Lec 9 Gradient Desent

↳. How to minimize $E_{in}(\underline{w}) = \frac{1}{N}\sum_{n=1}^{N} \log(1+e^{-y_n \underline{w}^T \underline{x}_n})$?

1. <u>Intro to gradient desent:</u>

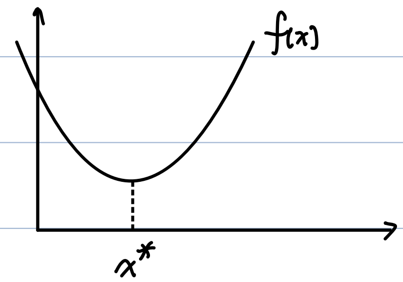
↳ Given differenable function $f: \mathbb{R}^n \to \mathbb{R}$,

we want to $\min_{\underline{x} \in \mathbb{R}^n} f(\underline{x})$ // here $\underline{x}$ is not data. Just a general term

(unconstrained)

↳ Gradient desent is a numerical approach.

↳ <u>example:</u>



- $x \in \mathbb{R}^n$. $n=1$ i.e. $x \in \mathbb{R}$

// Assume $f(\underline{x})$ is convex for now

- If $x = x^*$, then $f'(x) = 0$

  $x > x^*$, then $f'(x) > 0 \Rightarrow f(x)$ increases w.r.t x

  $x < x^*$, then $f'(x) < 0 \Rightarrow f(x)$ decreases w.r.t x

↳ <u>steps for $\mathbb{R} \in \mathbb{R}^1$:</u>

① initialize $x = x_0$

② If $f'(x) \approx 0$, then stop

③ If $f'(x) > 0$, then $x = x - \varepsilon$; If $f'(x) < 0$, then $x = x + \varepsilon$

④ Go to step 2

// ε : step size $\begin{cases} \text{large } \varepsilon: \text{ less accurate} \\ \text{small } \varepsilon: \text{ slow program} \end{cases}$

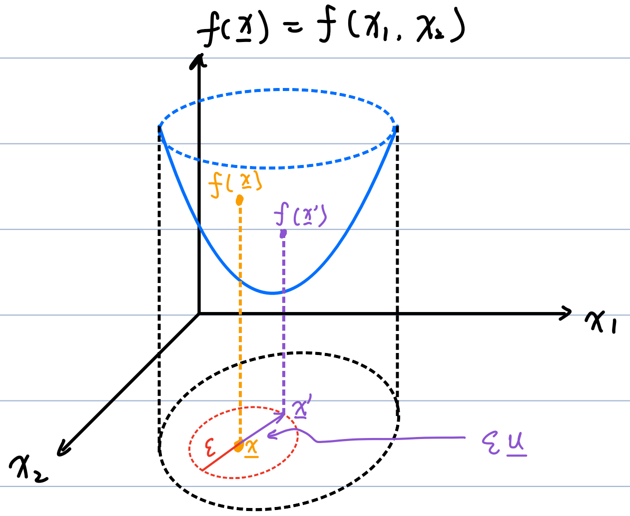
↳ <u>setup for step size ε:</u>

use ε_k, where k is the iteration number.

e.g. $\varepsilon_k \sim \frac{1}{k}$

2. General Gradient desent:

↳ e.g. $n = 2$:



Given current location $\underline{x}$, what's the next step to go?

$\Rightarrow \underline{x}' \leftarrow \underline{x} + \varepsilon \underline{u}$

unit vector, just for direction

↳ Idea: choose direction that maximize $f(\underline{x}) - f(\underline{x}')$

$\because f(\underline{x}') = f(\underline{x} + \varepsilon \underline{u})$

$= f(\underline{x}) + \varepsilon \underline{u}^T \nabla f(\underline{x}) + O(\varepsilon^2)$ // Taylor series expression

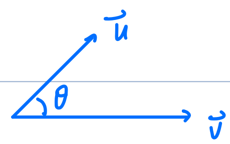
// $O(\varepsilon^2)$: smaller term of ε, negligible

$$\nabla f(\underline{x}) = \begin{bmatrix} \partial f / \partial x_1 \\ \vdots \\ \partial f / \partial x_n \end{bmatrix}$$

Pick $\underline{u}$ to minimize $\underline{u}^T \nabla f(\underline{x})$, called $\underline{u}^*$

// Recall: $\underline{u}^T \underline{v} = \|u\| \cdot \|v\| \cos\theta$

$\geq - \|u\| \cdot \|v\|$



$$\therefore \underline{u}^* = - \frac{\nabla f(\underline{x})}{\|\nabla f(\underline{x})\|}$$ ← Normalize so that $\|u\| = 1$

※ direction of $\underline{u}^*$ is the opposite of $\nabla f(\underline{x})$

↳ Note:

① $\nabla f(\underline{x})$ are points in the direction where $f(\underline{x})$ has max ascent.

② Magnitude of $\|f(\underline{x})\|$ in the denominator can be absorbed into ε.

3. Gradient Desent Algorithm:

- Initialize $\underline{x}_0$ (typically pick at random)
- For $t = 0, 1, 2 \cdots$

① Compute $\underline{g}_t = \nabla f(\underline{x}_t)$

② select direction $\underline{v}_t = -\underline{g}_t$

③ update $\underline{x}_{t+1} = \underline{x}_t + \varepsilon_t \underline{v}_t$ // Normalization part absorbed.

④ Go to step 1 until stopping criteria are reached

Note:
- "ε_t" : learning rate
- Condition for stopping : $\nabla f(\underline{x}_t) \approx 0$ is reasonable

4. GD of logistic regression

Recap: $\underline{x}_{t+1} = \underline{x}_t - \varepsilon_t \underbrace{\nabla f(\underline{x}_t)}$ → $\nabla_{\underline{x}} f(\underline{x}) \big|_{\underline{x} = \underline{x}_t}$

e.g. logistic regression,

how to minimize $E_{in}(\underline{w}) = \frac{1}{N} \sum_{n=1}^{N} \log(1 + e^{-y_n \underline{w}^T \underline{x}_n})$?

$$= \frac{1}{N} \sum_{n=1}^{N} e_n(\underline{w})$$

↳ GD: $\underline{w}_{k+1} = \underline{w}_k - \varepsilon_k \nabla E_{in}(\underline{w}_k)$

$$= \underline{w}_k - \varepsilon_k \frac{1}{N} \sum_{n=1}^{N} \nabla e_n(\underline{w}_k) \; ?$$

$$\nabla e_n(\underline{w}_k) = \nabla_{\underline{w}} \log\left(1 + e^{-y_n \underline{w}^T \underline{x}_n}\right)$$

$$= \frac{1}{1 + e^{-y_n \underline{w}^T \underline{x}_n}} \cdot \nabla_{\underline{w}}\left(1 + e^{-y_n \underline{w}^T \underline{x}_n}\right)$$

$$= \frac{e^{-y_n \underline{w}^T \underline{x}_n}}{1 + e^{-y_n \underline{w}^T \underline{x}_n}} \cdot \underbrace{\nabla_{\underline{w}}\left(-y_n \underline{w}^T \underline{x}_n\right)}_{= -y_n \cdot \underline{x}_n}$$

$$= \frac{-y_n \underline{x}_n}{1 + e^{y_n \underline{w}^T \underline{x}_n}}$$