A Walk Downhill

Ananda Biswas

ami.ananda@bhu.ac.in

Acknowledgements

- CS7015 Lectures[a] of Professor Mitesh Khapra, IIT Madras
- Videos[b] of Ryan Harris on Backpropagation

[a] YouTube
[b] YouTube

Target : Learning Parameters

$\boxed{\texttt{Supervised Learning}}$

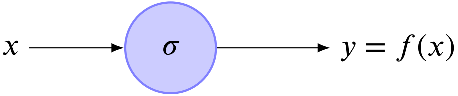
Target : Learning Parameters

$\boxed{\texttt{Supervised Learning}} \longrightarrow \boxed{\texttt{Loss}\ \mathscr{L}(\boldsymbol{\theta})}$

Target : Learning Parameters

$$\boxed{\texttt{Supervised Learning}} \longrightarrow \boxed{\texttt{Loss } \mathscr{L}(\boldsymbol{\theta})} \longrightarrow \boxed{\texttt{optimal } \boldsymbol{\theta}}$$

Set-up



$$f(x) = \frac{1}{1 + e^{-(wx+b)}}$$

Input for Training

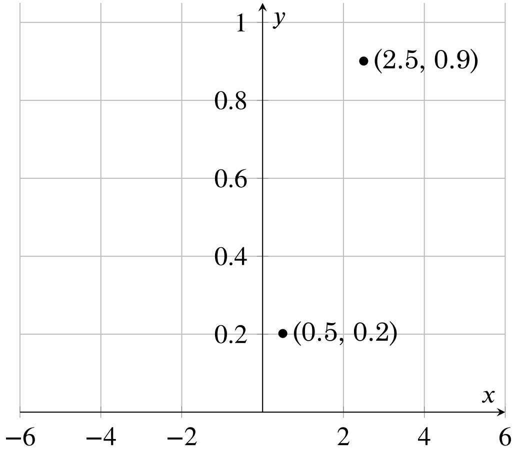
$(x_i, y_i)_{i=1}^{n} \rightarrow n$ pairs of (x, y)

Training Objective

Find w and b that minimizes

$$\mathscr{L}(w, b) = \sum_{i=1}^{n} \left(y_i - f(x_i)^2\right)$$

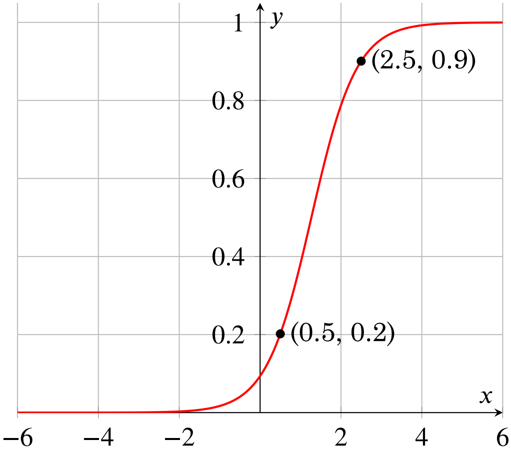
Set-up



What is training ?

At the end of the process we wish to get w^* and b^* so that $f(0.5) \to 0.2$ and $f(2.5) \to 0.9$.

Set-up

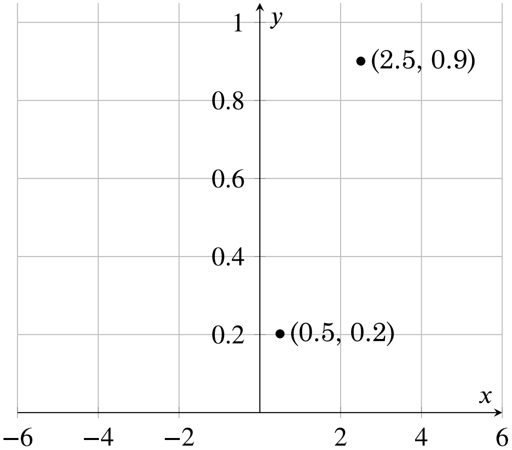


What is training ?

At the end of the process we wish to get w^* and b^* so that $f(0.5) \rightarrow 0.2$ and $f(2.5) \rightarrow 0.9$.

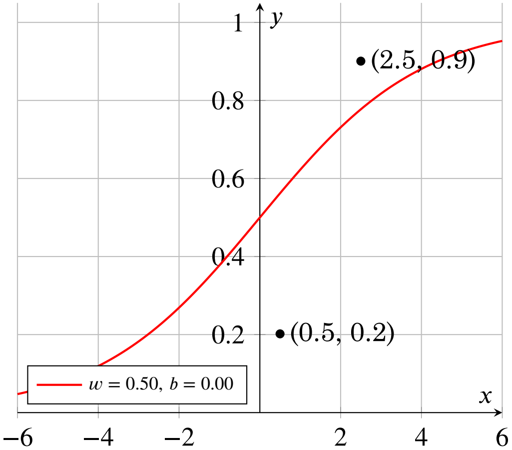
In other words, we hope to find a sigmoid function that satisfies $(0.5, 0.2)$ and $(2.5, 0.9)$.

Approach 1 : Guess Work



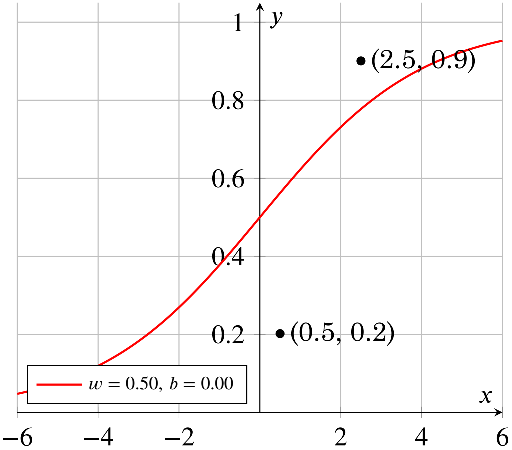
Let us try a random guess..

Approach 1 : Guess Work



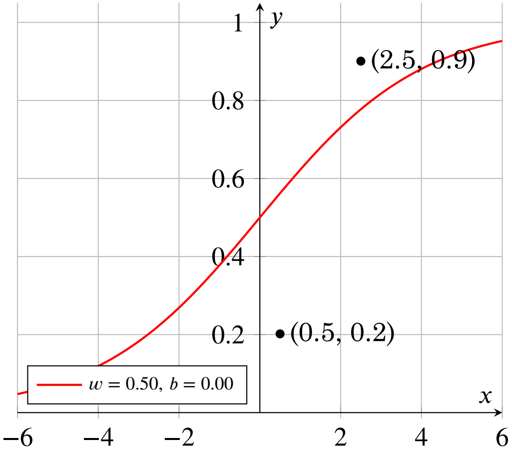
Let us try a random guess..
say, $w = 0.5, b = 0$

Approach 1 : Guess Work



Let us try a random guess..
say, $w = 0.5, b = 0$
Clearly not good, but how bad is it ?

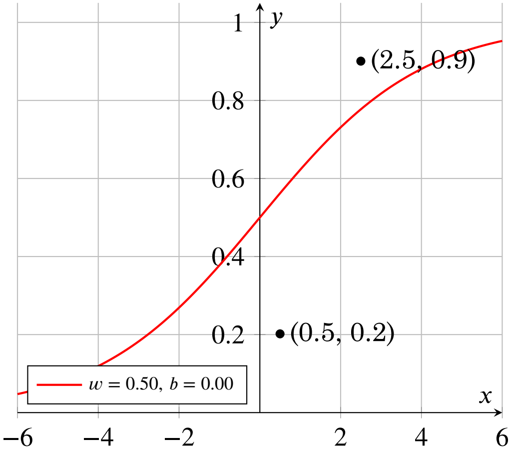
Approach 1 : Guess Work



Let us try a random guess..
say, $w = 0.5, b = 0$
Clearly not good, but how bad is it ?

$\mathscr{L}(w, b) = \sum_{i=1}^{n} \left(y_i - f(x_i)^2\right)$ will tell us.

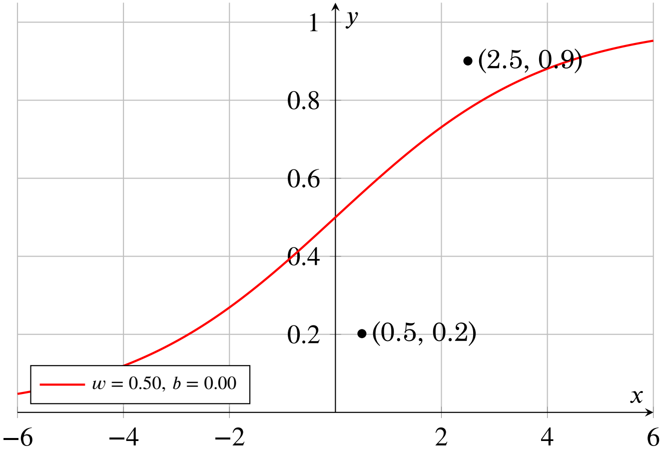
Approach 1 : Guess Work



$$\mathscr{L}(w,b) = \sum_{i=1}^{n} \left(y_i - f(x_i)^2\right)$$
$$= \frac{1}{2}\left[\left(y_1 - f(x_1)\right)^2 + \left(y_2 - f(x_2)\right)^2\right]$$
$$= \frac{1}{2}\left[(0.2 - f(0.5))^2 + (0.9 - f(0.2))^2\right]$$
$$= 0.073$$

We want $\mathscr{L}(w,b) = \sum_{i=1}^{n} \left(y_i - f(x_i)^2\right)$ to as close to 0 as possible.

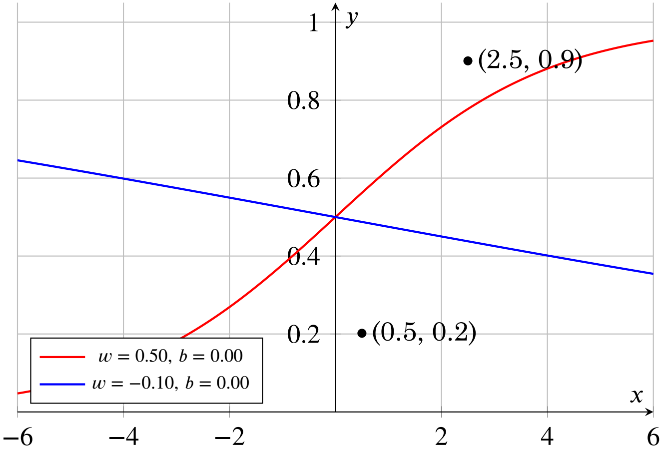
Approach 1 : Guess Work



We try with some other values of w and b.

w	b	$\mathscr{L}(w, b)$
0.50	0.00	0.0730

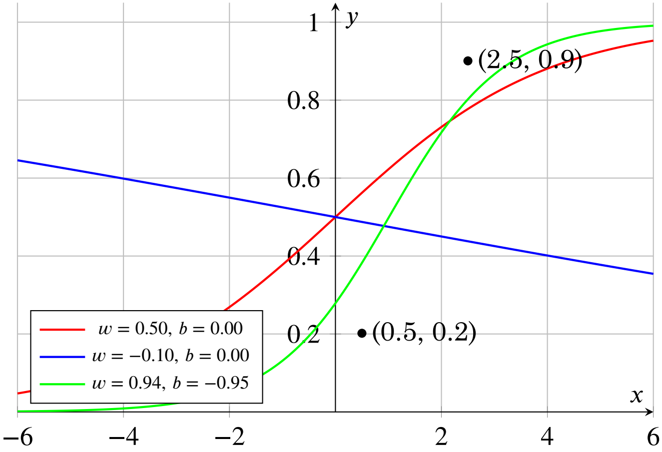
Approach 1 : Guess Work



We try with some other values of w and b.

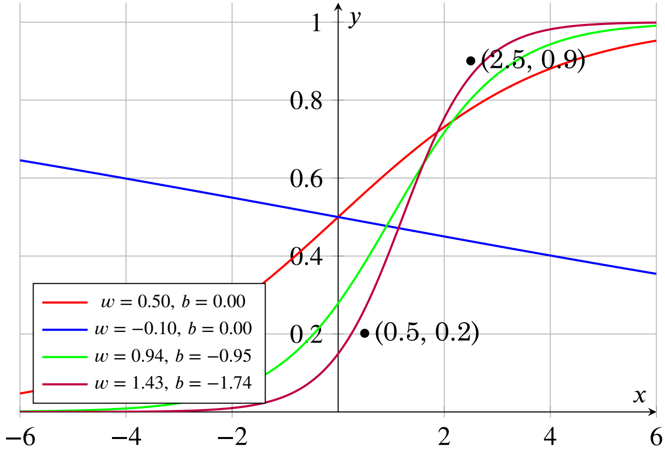
w	b	$\mathscr{L}(w, b)$
0.50	0.00	0.0730
−0.10	0.00	0.1481

Approach 1 : Guess Work



We try with some other values of w and b.

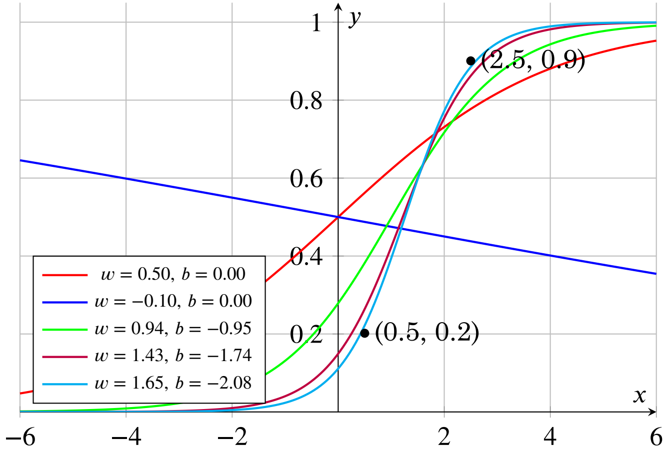
w	b	$\mathscr{L}(w,b)$
0.50	0.00	0.0730
−0.10	0.00	0.1481
0.94	−0.94	0.0214



We try with some other values of w and b.

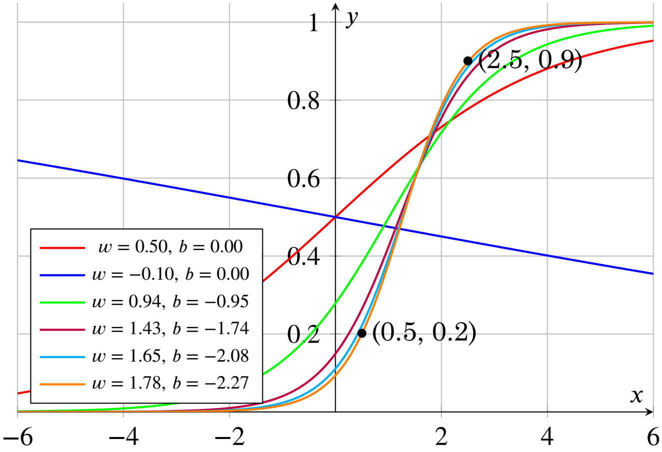
w	b	$\mathscr{L}(w, b)$
0.50	0.00	0.0730
−0.10	0.00	0.1481
0.94	−0.94	0.0214
1.42	−1.73	0.0028

Approach 1 : Guess Work



We try with some other values of w and b.

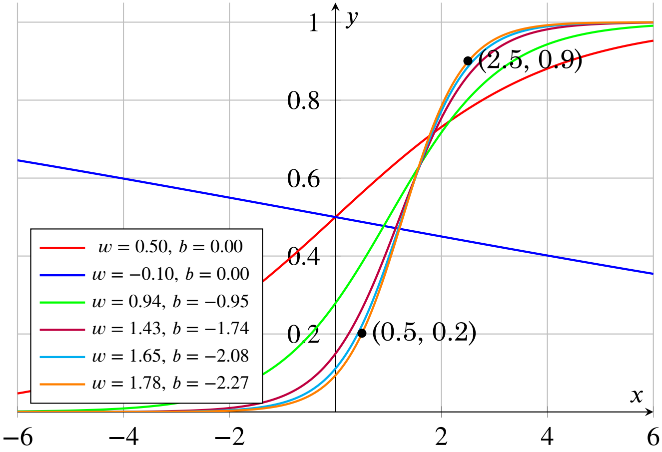
w	b	$\mathscr{L}(w,b)$
0.50	0.00	0.0730
−0.10	0.00	0.1481
0.94	−0.94	0.0214
1.42	−1.73	0.0028
1.65	−2.08	0.0003



We try with some other values of w and b.

w	b	$\mathscr{L}(w, b)$
0.50	0.00	0.0730
−0.10	0.00	0.1481
0.94	−0.94	0.0214
1.42	−1.73	0.0028
1.65	−2.08	0.0003
1.78	−2.27	0.0000

Approach 1 : Guess Work

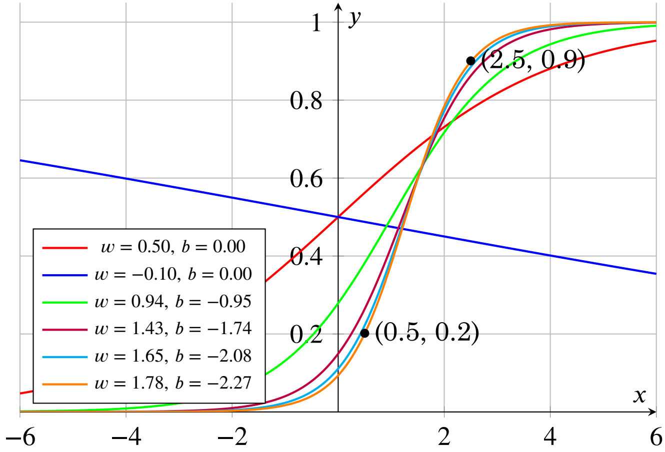


We try with some other values of w and b.

w	b	$\mathscr{L}(w, b)$
0.50	0.00	0.0730
−0.10	0.00	0.1481
0.94	−0.94	0.0214
1.42	−1.73	0.0028
1.65	−2.08	0.0003
1.78	−2.27	0.0000

Job done ! But

Approach 1 : Guess Work



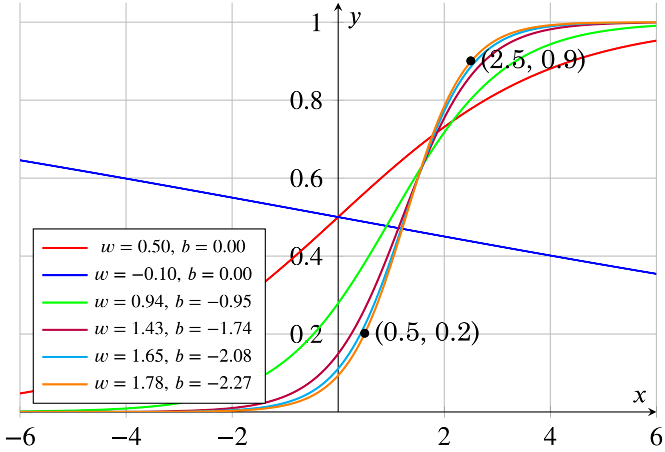
We try with some other values of w and b.

w	b	$\mathscr{L}(w,b)$
0.50	0.00	0.0730
−0.10	0.00	0.1481
0.94	−0.94	0.0214
1.42	−1.73	0.0028
1.65	−2.08	0.0003
1.78	−2.27	0.0000

Job done ! But

- Infeasible

Approach 1 : Guess Work



We try with some other values of w and b.

w	b	$\mathscr{L}(w, b)$
0.50	0.00	0.0730
−0.10	0.00	0.1481
0.94	−0.94	0.0214
1.42	−1.73	0.0028
1.65	−2.08	0.0003
1.78	−2.27	0.0000

Job done ! But

- Infeasible
- Does not gurantee correct solution

Approach 2 : Brute–force Search

$\boxed{\texttt{Compute } \mathscr{L}(w, b) \texttt{ for all possible } w \texttt{ and } b}$

Approach 2 : Brute–force Search

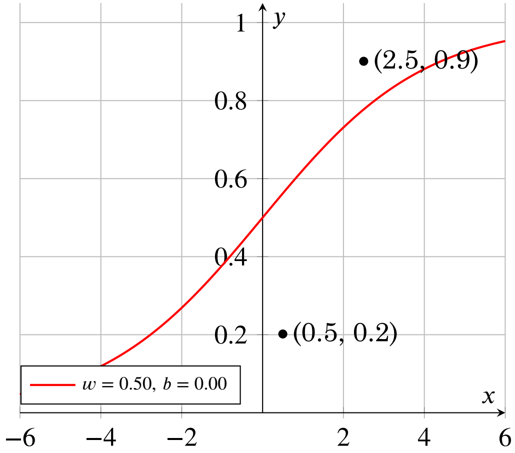
Compute $\mathscr{L}(w, b)$ for all possible w and b → Choose the best w and b

Approach 2 : Brute–force Search

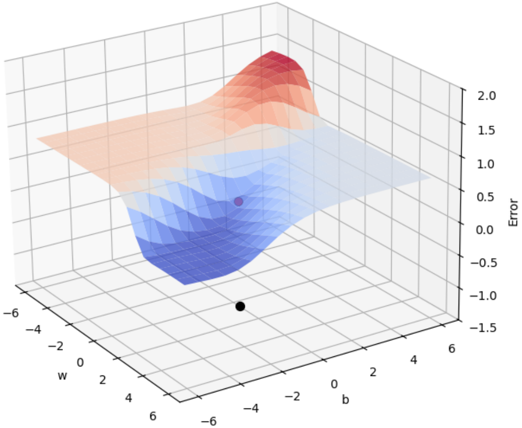
Compute $\mathscr{L}(w, b)$ for all possible w and b → Choose the best w and b

- Computationally infeasible

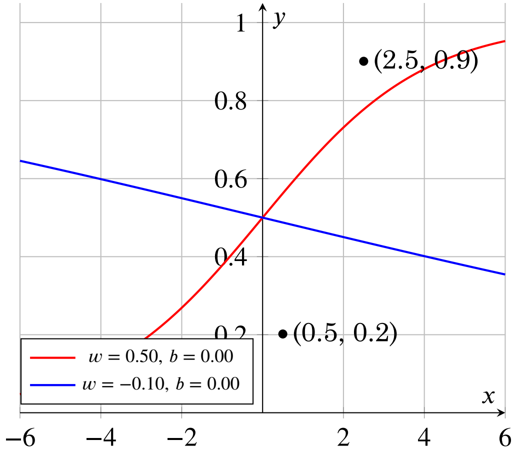
Revisit to Approach 1



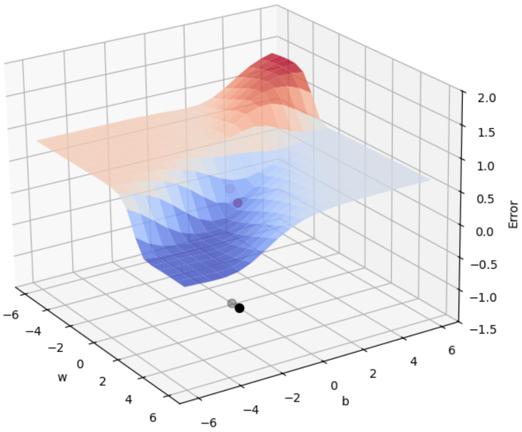
What we were actually doing !



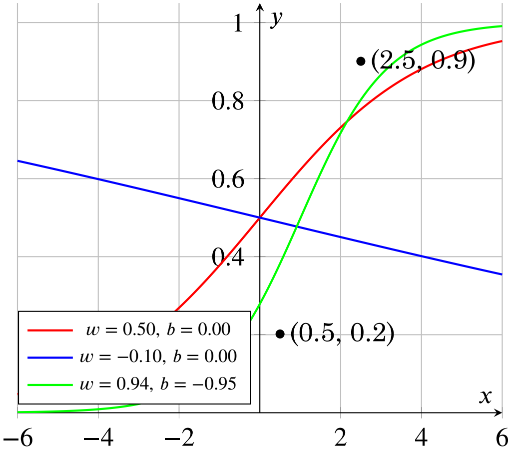
Revisit to Approach 1



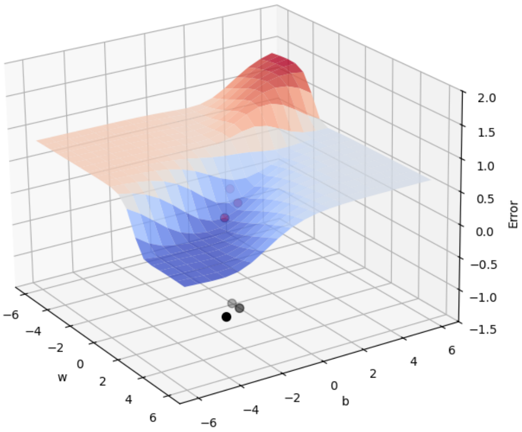
What we were actually doing !



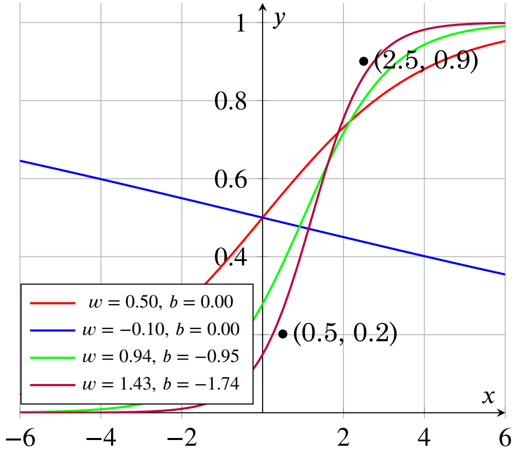
Revisit to Approach 1



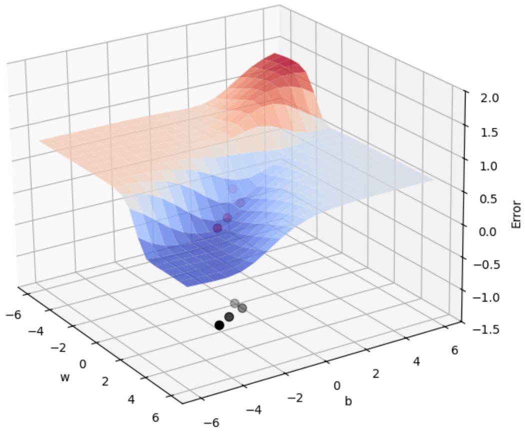
What we were actually doing !

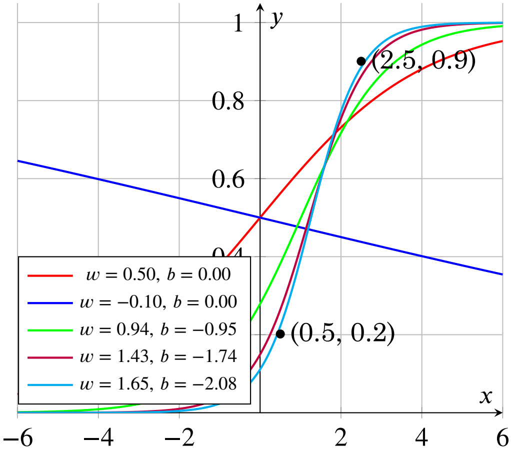


Revisit to Approach 1

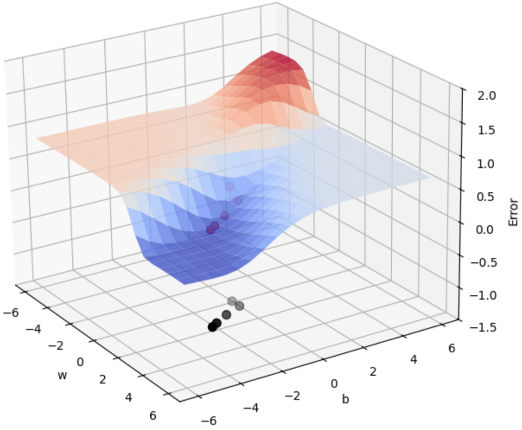


What we were actually doing !

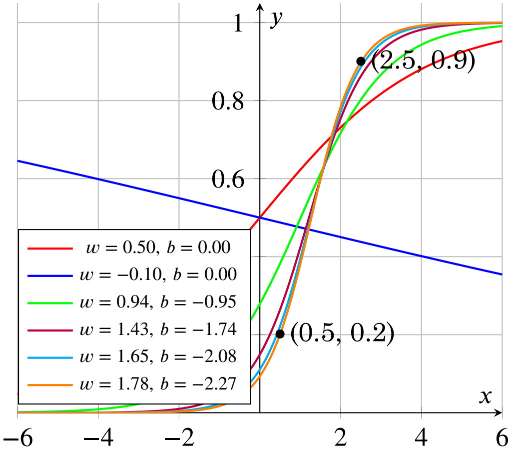




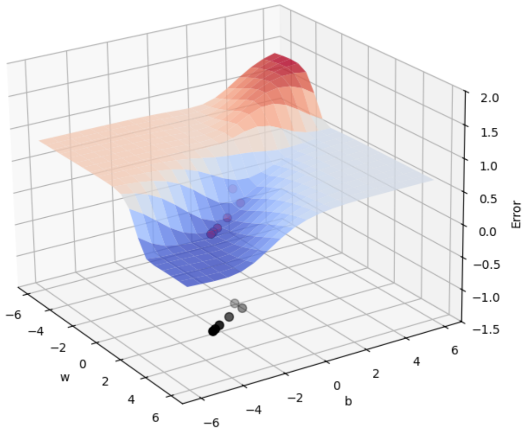
What we were actually doing !



Revisit to Approach 1



What we were actually doing !



So far ...

We were traversing the error surface by fiddling with w and b with mere intuition and guess work.

So far ...

We were traversing the error surface by fiddling with w and b with mere intuition and guess work.

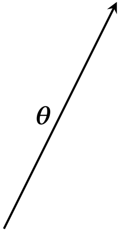
Up next ...

Looking for a more efficient and principled way of doing this.

Goal ahead

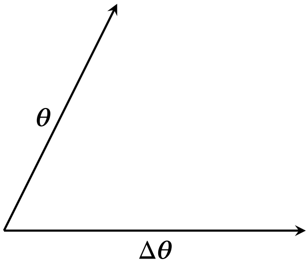
To find a better way of traversing the error surface so that we can reach the minimum value quickly without resorting to guess work or brute force search which is any how infeasible.

Approach 3 : A Principled Way



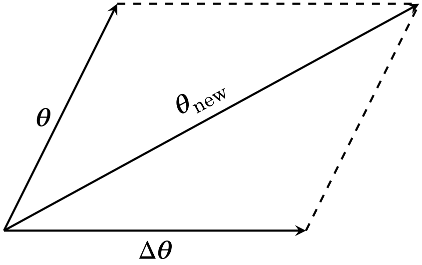
- Suppose we have a randomly initialized vector of parameters $\boldsymbol{\theta} = [w, b]$.

Approach 3 : A Principled Way



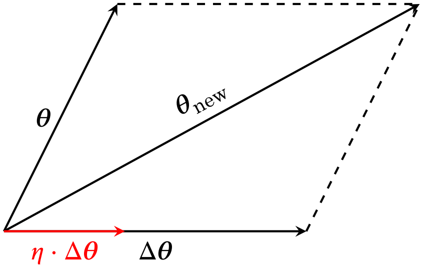
- Suppose we have a randomly initialized vector of parameters $\boldsymbol{\theta} = [w, b]$.
- Let $\Delta\boldsymbol{\theta} = [\Delta w, \Delta b]$ denote change in the values of w and b respectively.

Approach 3 : A Principled Way



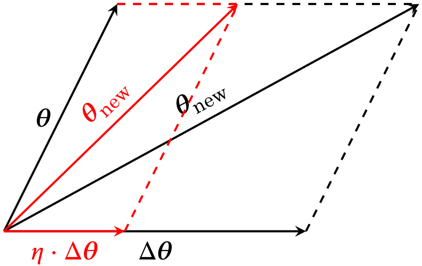
- Suppose we have a randomly initialized vector of parameters $\boldsymbol{\theta} = [w, b]$.
- Let $\Delta\boldsymbol{\theta} = [\Delta w, \Delta b]$ denote change in the values of w and b respectively.
- So we move in the direction of $\Delta\boldsymbol{\theta}$.

Approach 3 : A Principled Way



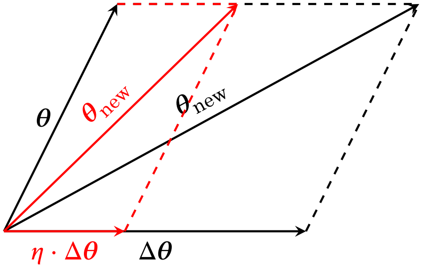
- Suppose we have a randomly initialized vector of parameters $\boldsymbol{\theta} = [w, b]$.
- Let $\Delta\boldsymbol{\theta} = [\Delta w, \Delta b]$ denote change in the values of w and b respectively.
- So we move in the direction of $\Delta\boldsymbol{\theta}$.
- Let us be a bit conservative: we move only by a small amount $\eta > 0$.

Approach 3 : A Principled Way



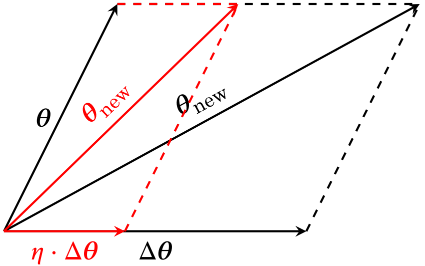
- Suppose we have a randomly initialized vector of parameters $\boldsymbol{\theta} = [w, b]$.
- Let $\Delta\boldsymbol{\theta} = [\Delta w, \Delta b]$ denote change in the values of w and b respectively.
- So we move in the direction of $\Delta\boldsymbol{\theta}$.
- Let us be a bit conservative: we move only by a small amount $\eta > 0$.
- $\boldsymbol{\theta}_{new} = \boldsymbol{\theta} + \eta \cdot \Delta\boldsymbol{\theta}$

Approach 3 : A Principled Way



- Suppose we have a randomly initialized vector of parameters $\theta = [w, b]$.
- Let $\Delta\theta = [\Delta w, \Delta b]$ denote change in the values of w and b respectively.
- So we move in the direction of $\Delta\theta$.
- Let us be a bit conservative: we move only by a small amount $\eta > 0$.
- $\theta_{new} = \theta + \eta \cdot \Delta\theta$
- Now the question is what is the right $\Delta\theta$ to use.

Approach 3 : A Principled Way



- Suppose we have a randomly initialized vector of parameters $\theta = [w, b]$.
- Let $\Delta\theta = [\Delta w, \Delta b]$ denote change in the values of w and b respectively.
- So we move in the direction of $\Delta\theta$.
- Let us be a bit conservative: we move only by a small amount $\eta > 0$.
- $\theta_{new} = \theta + \eta \cdot \Delta\theta$
- Now the question is what is the right $\Delta\theta$ to use.
- The answer comes from Taylor Series.

Taylor Series

For a function $\mathscr{F}(x)$ expanded around $a \in \mathbb{R}$, we have,

$$\mathscr{F}(x) = \mathscr{F}(a) + (x-a)\mathscr{F}'(a) + \frac{1}{2!}(x-a)^2 F''(a) + \cdots$$

Taylor Series

For a function $\mathscr{F}(x)$ expanded around $a \in \mathbb{R}$, we have,

$$\mathscr{F}(x) = \mathscr{F}(a) + (x-a)\mathscr{F}'(a) + \frac{1}{2!}(x-a)^2 F''(a) + \cdots$$

Multivariate Taylor Series

For a function $\mathscr{F}(\mathbf{x})$ expanded around $\mathbf{a} \in \mathbb{R}^n$, we have,

$$\mathscr{F}(\mathbf{x}) = \mathscr{F}(\mathbf{a}) + (\mathbf{x}-\mathbf{a})'\nabla\mathscr{F}(\mathbf{a}) + \frac{1}{2!}(\mathbf{x}-\mathbf{a})'\nabla^2\mathscr{F}(\mathbf{a})(\mathbf{x}-\mathbf{a}) + \cdots$$

$$\mathscr{F}(\mathbf{x}) = \mathscr{F}(\mathbf{a}) + (\mathbf{x}-\mathbf{a})'\nabla\mathscr{F}(\mathbf{a}) + \frac{1}{2!}(\mathbf{x}-\mathbf{a})'\nabla^2\mathscr{F}(\mathbf{a})(\mathbf{x}-\mathbf{a}) + \cdots$$

For ease of notation, take $\Delta\boldsymbol{\theta} = \boldsymbol{\mu}$.

$$\mathscr{F}(\mathbf{x}) = \mathscr{F}(\mathbf{a}) + (\mathbf{x}-\mathbf{a})'\nabla\mathscr{F}(\mathbf{a}) + \frac{1}{2!}(\mathbf{x}-\mathbf{a})'\nabla^2\mathscr{F}(\mathbf{a})(\mathbf{x}-\mathbf{a}) + \cdots$$

For ease of notation, take $\Delta\boldsymbol{\theta} = \boldsymbol{\mu}$.

With $\mathscr{F} = \mathscr{L}$ (our Loss Function),

$\mathbf{x} = \boldsymbol{\theta} + \eta\boldsymbol{\mu}$,

$\mathbf{a} = \boldsymbol{\theta}$; from the Taylor Series we have,

$$\mathscr{F}(\mathbf{x}) = \mathscr{F}(\mathbf{a}) + (\mathbf{x}-\mathbf{a})'\nabla\mathscr{F}(\mathbf{a}) + \frac{1}{2!}(\mathbf{x}-\mathbf{a})'\nabla^2\mathscr{F}(\mathbf{a})(\mathbf{x}-\mathbf{a}) + \cdots$$

For ease of notation, take $\Delta\boldsymbol{\theta} = \boldsymbol{\mu}$.

With $\mathscr{F} = \mathscr{L}$ (our Loss Function),

$\mathbf{x} = \boldsymbol{\theta} + \eta\boldsymbol{\mu}$,

$\mathbf{a} = \boldsymbol{\theta}$; from the Taylor Series we have,

$$\mathscr{L}$$

$$\mathscr{F}(\mathbf{x}) = \mathscr{F}(\mathbf{a}) + (\mathbf{x}-\mathbf{a})'\nabla\mathscr{F}(\mathbf{a}) + \frac{1}{2!}(\mathbf{x}-\mathbf{a})'\nabla^2\mathscr{F}(\mathbf{a})(\mathbf{x}-\mathbf{a}) + \cdots$$

For ease of notation, take $\Delta\boldsymbol{\theta} = \boldsymbol{\mu}$.

With $\mathscr{F} = \mathscr{L}$ (our Loss Function),

$\mathbf{x} = \boldsymbol{\theta} + \eta\boldsymbol{\mu}$,

$\mathbf{a} = \boldsymbol{\theta}$; from the Taylor Series we have,

$$\mathscr{L}(\boldsymbol{\theta} + \eta\boldsymbol{\mu}) =$$

$$\mathscr{F}(\mathbf{x}) = \mathscr{F}(\mathbf{a}) + (\mathbf{x} - \mathbf{a})'\nabla\mathscr{F}(\mathbf{a}) + \frac{1}{2!}(\mathbf{x} - \mathbf{a})'\nabla^2\mathscr{F}(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \cdots$$

For ease of notation, take $\Delta\boldsymbol{\theta} = \boldsymbol{\mu}$.

With $\mathscr{F} = \mathscr{L}$ (our Loss Function),
$\mathbf{x} = \boldsymbol{\theta} + \eta\boldsymbol{\mu}$,
$\mathbf{a} = \boldsymbol{\theta}$; from the Taylor Series we have,

$$\mathscr{L}(\boldsymbol{\theta} + \eta\boldsymbol{\mu}) = \mathscr{L}(\boldsymbol{\theta}) +$$

$$\mathscr{F}(\mathbf{x}) = \mathscr{F}(\mathbf{a}) + (\mathbf{x}-\mathbf{a})'\nabla\mathscr{F}(\mathbf{a}) + \frac{1}{2!}(\mathbf{x}-\mathbf{a})'\nabla^2\mathscr{F}(\mathbf{a})(\mathbf{x}-\mathbf{a}) + \cdots$$

For ease of notation, take $\Delta\boldsymbol{\theta} = \boldsymbol{\mu}$.

With $\mathscr{F} = \mathscr{L}$ (our Loss Function),

$\mathbf{x} = \boldsymbol{\theta} + \eta\boldsymbol{\mu}$,

$\mathbf{a} = \boldsymbol{\theta}$; from the Taylor Series we have,

$$\mathscr{L}(\boldsymbol{\theta} + \eta\boldsymbol{\mu}) = \mathscr{L}(\boldsymbol{\theta}) + (\eta\boldsymbol{\mu})'$$

$$\mathscr{F}(\mathbf{x}) = \mathscr{F}(\mathbf{a}) + (\mathbf{x}-\mathbf{a})'\nabla\mathscr{F}(\mathbf{a}) + \frac{1}{2!}(\mathbf{x}-\mathbf{a})'\nabla^2\mathscr{F}(\mathbf{a})(\mathbf{x}-\mathbf{a}) + \cdots$$

For ease of notation, take $\Delta\boldsymbol{\theta} = \boldsymbol{\mu}$.

With $\mathscr{F} = \mathscr{L}$ (our Loss Function),

$\mathbf{x} = \boldsymbol{\theta} + \eta\boldsymbol{\mu}$,

$\mathbf{a} = \boldsymbol{\theta}$; from the Taylor Series we have,

$$\mathscr{L}(\boldsymbol{\theta} + \eta\boldsymbol{\mu}) = \mathscr{L}(\boldsymbol{\theta}) + (\eta\boldsymbol{\mu})'\nabla\mathscr{L}(\boldsymbol{\theta}) +$$

$$\mathscr{F}(\mathbf{x}) = \mathscr{F}(\mathbf{a}) + (\mathbf{x}-\mathbf{a})'\nabla\mathscr{F}(\mathbf{a}) + \frac{1}{2!}(\mathbf{x}-\mathbf{a})'\nabla^2\mathscr{F}(\mathbf{a})(\mathbf{x}-\mathbf{a}) + \cdots$$

For ease of notation, take $\Delta\boldsymbol{\theta} = \boldsymbol{\mu}$.

With $\mathscr{F} = \mathscr{L}$ (our Loss Function),

$\mathbf{x} = \boldsymbol{\theta} + \eta\boldsymbol{\mu}$,

$\mathbf{a} = \boldsymbol{\theta}$; from the Taylor Series we have,

$$\mathscr{L}(\boldsymbol{\theta}+\eta\boldsymbol{\mu}) = \mathscr{L}(\boldsymbol{\theta}) + (\eta\boldsymbol{\mu})'\nabla\mathscr{L}(\boldsymbol{\theta}) + \frac{1}{2!}(\eta\boldsymbol{\mu})'\nabla^2\mathscr{L}(\boldsymbol{\theta})(\eta\boldsymbol{\mu}) + \cdots$$

$$\mathscr{L}(\boldsymbol{\theta}+\eta\boldsymbol{\mu}) = \mathscr{L}(\boldsymbol{\theta}) + (\eta\boldsymbol{\mu})'\nabla\mathscr{L}(\boldsymbol{\theta}) + \frac{1}{2!}(\eta\boldsymbol{\mu})'\nabla^2\mathscr{L}(\boldsymbol{\theta})(\eta\boldsymbol{\mu}) + \cdots$$

$$
\begin{aligned}
\mathscr{L}(\boldsymbol{\theta} + \eta\boldsymbol{\mu}) &= \mathscr{L}(\boldsymbol{\theta}) + (\eta\boldsymbol{\mu})'\nabla\mathscr{L}(\boldsymbol{\theta}) + \frac{1}{2!}(\eta\boldsymbol{\mu})'\nabla^2\mathscr{L}(\boldsymbol{\theta})(\eta\boldsymbol{\mu}) + \cdots \\
&= \mathscr{L}(\boldsymbol{\theta}) + \eta\,\boldsymbol{\mu}'\nabla\mathscr{L}(\boldsymbol{\theta}) + \frac{\eta^2}{2!}\,\boldsymbol{\mu}'\nabla^2\mathscr{L}(\boldsymbol{\theta})\boldsymbol{\mu} + \cdots
\end{aligned}
$$

$$
\begin{aligned}
\mathscr{L}(\boldsymbol{\theta} + \eta \boldsymbol{\mu}) &= \mathscr{L}(\boldsymbol{\theta}) + (\eta \boldsymbol{\mu})' \nabla \mathscr{L}(\boldsymbol{\theta}) + \frac{1}{2!}(\eta \boldsymbol{\mu})' \nabla^2 \mathscr{L}(\boldsymbol{\theta})(\eta \boldsymbol{\mu}) + \cdots \\
&= \mathscr{L}(\boldsymbol{\theta}) + \eta\, \boldsymbol{\mu}' \nabla \mathscr{L}(\boldsymbol{\theta}) + \frac{\eta^2}{2!}\, \boldsymbol{\mu}' \nabla^2 \mathscr{L}(\boldsymbol{\theta}) \boldsymbol{\mu} + \cdots \\
&= \mathscr{L}(\boldsymbol{\theta}) + \eta\, \boldsymbol{\mu}' \nabla \mathscr{L}(\boldsymbol{\theta}) \left[\eta \text{ typically being small } \eta^2, \eta^3, \ldots \to 0\right]
\end{aligned}
$$

$$\mathscr{L}(\boldsymbol{\theta} + \eta \boldsymbol{\mu}) = \mathscr{L}(\boldsymbol{\theta}) + \eta\, \boldsymbol{\mu}' \nabla \mathscr{L}(\boldsymbol{\theta})$$

$$\mathscr{L}(\boldsymbol{\theta} + \eta \boldsymbol{\mu}) = \mathscr{L}(\boldsymbol{\theta}) + \eta\, \boldsymbol{\mu}' \nabla \mathscr{L}(\boldsymbol{\theta})$$

Recall that our move was $(\eta \boldsymbol{\mu})$.

$$\mathscr{L}(\boldsymbol{\theta} + \eta\boldsymbol{\mu}) = \mathscr{L}(\boldsymbol{\theta}) + \eta\,\boldsymbol{\mu}'\nabla\mathscr{L}(\boldsymbol{\theta})$$

Recall that our move was $(\eta\boldsymbol{\mu})$.

The move $(\eta\boldsymbol{\mu})$ would be favourable only if,

$$\mathscr{L}(\boldsymbol{\theta} + \eta \boldsymbol{\mu}) = \mathscr{L}(\boldsymbol{\theta}) + \eta\, \boldsymbol{\mu}' \nabla \mathscr{L}(\boldsymbol{\theta})$$

Recall that our move was $(\eta \boldsymbol{\mu})$.

The move $(\eta \boldsymbol{\mu})$ would be favourable only if,

$$\mathscr{L}(\boldsymbol{\theta} + \eta \boldsymbol{\mu}) - \mathscr{L}(\boldsymbol{\theta}) < 0$$

$$\mathscr{L}(\boldsymbol{\theta}+\eta\boldsymbol{\mu}) = \mathscr{L}(\boldsymbol{\theta}) + \eta\,\boldsymbol{\mu}'\,\nabla\mathscr{L}(\boldsymbol{\theta})$$

Recall that our move was $(\eta\boldsymbol{\mu})$.

The move $(\eta\boldsymbol{\mu})$ would be favourable only if,

$$\mathscr{L}(\boldsymbol{\theta}+\eta\boldsymbol{\mu}) - \mathscr{L}(\boldsymbol{\theta}) < 0$$

i.e. if the new loss is less than the previous loss.

$$\mathscr{L}(\boldsymbol{\theta} + \eta\boldsymbol{\mu}) = \mathscr{L}(\boldsymbol{\theta}) + \eta\,\boldsymbol{\mu}'\nabla\mathscr{L}(\boldsymbol{\theta})$$

Recall that our move was $(\eta\boldsymbol{\mu})$.

The move $(\eta\boldsymbol{\mu})$ would be favourable only if,

$$\mathscr{L}(\boldsymbol{\theta} + \eta\boldsymbol{\mu}) - \mathscr{L}(\boldsymbol{\theta}) < 0$$

i.e. if the new loss is less than the previous loss.

This implies

$$\boldsymbol{\mu}'\nabla\mathscr{L}(\boldsymbol{\theta}) < 0$$

- More the negative $\boldsymbol{\mu}'\nabla\mathscr{L}(\boldsymbol{\theta})$ is,

- More the negative $\boldsymbol{\mu}' \nabla \mathscr{L}(\boldsymbol{\theta})$ is,

  the more favourable is the move $\eta \boldsymbol{\mu}$.

Got that?

- Now let us find out the range of $\boldsymbol{\mu}' \nabla \mathscr{L}(\boldsymbol{\theta})$.

- Let β be the angle between $\boldsymbol{\mu}$ and $\nabla_{\boldsymbol{\theta}}\mathscr{L}(\boldsymbol{\theta})$, then we know that,

$$-1 \leq \cos(\beta) = \frac{\boldsymbol{\mu}^T \nabla_{\boldsymbol{\theta}}\mathscr{L}(\boldsymbol{\theta})}{\|\boldsymbol{\mu}\| \cdot \|\nabla_{\boldsymbol{\theta}}\mathscr{L}(\boldsymbol{\theta})\|} \leq 1$$

- Let β be the angle between $\boldsymbol{\mu}$ and $\nabla_{\boldsymbol{\theta}}\mathscr{L}(\boldsymbol{\theta})$, then we know that,

$$-1 \leq \cos(\beta) = \frac{\boldsymbol{\mu}^T \nabla_{\boldsymbol{\theta}}\mathscr{L}(\boldsymbol{\theta})}{\|\boldsymbol{\mu}\| \cdot \|\nabla_{\boldsymbol{\theta}}\mathscr{L}(\boldsymbol{\theta})\|} \leq 1$$

- Multiplying throughout by $k = \|\boldsymbol{\mu}\| \cdot \|\nabla_{\boldsymbol{\theta}}\mathscr{L}(\boldsymbol{\theta})\|$:

$$-k \leq k \cdot \cos(\beta) = \boldsymbol{\mu}^T \nabla_{\boldsymbol{\theta}}\mathscr{L}(\boldsymbol{\theta}) \leq k$$

- Let β be the angle between $\boldsymbol{\mu}$ and $\nabla_{\boldsymbol{\theta}}\mathscr{L}(\boldsymbol{\theta})$, then we know that,

$$-1 \leq \cos(\beta) = \frac{\boldsymbol{\mu}^T \nabla_{\boldsymbol{\theta}}\mathscr{L}(\boldsymbol{\theta})}{\|\boldsymbol{\mu}\| \cdot \|\nabla_{\boldsymbol{\theta}}\mathscr{L}(\boldsymbol{\theta})\|} \leq 1$$

- Multiplying throughout by $k = \|\boldsymbol{\mu}\| \cdot \|\nabla_{\boldsymbol{\theta}}\mathscr{L}(\boldsymbol{\theta})\|$:

$$-k \leq k \cdot \cos(\beta) = \boldsymbol{\mu}^T \nabla_{\boldsymbol{\theta}}\mathscr{L}(\boldsymbol{\theta}) \leq k$$

- Thus,

$$\mathscr{L}(\boldsymbol{\theta} + \eta\boldsymbol{\mu}) - \mathscr{L}(\boldsymbol{\theta}) = \boldsymbol{\mu}^T \nabla_{\boldsymbol{\theta}}\mathscr{L}(\boldsymbol{\theta}) = k \cdot \cos(\beta)$$

will be most negative when $\cos(\beta) = -1$, *i.e.*, when β is 180°.

Best Update Rule

So our best move is at 180° w.r.t. the gradient $\nabla\mathscr{L}(\boldsymbol{\theta})$.

Best Update Rule

So our best move is at 180° w.r.t. the gradient $\nabla \mathscr{L}(\boldsymbol{\theta})$.

In other words, we should move in a direction opposite to the gradient *i.e.*

$$\eta \Delta \boldsymbol{\theta} = -\eta \nabla \mathscr{L}(\boldsymbol{\theta})$$

Best Update Rule

So our best move is at 180° w.r.t. the gradient $\nabla\mathscr{L}(\boldsymbol{\theta})$.

In other words, we should move in a direction opposite to the gradient *i.e.*

$$\eta\Delta\boldsymbol{\theta} = -\eta\nabla\mathscr{L}(\boldsymbol{\theta})$$

Parameter Update Equations

$$w_{t+1} = w_t - \eta\nabla w_t$$
$$b_{t+1} = b_t - \eta\nabla b_t$$

$$\text{where, } \nabla w_t = \left.\frac{\partial\mathscr{L}(w,b)}{\partial w}\right|_{w=w_t,\, b=b_t}, \quad \nabla b_t = \left.\frac{\partial\mathscr{L}(w,b)}{\partial b}\right|_{w=w_t,\, b=b_t}$$

Naming the Algorithm

Naming the Algorithm

We are descending along the error surface

Naming the Algorithm

We are descending along the error surface driven by gradients.

Naming the Algorithm

We are descending along the error surface

driven by gradients.

So we call it **Gradient Descent**.

A Formal Definition : Gradient Descent

Definition

A Formal Definition : Gradient Descent

Definition

Gradient Descent is a

A Formal Definition : Gradient Descent

Definition

Gradient Descent is a **first-order**

A Formal Definition : Gradient Descent

Definition

Gradient Descent is a **first-order iterative** algorithm for

A Formal Definition : Gradient Descent

Definition

Gradient Descent is a **first-order iterative** algorithm for **minimizing**

A Formal Definition : Gradient Descent

Definition

Gradient Descent is a **first-order iterative** algorithm for **minimizing** a **differentiable** multivariate function.

A Formal Definition : Gradient Descent

Definition

Gradient Descent is a **first-order iterative** algorithm for **minimizing** a **differentiable** multivariate function.

The idea is to take repeated steps in the opposite direction of the gradient (or approximate gradient) of the function at the current point, because this is the direction of steepest descent.

Batch / Vanilla 🍦 Gradient Descent

Let's formulating the algorithm.

Algorithm 1: `gradient_descent()`

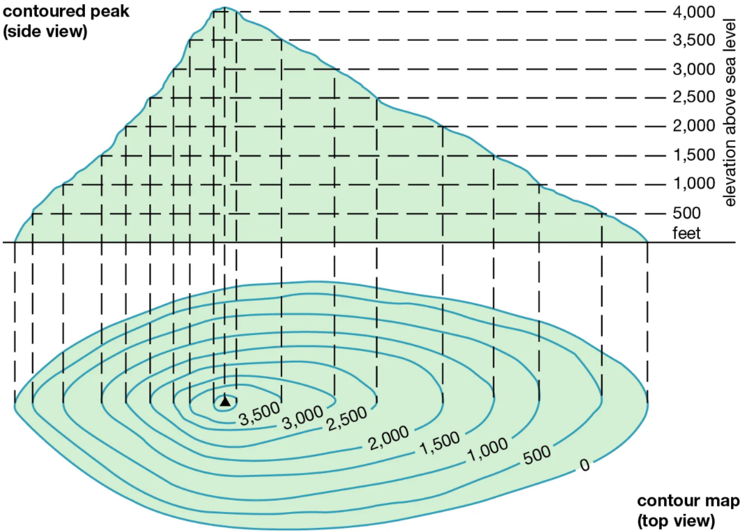
```
t ← 0;
max_iterations ← 1000;
while t < max_iterations do
    w_{t+1} ← w_t − η ∇w_t;
    b_{t+1} ← b_t − η ∇b_t;
    t ← t + 1;
end
```

Visualizations

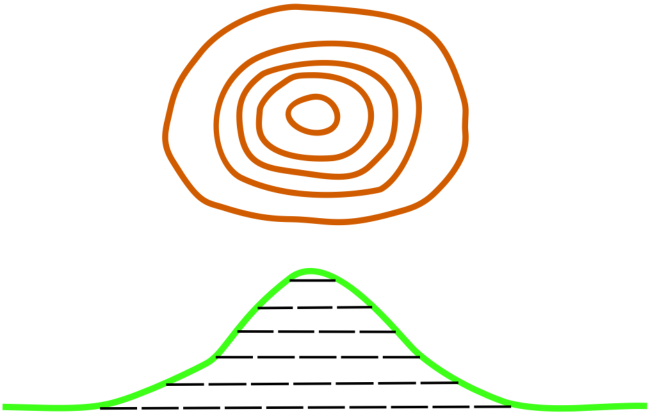
Visualizations

A Digression to Contours

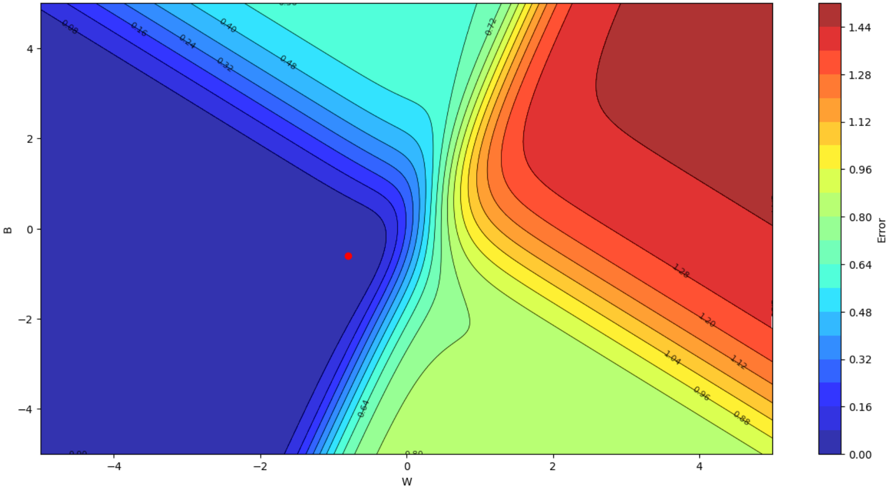
A Digression to Contours



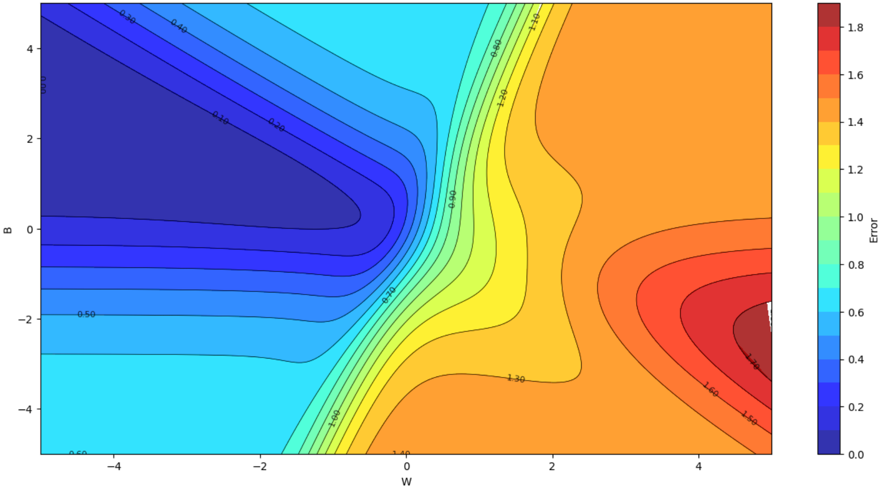
A Digression to Contours



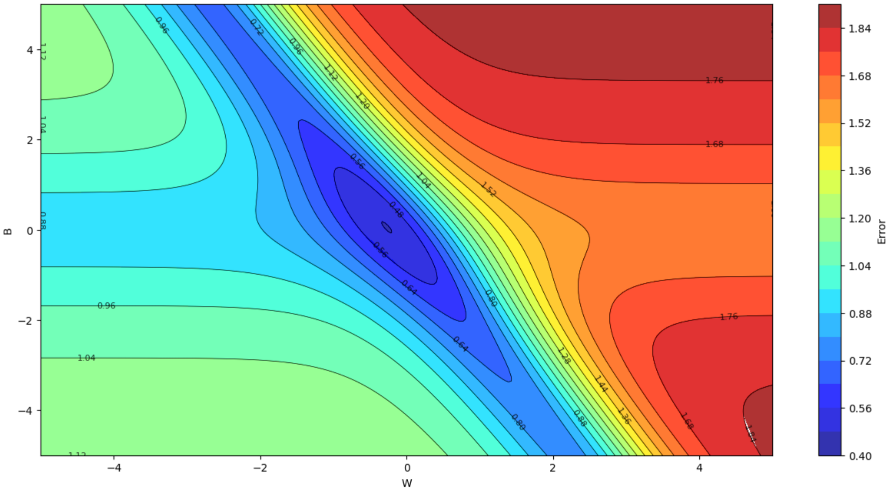
A Digression to Contours



A Digression to Contours



A Digression to Contours



Some Observations in Vanilla Gradient Descent

Some Observations in Vanilla Gradient Descent

- It takes a lot of time to navigate regions having a gentle slope.

Some Observations in Vanilla Gradient Descent

- It takes a lot of time to navigate regions having a gentle slope.
- This is because the gradient in these regions is very small.

Some Observations in Vanilla Gradient Descent

- It takes a lot of time to navigate regions having a gentle slope.
- This is because the gradient in these regions is very small.
- We have to do something better.

Questions ?

Intuition

- If I am repeatedly being asked to move in the same direction then I should probably gain some confidence and start taking bigger steps in that direction.

Intuition

- If I am repeatedly being asked to move in the same direction then I should probably gain some confidence and start taking bigger steps in that direction.
- Just as a ball gains momentum while rolling down a slope

Momentum Based Gradient Descent

A New Update Rule

$$\text{update}_t = \gamma \cdot \text{update}_{t-1} + \eta \cdot \nabla w_t$$

$$w_{t+1} = w_t - \text{update}_t$$

Momentum Based Gradient Descent

A New Update Rule

$$\text{update}_t = \gamma \cdot \text{update}_{t-1} + \eta \cdot \nabla w_t$$
$$w_{t+1} = w_t - \text{update}_t$$

- We have a similar update rule for b.

Momentum Based Gradient Descent

A New Update Rule

$$\text{update}_t = \gamma \cdot \text{update}_{t-1} + \eta \cdot \nabla w_t$$
$$w_{t+1} = w_t - \text{update}_t$$

- We have a similar update rule for b.
- In addition to the current update, also look at the history of updates.

$$
\begin{aligned}
\text{update}_t &= \gamma \cdot \text{update}_{t-1} + \eta \nabla w_t \\
w_{t+1} &= w_t - \text{update}_t
\end{aligned}
$$

$$\text{update}_t = \gamma \cdot \text{update}_{t-1} + \eta \nabla w_t$$
$$w_{t+1} = w_t - \text{update}_t$$

$\text{update}_0 = 0$

$$\text{update}_t = \gamma \cdot \text{update}_{t-1} + \eta \nabla w_t$$
$$w_{t+1} = w_t - \text{update}_t$$

$$\text{update}_0 = 0$$
$$\text{update}_1 = \gamma \cdot \text{update}_0 + \eta \nabla w_1 = \eta \nabla w_1$$

$$\mathrm{update}_t = \gamma \cdot \mathrm{update}_{t-1} + \eta \nabla w_t$$
$$w_{t+1} = w_t - \mathrm{update}_t$$

$$\mathrm{update}_0 = 0$$
$$\mathrm{update}_1 = \gamma \cdot \mathrm{update}_0 + \eta \nabla w_1 = \eta \nabla w_1$$
$$\mathrm{update}_2 = \gamma \cdot \mathrm{update}_1 + \eta \nabla w_2 = \gamma \cdot \eta \nabla w_1 + \eta \nabla w_2$$

$$\text{update}_t = \gamma \cdot \text{update}_{t-1} + \eta \nabla w_t$$
$$w_{t+1} = w_t - \text{update}_t$$

$$\begin{aligned}
\text{update}_0 &= 0 \\
\text{update}_1 &= \gamma \cdot \text{update}_0 + \eta \nabla w_1 = \eta \nabla w_1 \\
\text{update}_2 &= \gamma \cdot \text{update}_1 + \eta \nabla w_2 = \gamma \cdot \eta \nabla w_1 + \eta \nabla w_2 \\
\text{update}_3 &= \gamma \cdot \text{update}_2 + \eta \nabla w_3 = \gamma(\gamma \cdot \eta \nabla w_1 + \eta \nabla w_2) + \eta \nabla w_3 \\
&= \gamma^2 \cdot \eta \nabla w_1 + \gamma \cdot \eta \nabla w_2 + \eta \nabla w_3
\end{aligned}$$

$$\mathrm{update}_t = \gamma \cdot \mathrm{update}_{t-1} + \eta \nabla w_t$$
$$w_{t+1} = w_t - \mathrm{update}_t$$

$$\begin{aligned}
\mathrm{update}_0 &= 0 \\
\mathrm{update}_1 &= \gamma \cdot \mathrm{update}_0 + \eta \nabla w_1 = \eta \nabla w_1 \\
\mathrm{update}_2 &= \gamma \cdot \mathrm{update}_1 + \eta \nabla w_2 = \gamma \cdot \eta \nabla w_1 + \eta \nabla w_2 \\
\mathrm{update}_3 &= \gamma \cdot \mathrm{update}_2 + \eta \nabla w_3 = \gamma(\gamma \cdot \eta \nabla w_1 + \eta \nabla w_2) + \eta \nabla w_3 \\
&= \gamma^2 \cdot \eta \nabla w_1 + \gamma \cdot \eta \nabla w_2 + \eta \nabla w_3 \\
\mathrm{update}_4 &= \gamma \cdot \mathrm{update}_3 + \eta \nabla w_4 = \gamma^3 \cdot \eta \nabla w_1 + \gamma^2 \cdot \eta \nabla w_2 + \gamma \cdot \eta \nabla w_3 + \eta \nabla w_4
\end{aligned}$$

$$\text{update}_t = \gamma \cdot \text{update}_{t-1} + \eta \nabla w_t$$
$$w_{t+1} = w_t - \text{update}_t$$

$$\begin{aligned}
\text{update}_0 &= 0 \\
\text{update}_1 &= \gamma \cdot \text{update}_0 + \eta \nabla w_1 = \eta \nabla w_1 \\
\text{update}_2 &= \gamma \cdot \text{update}_1 + \eta \nabla w_2 = \gamma \cdot \eta \nabla w_1 + \eta \nabla w_2 \\
\text{update}_3 &= \gamma \cdot \text{update}_2 + \eta \nabla w_3 = \gamma(\gamma \cdot \eta \nabla w_1 + \eta \nabla w_2) + \eta \nabla w_3 \\
&= \gamma^2 \cdot \eta \nabla w_1 + \gamma \cdot \eta \nabla w_2 + \eta \nabla w_3 \\
\text{update}_4 &= \gamma \cdot \text{update}_3 + \eta \nabla w_4 = \gamma^3 \cdot \eta \nabla w_1 + \gamma^2 \cdot \eta \nabla w_2 + \gamma \cdot \eta \nabla w_3 + \eta \nabla w_4 \\
&\vdots \\
\text{update}_t &= \gamma \cdot \text{update}_{t-1} + \eta \nabla w_t \\
&= \gamma^{t-1} \cdot \eta \nabla w_1 + \gamma^{t-2} \cdot \eta \nabla w_2 + \cdots + \eta \nabla w_t
\end{aligned}$$

Some observations and questions

- Even in the regions having gentle slopes, momentum based gradient descent is able to take large steps because the momentum carries it along.

Some observations and questions

- Even in the regions having gentle slopes, momentum based gradient descent is able to take large steps because the momentum carries it along.
- Is moving fast always good? Would there be situations where momentum would cause us to run pass our goal?

- Momentum based gradient descent oscillates in and out of the minima valley as the momentum carries it out of the valley.

- Momentum based gradient descent oscillates in and out of the minima valley as the momentum carries it out of the valley.
- Takes a lot of u-turns before finally converging.

- Momentum based gradient descent oscillates in and out of the minima valley as the momentum carries it out of the valley.
- Takes a lot of u-turns before finally converging.
- Despite these u-turns it still converges faster than Vanilla Gradient Descent.

Question

- Can we do something to reduce these oscillations ?

☞ Of Course !

Recall in Momentum Gradient Descent

- We had $w_{t+1} = w_t - \text{update}_t$ where

Recall in Momentum Gradient Descent

- We had $w_{t+1} = w_t - \text{update}_t$ where

$$\text{update}_t = \underbrace{\gamma \cdot \text{update}_{t-1}}_{} + \underbrace{\eta \nabla w_t}_{}$$

Recall in Momentum Gradient Descent

- We had $w_{t+1} = w_t - \text{update}_t$ where

$$\text{update}_t = \underbrace{\gamma \cdot \text{update}_{t-1}}_{\substack{\text{momentum derived} \\ \text{from past updates}}} + \underbrace{\eta \nabla w_t}$$

Recall in Momentum Gradient Descent

- We had $w_{t+1} = w_t - \text{update}_t$ where

$$\text{update}_t = \underbrace{\gamma \cdot \text{update}_{t-1}}_{\substack{\text{momentum derived} \\ \text{from past updates}}} + \underbrace{\eta \nabla w_t}_{\substack{\text{push from the} \\ \text{current gradient}}}$$

Recall in Momentum Gradient Descent

- We had $w_{t+1} = w_t - \text{update}_t$ where

$$\text{update}_t = \underbrace{\gamma \cdot \text{update}_{t-1}}_{\substack{\text{momentum derived}\\ \text{from past updates}}} + \underbrace{\eta \nabla w_t}_{\substack{\text{push from the}\\ \text{current gradient}}}$$

- So we know that we are going to move by at least by update_{t-1} and then a bit more by $\eta \nabla w_t$.

Intuition

- Look before you leap.

Intuition

- Look before you leap.
- If we already know the direction we're moving in, why not "look ahead" to where the momentum will take us before calculating the gradient?

Intuition

- Look before you leap.
- If we already know the direction we're moving in, why not "look ahead" to where the momentum will take us before calculating the gradient?
- $w_{\text{look\_ahead}} = w_t - \gamma \cdot \text{update}_{t-1}$

Intuition

- Look before you leap.
- If we already know the direction we're moving in, why not "look ahead" to where the momentum will take us before calculating the gradient?
- $w_{\text{look\_ahead}} = w_t - \gamma \cdot \text{update}_{t-1}$
- Why not calculate the gradient at this partially updated value of w *i.e.* $w_{\text{look\_ahead}}$ instead of calculating it using the current value w_t ?

Nesterov Accelerated Gradient Descent

A New Update Rule

$$
\begin{aligned}
w_{\text{look\_ahead}} &= w_t - \gamma \cdot \text{update}_{t-1} \\
\text{update}_t &= \gamma \cdot \text{update}_{t-1} + \eta \nabla w_{\text{look\_ahead}} \\
w_{t+1} &= w_t - \text{update}_t
\end{aligned}
$$

We have similar update rule for b.