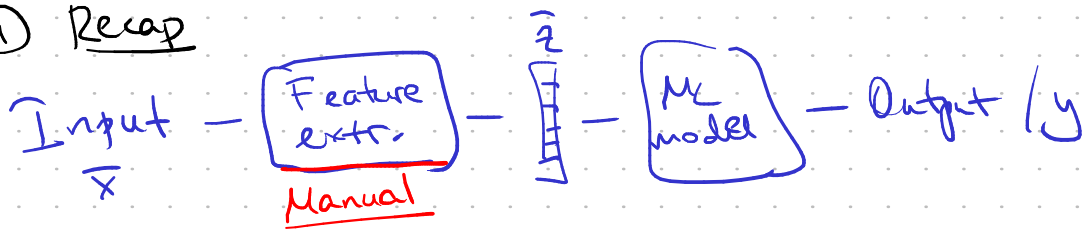
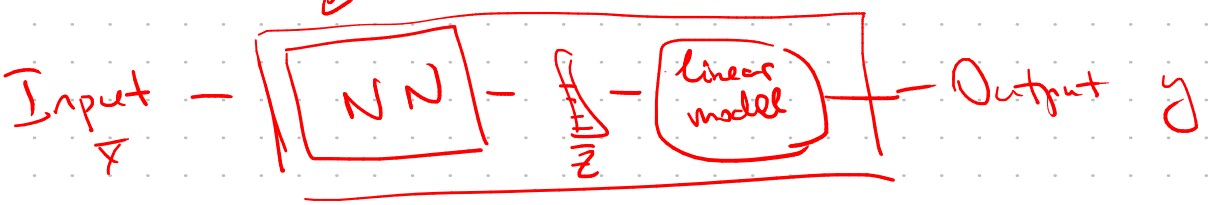
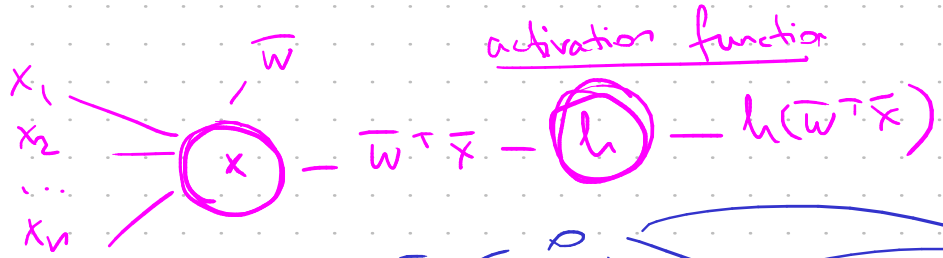
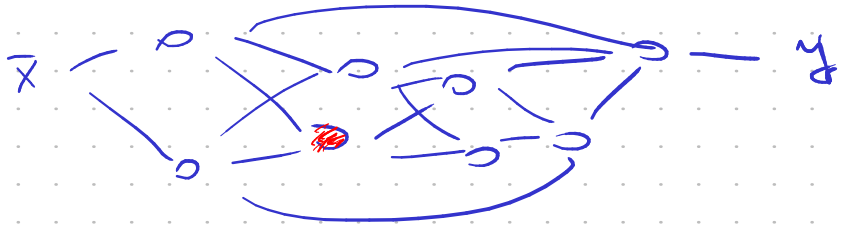
① Recap



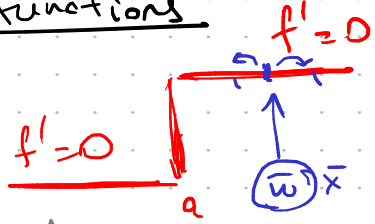






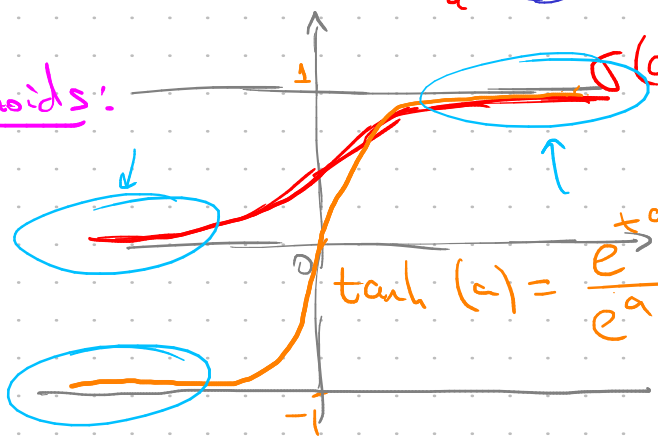
② Activation functions

1943, 1958 :



if $\bar{w}^T\bar{x} > a$ then 1 else 0

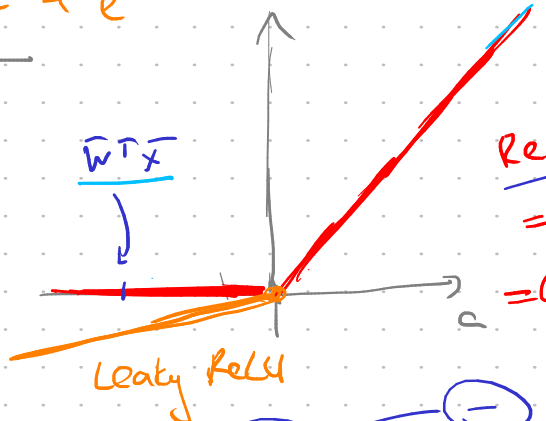
Sigmoids:



$$\sigma(a) = \frac{1}{1+e^{-a}}$$

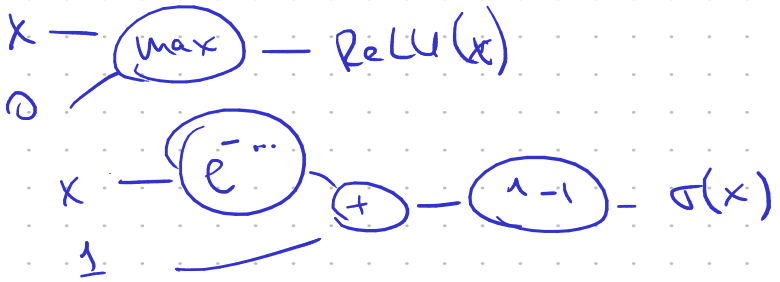
$$\tanh(a) = \frac{e^{+a} - e^{-a}}{e^{a} + e^{-a}}$$

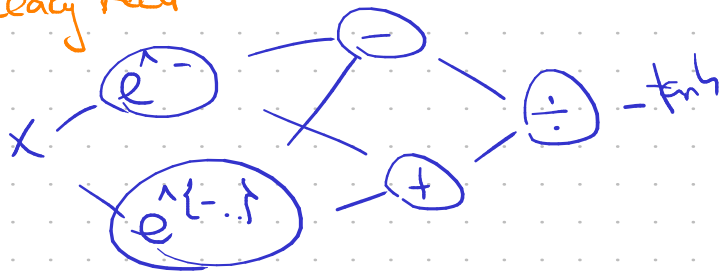
2014, 2012 ReLU rectified linear unit



$$ReLU(a) = \max(0, a) = a \text{ if } a > 0 \text{ else } 0$$

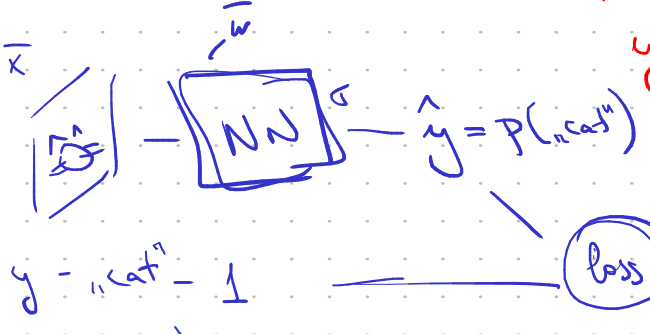
ReLU / Swish / GELU / Leaky ReLU





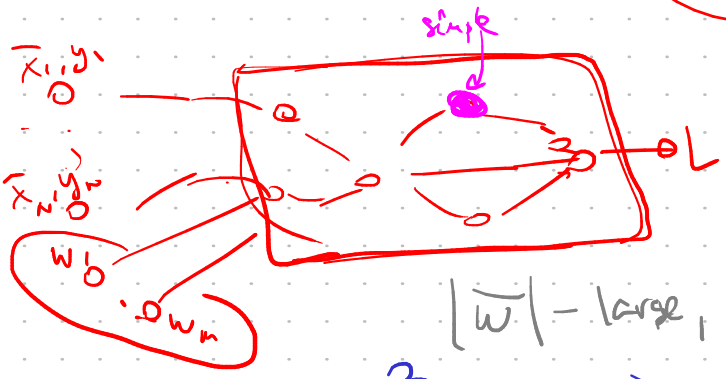
(3) Backpropagation

$\bar{x} \to f \to l(\bar{x}, y, \bar{w})$, f takes y and $\bar{w}$



$$L(\bar{w}) = \frac{1}{N} \sum_{n=1}^{N} l(\bar{x}_n, y_n, \bar{w}) \to \min_{\bar{w}}$$

gradient descent: $\bar{w}^{(k+1)} := \bar{w}^{(k)} - \eta \nabla_{\bar{w}} L(\bar{w}^{(k)})$

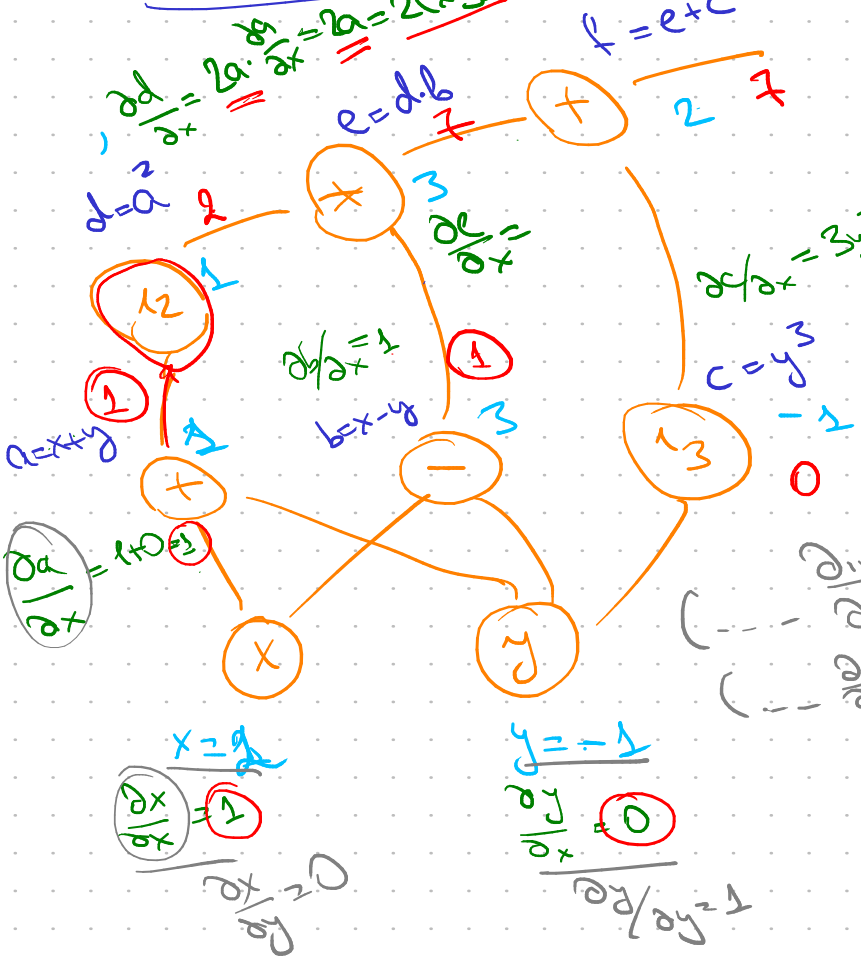


$\nabla_{\bar{w}} L|_{\_\_} = ?$

$|\bar{w}|$ - large, millions

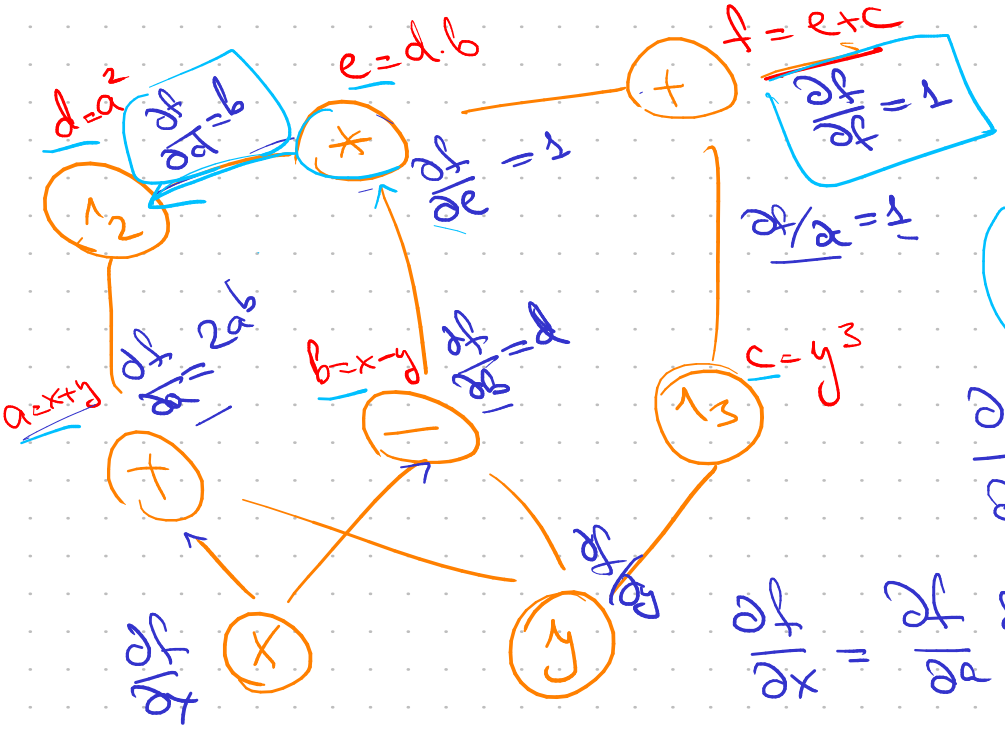
$f(x,y) = (x+y)^2 \cdot (x-y) + y$ $\quad \nabla f = \left(\frac{\partial f}{\partial x} \; \frac{\partial f}{\partial y}\right)$

forward propagation / fprop



$$\frac{\partial f}{\partial x} = \frac{\partial e}{\partial x} + \frac{\partial c}{\partial x} = 2(x+y)(x-y) + 1 \cdot (x+y)^2$$

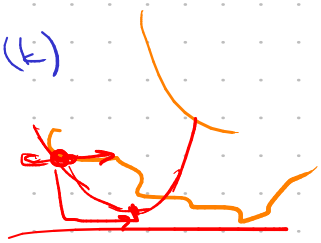
$$= (x+y)^2 + 2(x+y)(x-y)$$



(4) Stochastic gradient descent

~~GD~~ - first-order approx.

$$L(\bar{w}) = L(\bar{w}^{(k)}) + (\bar{w}-\bar{w}^{(k)})^T \underbrace{\nabla_{\bar{w}} L\big|_{\bar{w}^{(k)}}}_{=\bar{g}_k} + \frac{1}{2}(\bar{w}-\bar{w}^{(k)})^T \cdot \underbrace{H_k}_{\text{Hessian}} (\bar{w}-\bar{w}^{(k)}) + O(\dots)$$



$d = 10^6$

$\bar{w} \in \mathbb{R}^d$

$$H_k = \left(\frac{\partial^2 L}{\partial w_i \partial w_j} \right)_{i,j}$$

$$\bar{g}_k + H_k(\bar{w}-\bar{w}^{(k)}) = 0$$

Newton's method

$$\bar{w}^{(k+1)} = \bar{w}^{(k)} - H_k^{-1} \cdot \bar{g}_k$$

quasi-Newton algorithms

L-BFGS

$\bar{w}_1, \bar{w}_2, \dots$

$\bar{g}_1, \bar{g}_2, \dots$

update $\hat{H}^{-1}$

$k \to k+1$

$$L(\bar{w}) = \frac{1}{N} \sum_{n=1}^{N} l(\bar{x}_n, y_n, \bar{w}) = -\frac{1}{N} \sum_{n=1}^{N} \log p(\bar{w} \mid \bar{x}_n, y_n)$$

NN

$$\bar{g}_k = \frac{1}{N} \sum_{n} \nabla_{\bar{w}} l(\bar{x}_n, y_n, \bar{w}) \Big|_{\bar{w}^{(k)}}$$

Stochastic gradient descent

$$L(\bar{w}) = \mathbb{E}_D \left[l(\bar{x}, y, \bar{w}) \right] \to \min_{\bar{w}}$$

$(\bar{x}, y) \sim \text{unif}(D)$

Stochastic optimization

$$\bar{g}_k = \mathbb{E}_D \left[\nabla_{\bar{w}} l(\bar{x}, y, \bar{w}) \right] \approx \hat{g}_k = \frac{1}{m} \sum_{i=1}^{m} \nabla_{\bar{w}} l(\bar{x}_i, y_i, \bar{w})$$

$|D| = 10^6$

$m = 4, 8$

mini-batch

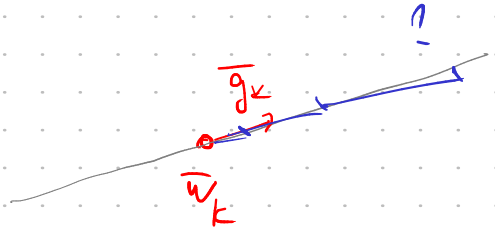
$$\bar{w}^{(k+1)} := \bar{w}^{(k)} - \eta_k \hat{g}_k$$ - SGD

$k \to \infty$? $\eta_k \to 0$

$\eta_k = \frac{1}{k}$

$\sum \eta_k \to \infty$

$\sum \eta_k^2 < \infty$



$f(\bar{x}) \to \min$

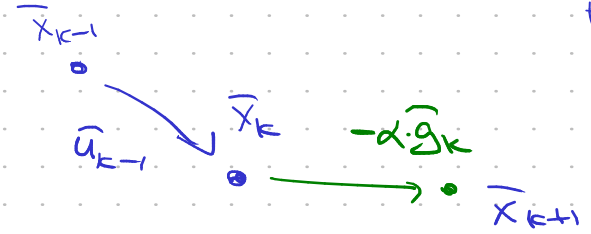
$F(\bar{x} + \alpha \cdot \bar{q}) = f(\alpha) \to \min$

5 SGD improvements & variations

examples: $f(x,y) = x^2 + y \cdot y^2 \xrightarrow[x,y]{} \min$

$F(\bar{x}) \xrightarrow[\bar{x}]{} \min$

a) Notation

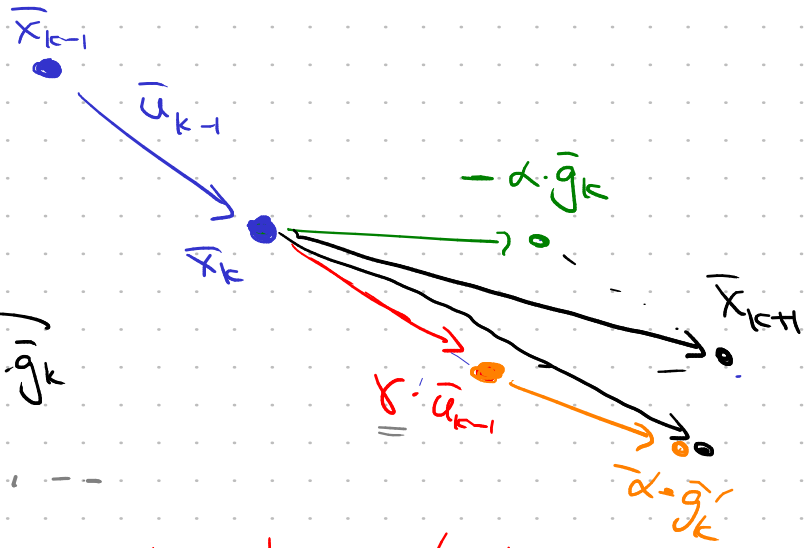


$\bar{u}_{k-1} = \bar{x}_k - \bar{x}_{k-1}$

SGD: $\bar{x}_{k+1} = \bar{x}_k - \alpha \cdot \bar{g}_k$

$\bar{u}_k = -\alpha \cdot \bar{g}_k$

1) Momentum



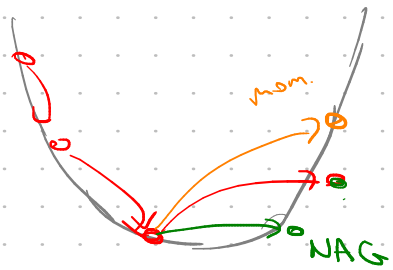
$\bar{u}_k$

$\bar{x}_{k+1} = \bar{x}_k + \underbrace{\gamma \cdot \bar{u}_{k-1} - \alpha \cdot \bar{g}_k}$

$\gamma = .95, .99, \ldots$

2) Nesterov accelerated gradients / NAG

$\bar{x}_{k+1} = \bar{x}_k + \underbrace{\gamma \cdot \bar{u}_{k-1} - \alpha \cdot \nabla_{\bar{x}} F(\bar{x}_k + \gamma \cdot \bar{u}_{k-1})}_{\bar{u}_k}$



$F(\bar{x})$ $(x_1, x_2, \ldots, x_d)$ } adaptive SGD

$\alpha_{1k}\ \alpha_{2k}\ \ldots\ \alpha_{dk}$

3) Adaptive SGD variations

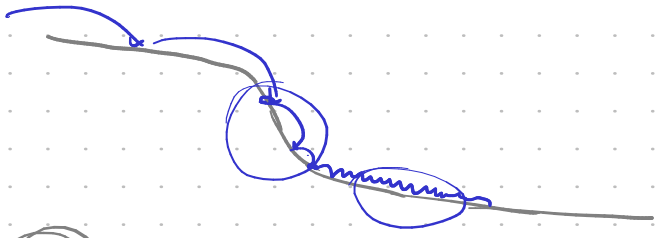
AdaGrad

$$x_{k+1,i} = x_{k,i} - \frac{\alpha}{\sqrt{G_{k,i} + \varepsilon}} \cdot g_{k,i}$$

$$G_{k,i} = G_{k-1,i} + g_{k,i}^2$$

Exponential Moving Average

$$G_{k,i} = \gamma \cdot G_{k-1,i} + (1-\gamma) \cdot g^2_{k,i} =$$



$$= (1-\gamma) g^2_{k,i} + \gamma \cdot (1-\gamma) \cdot g^2_{k-1,i} + \gamma^2 \cdot (1-\gamma) \cdot g^2_{k-2,i} + \dots$$

$\gamma < 1$

RMSprop

$\frac{\partial f}{\partial x}$ $\quad \bar{g}$ $\quad m/s$ $\quad \bar{x}$ - s, $\quad F(\bar{x})$ - m

$$\bar{x} := \bar{x} - \alpha \cdot \bar{g}$$

s $\quad$ s $\quad$ 1 $\quad$ m/s

$$\bar{x} := \bar{x} - \frac{\alpha}{\sqrt{G}} \cdot \bar{g}$$

s $\quad$ s $\quad$ m/s

$$\bar{x} := \bar{x} - H^{-1} \cdot \bar{g}$$

s $\quad$ s $\quad$ $(m/s^2)^{-1}$ $\quad$ m/s

Adadelta $\quad u = u^2$

$$\frac{\alpha \cdot \sqrt{u+\varepsilon}}{}$$

- **Adam**

$$G_{k,i} = \beta_2 G_{k-1,i} + (1-\beta_2) g^2_{k,i}$$

$$x_{k+1,i} = x_{k,i} - \frac{\alpha}{\sqrt{G_{k,i} + \varepsilon}} \cdot m_{k,i} \qquad m_{k,i} = \beta_1 m_{k-1,i} + (1-\beta_1) \cdot g_{k,i}$$

$\beta_1 = 0.9, \quad \beta_2 = 0.999, \quad \varepsilon = 10^{-8}$

- **weight decay** $\quad L'(\bar{w}) := L(\bar{w}) + \alpha \cdot \|\bar{w}\|^2 \longrightarrow \min$

$$\bar{w}_{k+1} := \bar{w}_k \cdot \gamma - \alpha \cdot \bar{g}_k =$$

$\gamma < 1$

AdamW

$$= \bar{w}_k - (1-\gamma)\bar{w}_k - \alpha \cdot \nabla_{\bar{w}} L(\bar{w}_k) =$$

$$= \bar{w}_k - \alpha \cdot \nabla_{\bar{w}} \left[L(\bar{w}) + \frac{1-\gamma}{\alpha} \bar{w}^T \bar{w} \right]_{\bar{w}_k}$$

Lion, **Muon**