Lec 12 GD / SGD non-convex func

1. For convex functions:

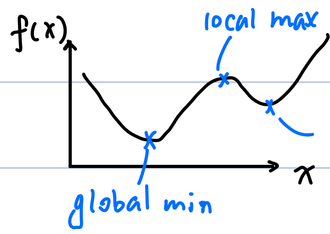
e.g. logistic regression. linear regression

the $E_{in}(\underline{w})$ is convex w.r.t the parameter w.

$\Rightarrow$ we will get a global minimum when $\nabla E_{in}(w) = 0$

2. Non-convex functions:

↳ e.g.



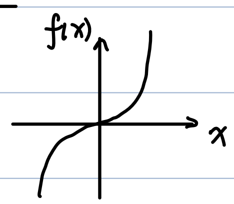
↳ those points are called saddle points:

$\nabla f(\underline{x}) = \underline{0}$, but $\underline{x}$ is neither a local min or local max

include local min & global min

↳ example in 1-D space:

$f(x) = x^3$



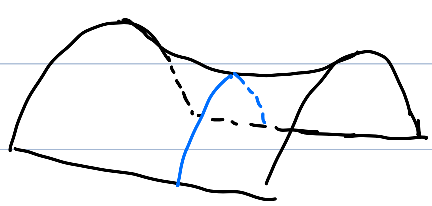
$\Rightarrow$ very slow progress at saddle point

$\Rightarrow$ Need to pre-set No. of iteration runs out

↳ In n-D case:

① 1-D saddle point in any direction

② local min in some direction, and local max in some other direction



saddle shape

③ both ① and ②

⇒ highly likely to have saddle points in high-dimensional space

↳ Stopping conditions:

- aim: to reduce the chance of stopping at a saddle point
- is saddle point if
  - ① $\nabla E_{in}(\underline{w}) = 0$
  - ② $\nabla E_{in}(\underline{w})$ is small
  - ③ No. of iteration is large

⇒ Solution: SGD with momentum

3. SGD with Momentum: (By Polyak, 1964)

↳ Recall basic SGD:

$$\underline{g}_k = \nabla e_n(\underline{w}_k)$$

$$\underline{w}_{k+1} = \underline{w}_k - \varepsilon_k \underline{g}_k$$

↳ SGD with momen

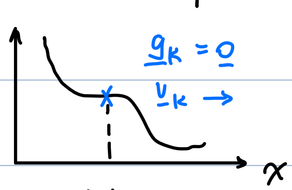
$$\boxed{\begin{aligned} \underline{g}_k &= \nabla e_n(\underline{w}_k) \\ \underline{v}_k &= -\varepsilon_k \underline{g}_k + \underbrace{\mu \underline{v}_{k-1}}_{\text{accumulation of prev momentums}} \\ \underline{w}_{k+1} &= \underline{w}_k + \underline{v}_k \end{aligned}}$$

// commonly $\mu = 0.9$

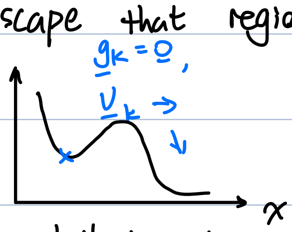
→ accumulation of prev momentums

↳ Usage:

① Momentum helps SGD escape that regions



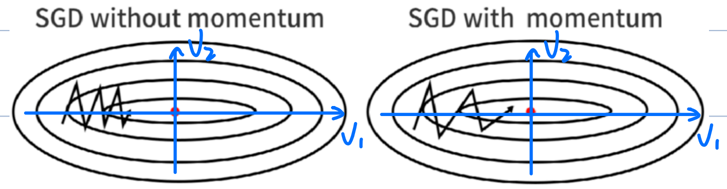
saddle point



shallow local min

"heavy ball momentum"

③ Momentum can lead to faster convergence, even for convex function



$f(x)$. $d=2$.

GD strongly favors the V_2 direction
⇒ Not ideal

Moving along v_1 direction

4. <u>Nesterov momentum</u> (1983)

↳ $\underline{v}_k = -\varepsilon_k \nabla e_n(\underline{w}_k + \mu \underline{v}_{k-1}) + \mu \underline{v}_{k-1}$

$\underline{w}_{k+1} = \underline{w}_k + \underline{v}_k$

↳ compare:

original version is: $\underline{v}_k = -\varepsilon_k \nabla e_n(\underbrace{\underline{w}_k}_{\text{here is different}}) + \mu \underline{v}_{k-1}$

↳ Can prove better convergence.

For convex func,

→ Full GD: distance between $\underline{w}_k$ and $\underline{w}^*$ is $O(\frac{1}{k})$

→ with Nesterov momentum: $O(\frac{1}{k^2})$