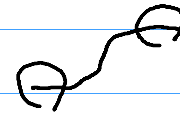
* Initialization for Deep Learning

activation ⇒ bad $W_{ij}^{(\ell)}$ properties

e.g. tanh : $|W_{ij}^{(\ell)}|$ too big → saturation 

$\boxed{W_{ij}^{(\ell)} = 0, \text{ or constant}}$ → saddle, symmetry

ReLU : $\boxed{W_{ij}^{(\ell)} \text{ too negative}}$ → saturation (signal shutdown)

* want : random small

usually 0-mean

uniform $[-L, +L]$

(truncated) Gaussian $(0, \sigma^2)$

ReLU $\max(s, 0)$

weight variance $\frac{2}{d^{(\ell-1)}}$

variance of $x_i^{(\ell-1)}$ ≈ variance of $x_j^{(\ell)}$

w/ assumptions

He init.

tanh

weight variance $\frac{1}{d^{(\ell-1)}}$

tanh(s) ≈ s

weight variance $\frac{1}{d^{(\ell)}}$

variance of $\delta^{(\ell-1)}$ ≈ variance of $\delta^{(\ell)}$

when tanh(s) ≈ s

weight variance $\frac{2}{d^{(\ell-1)} + d^{(\ell)}}$

Glorot init.

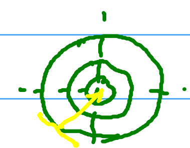
$[-U, U]$ uniform

$\sqrt{\frac{6}{\boxed{1 + d^{(\ell-1)}} + d^{(\ell)}}}$

* difficulty in DL optimization

surface
- local min: not as bad as imagined
- saddle / local max: easily escapable (esp. w/ SGD)
- plateau: need larger η (learning rate)
- ravine: avoid oscillation

hessian $\underline{H}^{-1}$ #w #w



(Quasi-)Newton, not feasible for DL

- slow computation of gradient: SGD on minibatch

$\Downarrow$

"instable" estimate of gradient

* running average estimate of SG

(SGD) with momentum

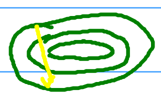
$$\underset{\text{current average}}{\underline{V}_t} = \beta \cdot \underset{\text{previous average}}{\underline{V}_{t-1}} + (1-\beta)\underline{\Delta}_t$$

$\underline{\Delta}_t$: SG at t-th iteration

update: $\underline{W}_t = \underline{W}_{t-1} - \eta \underline{V}_t$

- stochastic instability

pass ravine faster
plateau
shall local min



* ravine

larger gradient component
want: smaller step

per-component learning rate

update: $\underline{W}_t = \underline{W}_{t-1} - \eta \cdot \underline{\Delta}_t \boxed{\cdot/}$ "magnitude($\underline{\Delta}_t$)"

$\parallel$

$\sqrt{\underline{U}_t + \epsilon}$

running average

$$\underline{U}_t = \beta \cdot \underline{U}_{t-1} + (1-\beta)\underbrace{\underline{\Delta}_t \boxed{\cdot *} \underline{\Delta}_t}_{\text{RMS prop}}$$

* Adam = SGD + momentum + RMSProp + (other tricks)

0.9

$$\underline{v}_t = \beta_1 \underline{v}_{t-1} + (1-\beta_1)\,\underline{\Delta}_t$$

0.001

$$\underline{u}_t = \beta_2 \underline{u}_{t-1} + (1-\beta_2)\left(\underline{\Delta}_t \;.\!*\; \underline{\Delta}_t\right)$$

$\dfrac{\eta}{\sqrt{t/N}}$ decaying

0.999

$$\underline{W}_t = \underline{W}_{t-1} - \frac{\eta_t}{\sqrt{\underline{u}_t + \epsilon}} \;.\!*\; \underline{v}_t$$

10^{-8}