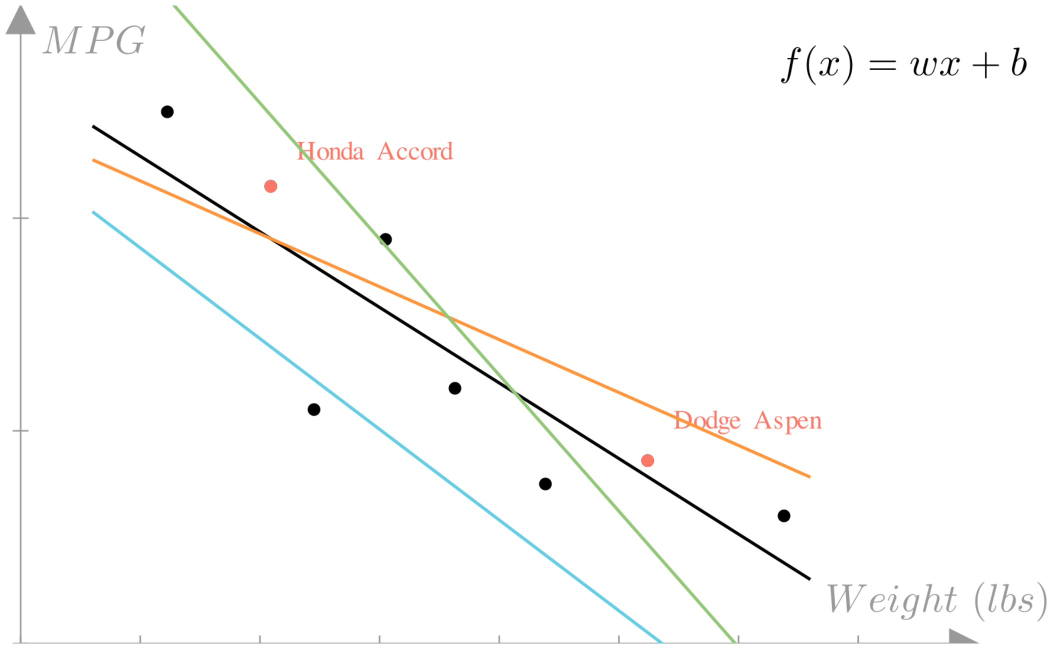
$$P(e_i \mid x_i, w) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}\left(e_i - (y_i - x_i w)\right)^2\right)$$

$$e_i = y_i - x_i^T w$$



Maximum likelihood

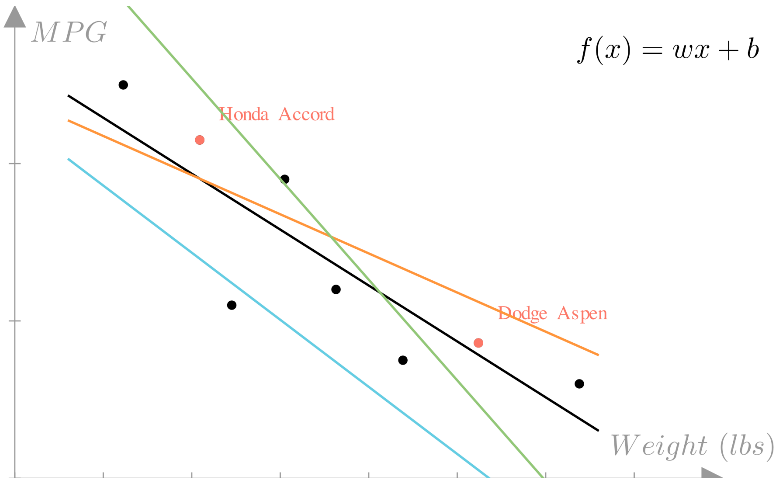
$$\mathbf{w}^* = \underset{\mathbf{w}}{\text{argmax}}\; p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \underset{\mathbf{w}}{\text{argmax}}\; p(y_1, \ldots, y_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w})$$



Maximum likelihood

$$\mathbf{w}^* = \underset{\mathbf{w}}{\text{argmax}}\, p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \underset{\mathbf{w}}{\text{argmax}}\, p(y_1, \ldots, y_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w})$$

$$p(y_1, \ldots, y_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{w}) = \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, \mathbf{w})$$



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

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

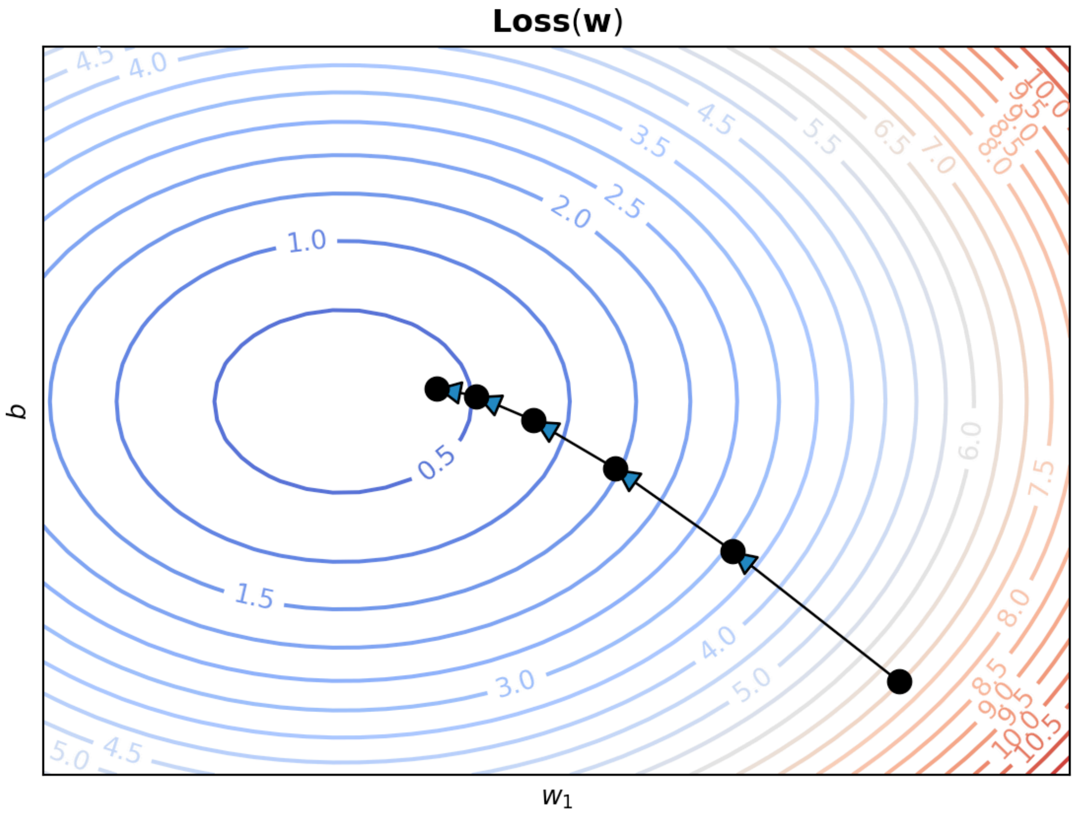
Recall that it's minimum value $\mathbf{w}^*$, a function f *must* have a gradient of $\mathbf{0}$.

$$\nabla f(\mathbf{w}^*) = \mathbf{0}$$

It follows that:

$$\mathbf{w}^* = \mathbf{w}^* - \nabla f(\mathbf{w}^*)$$

$$\text{While } ||\nabla f(\mathbf{w}^{(i)})||_2 > \epsilon : \quad \mathbf{w}^{(i+1)} \leftarrow \mathbf{w}^{(i)} - \nabla f(\mathbf{w}^{(i)})$$



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

$$p(y_i \mid \mathbf{x}_i, \mathbf{w}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} (y_i - \mathbf{x}_i^T \mathbf{w})^2 \right)$$

$$\mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = -\sum_{i=1}^{N} \log \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} (y_i - \mathbf{x}_i^T \mathbf{w})^2 \right) \right]$$

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

$$\mathbf{w}^* = \underset{\mathbf{w}}{\text{argmin}}\ \mathbf{Loss}(\mathbf{w})$$

$$\nabla_{\mathbf{w}} \mathbf{MSE}(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \frac{2}{N} \sum_{i=1}^{N} (\mathbf{x}_i^T \mathbf{w} - y_i)\mathbf{x}_i$$

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

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbf{Loss}(\mathbf{w})$$

$$\nabla_{\mathbf{w}} \mathbf{MSE}(\mathbf{w}, \mathbf{X}, \mathbf{y}) \quad = \frac{2}{N} \sum_{i=1}^{N} (\mathbf{x}_i^T \mathbf{w} - y_i) \mathbf{x}_i$$

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

$$\mathbf{0} = \left(\frac{2}{N} \right) \sum_{i=1}^{N} (\mathbf{x}_i^T \mathbf{w} - y_i) \mathbf{x}_i$$

$$\underset{\mathbf{w}}{\operatorname{argmin}}\, MSE(\mathbf{w}, \mathbf{X}, \mathbf{y}) = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbf{NLL}(\mathbf{w}, \mathbf{X}, \mathbf{y})$$