Gradient Descent

$\min f(x)$ $\quad x \in \mathbb{R}^n$

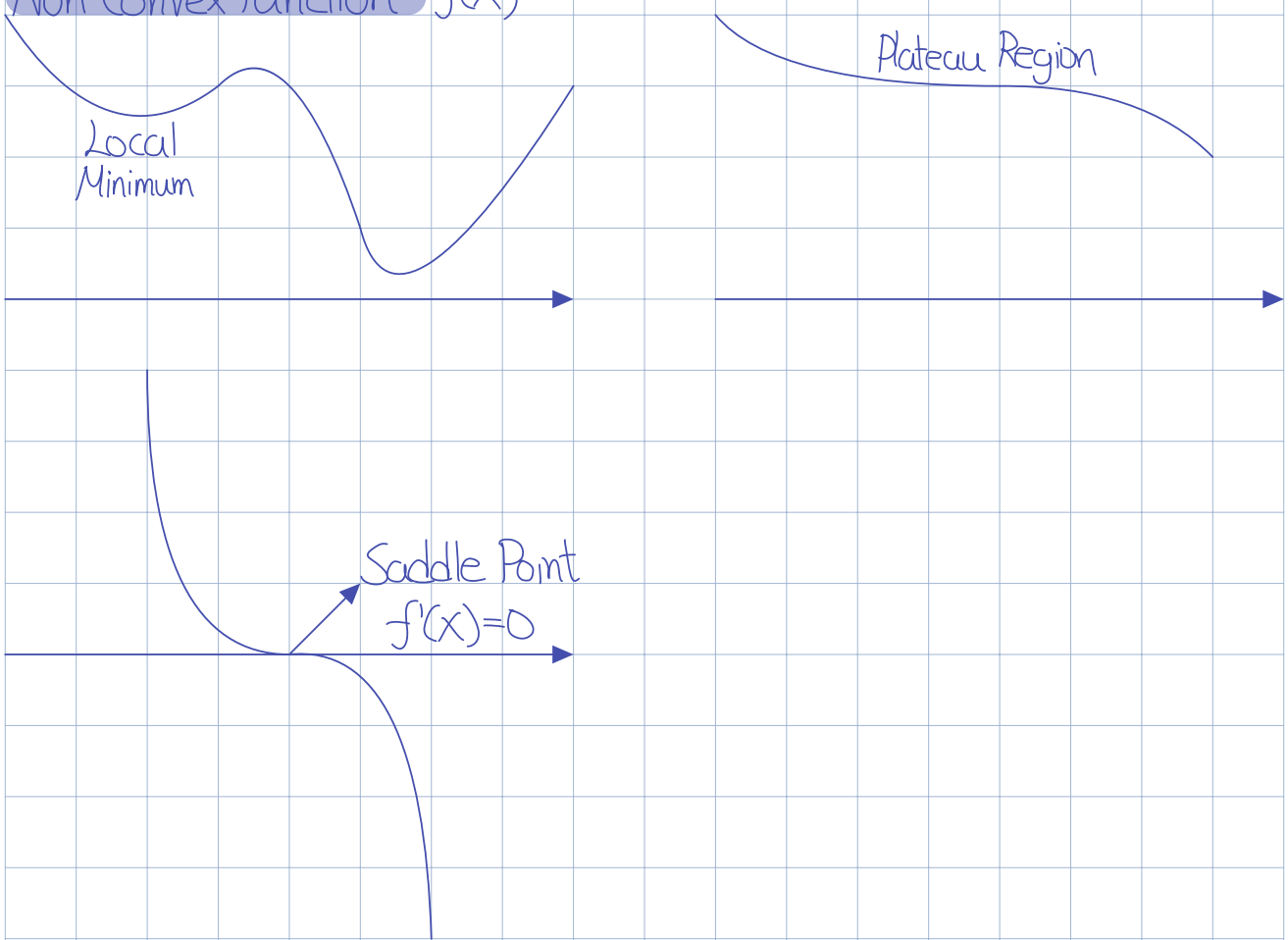
$$x_{t+1} = x_t - \varepsilon_t \nabla f(x_t)$$

Convex Function

$\| \nabla f(x_t) \|_2 = 0$

Can be the stopping criteria

Non Convex Function $f(x)$



Rule of Thumb

1. Large number of iterations
2. $f(x)$ must be small
3. $\|\nabla f(x_t)\|_2$ must be close to 0

Momentum Based Model

Recap Basic SGD Rule

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

$$g_t = \nabla e_n(w_t)$$

$$w_{t+1} = w_t + \varepsilon_t g_t$$

SGD + Momentum

Velocity Vector: v_t $\quad v_0 = 0$

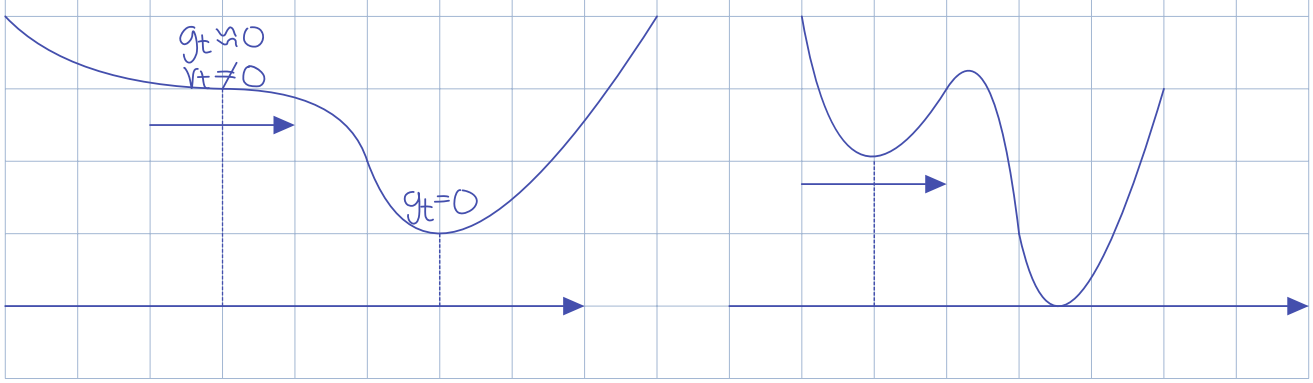
$$g_t = \nabla e_n(w_t)$$

$$v_t = -\varepsilon_t g_t + \mu v_{t-1}$$

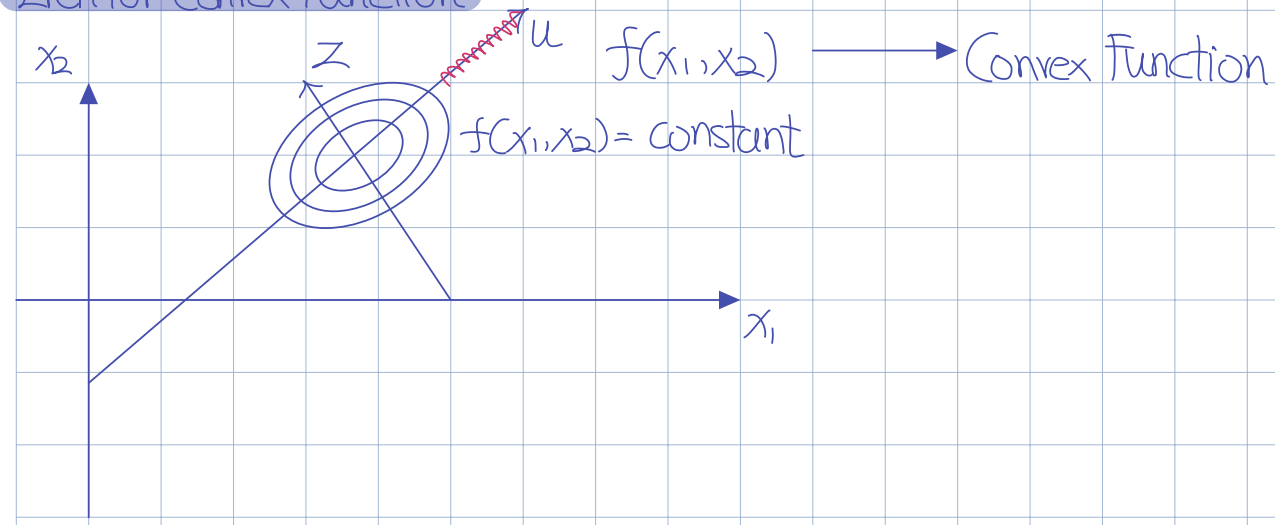
$$w_{t+1} = w_t + v_t$$

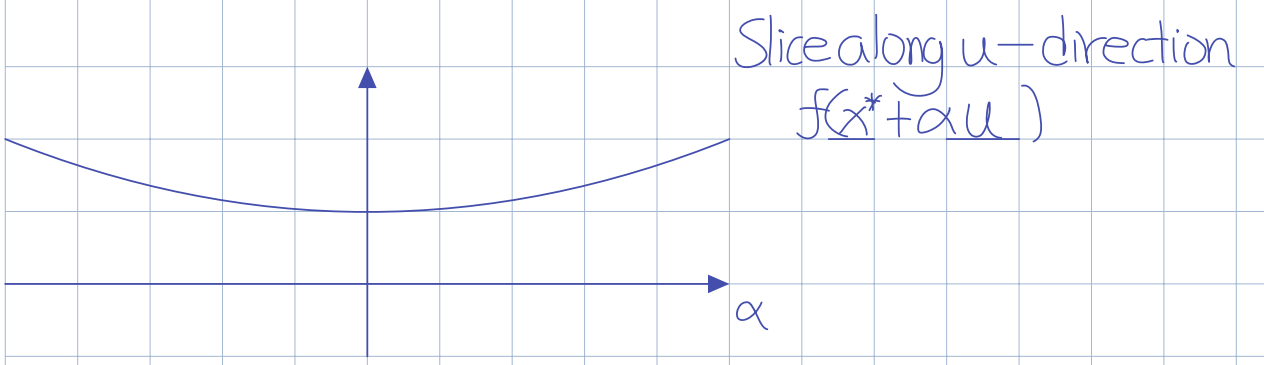
μ = momentum coefficient

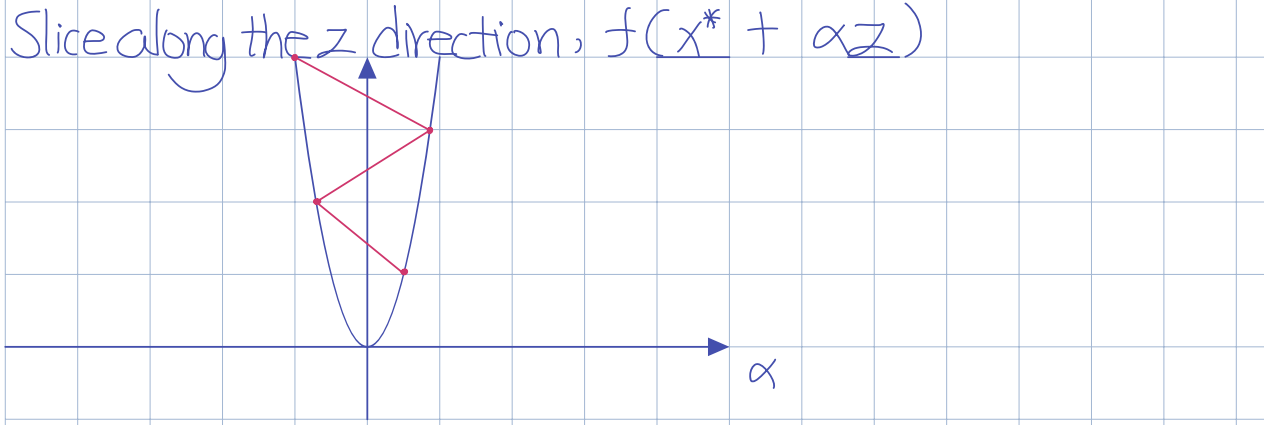
$\mu \approx 0.9$



Momentum Based Method Can Led to Faster Convergence Even For Convex Function



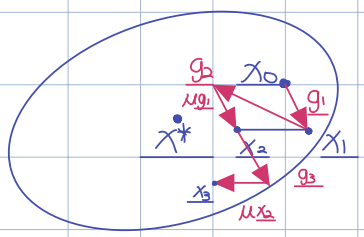




Some direction, small gradient

Some direction, large gradient

Application of Momentum Method ($\epsilon_t = 1$)



Step 1, $v_0 = 0$

$g_1 = -\nabla f(x_0)$

$v_1 = g_1$

$x_1 = x_0 + v_1$

Step 2,

$g_2 = -\nabla f(x_1)$

$v_2 = g_2 + \mu g_1$

$x_2 = x_1 + v_2$

Step 3,

$g_3 = -\nabla f(x_2)$

$v_3 = g_3 + \mu g_2$

$x_3 = x_2 + v_3$

Step 4,

$g_4 = -\nabla f(x_3)$

$v_4 = g_4 + \mu g_3$

$x_4 = x_3 + v_4$

Nestror Momentum

$$v_t = \mu v_{t-1} - \varepsilon_t \nabla e_n(w_t + \mu v_{t-1})$$

Intermedicute Point Between w_t and w_{t+1}

$$w_{t+1} = w_t + v_t$$
$$= w_t + \mu v_{t-1} - \varepsilon_t \nabla e_n(w_t + \mu v_{t-1})$$

For Convex Functions, "Convergence Rate" (1983)

1. GD, $O(1/k)$
2. Nestror Momentum, $O(1/k^2)$

AdaGrad (adaptive step size)

$$g_t = \nabla e_n(w_t) \in \mathbb{R}^d$$

$$r_t = r_{t-1} + g_t \odot g_t$$ Componentwise Vector Multiplication

$$r_t(j) = \sum_{q=1}^{t} g_q^2(j), \; j = 1, 2, \ldots, d$$

$$w_{t+1}(j) = w_t(j) - \frac{\varepsilon_t}{\sqrt{r_t(j)}} g_t(j)$$

RMS – PROP

$$r_t = (1-\beta)\, r_{t-1} + \beta\, g(t) \odot g(t)$$

β = decay factor

$$w_{t+1} = w_t - \frac{\varepsilon_t}{\sqrt{r(t)}} \odot g(t)$$

ADAM (RMS – PROP + Momentum)

$$g_t = \nabla en(w_t)$$

$$r_t = \beta\, r_{t-1} + (1-\beta)\, g_t \odot g_t$$

$$v_t = \mu\, v_{t-1} - \frac{\varepsilon_t}{\sqrt{r(t)}} \odot g(t)$$

$$w_{t+1} = w_t + v_t$$