

# Redes Neurais e Aprendizagem Profunda

## ATUALIZAÇÃO DE PESOS *NAG – NESTEROV ACCELERATED GRADIENT*

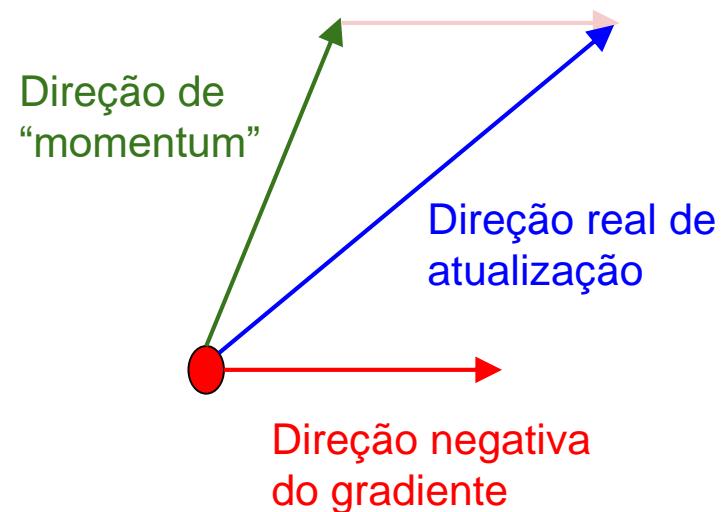
---

Zenilton K. G. Patrocínio Jr

[zenilton@pucminas.br](mailto:zenilton@pucminas.br)

# Atualização pelo “*Momentum*” de Nesterov

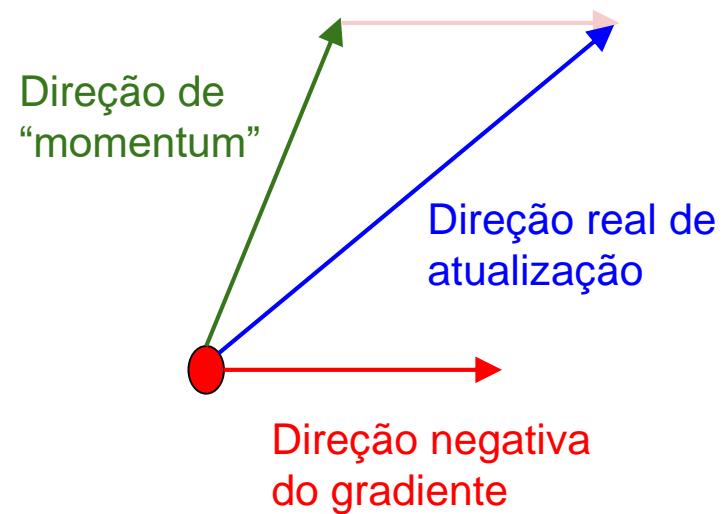
## Atualização pelo “*Momentum*”



```
# Momentum update
v = mu * v - learning_rate * dx
x += v # integrate position
```

# Atualização pelo “*Momentum*” de Nesterov

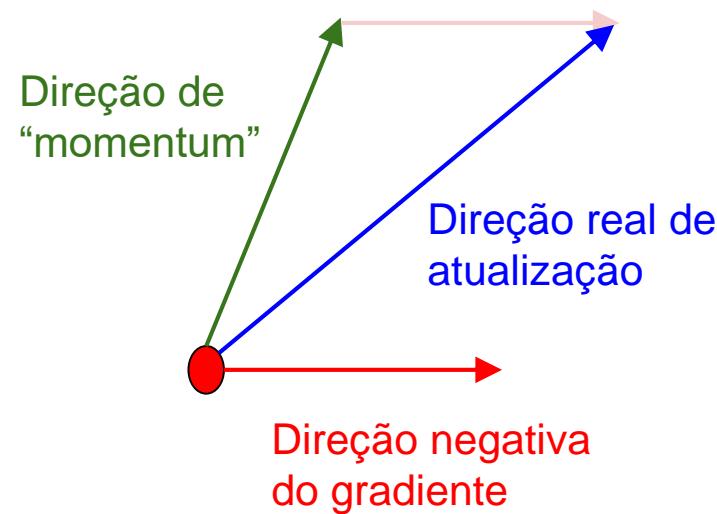
## Atualização pelo “*Momentum*”



$$\begin{aligned} p^{(t+1)} &= \mu p^{(t)} - \alpha \nabla_W L(W^{(t)}) \\ W^{(t+1)} &= W^{(t)} + p^{(t+1)} \end{aligned}$$

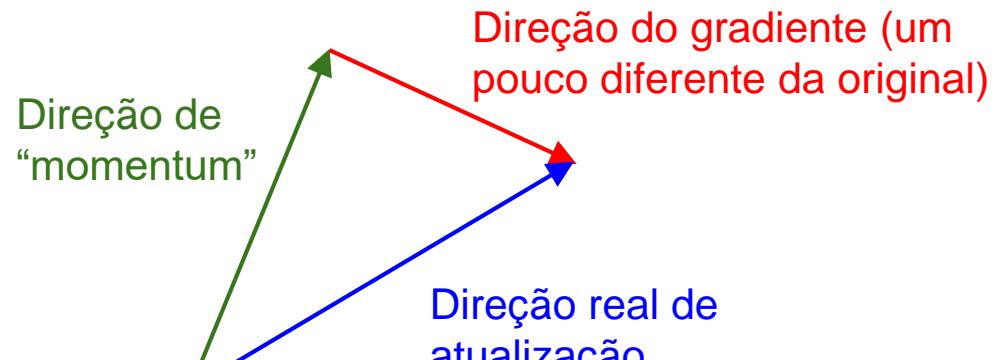
# Atualização pelo “*Momentum*” de Nesterov

Atualização pelo “*Momentum*”



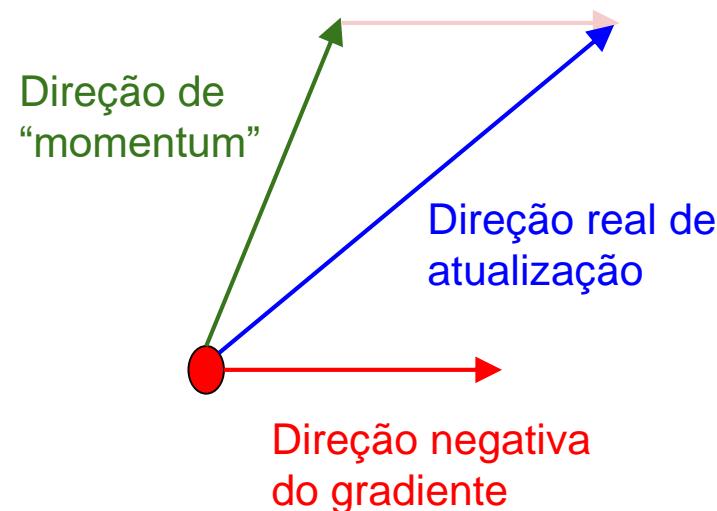
$$\begin{aligned} p^{(t+1)} &= \mu p^{(t)} - \alpha \nabla_W L(W^{(t)}) \\ W^{(t+1)} &= W^{(t)} + p^{(t+1)} \end{aligned}$$

Atualização pelo “*Momentum*” de Nesterov



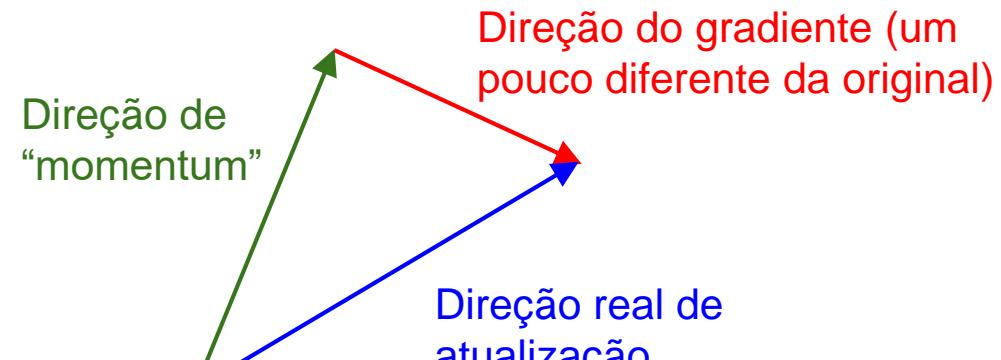
# Atualização pelo “Momentum” de Nesterov

Atualização pelo “Momentum”



$$\begin{aligned} p^{(t+1)} &= \mu p^{(t)} - \alpha \nabla_W L(W^{(t)}) \\ W^{(t+1)} &= W^{(t)} + p^{(t+1)} \end{aligned}$$

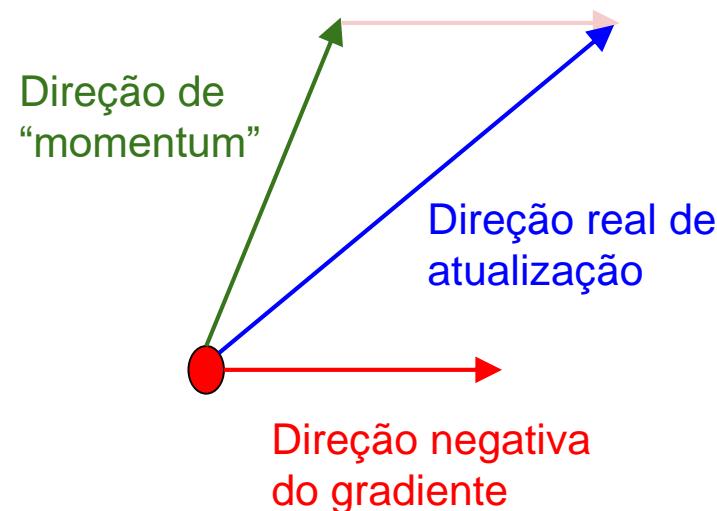
Atualização pelo “Momentum” de Nesterov



$$\begin{aligned} p^{(t+1)} &= \mu p^{(t)} - \alpha \nabla_W L(W^{(t)} + \mu p^{(t)}) \\ W^{(t+1)} &= W^{(t)} + p^{(t+1)} \end{aligned}$$

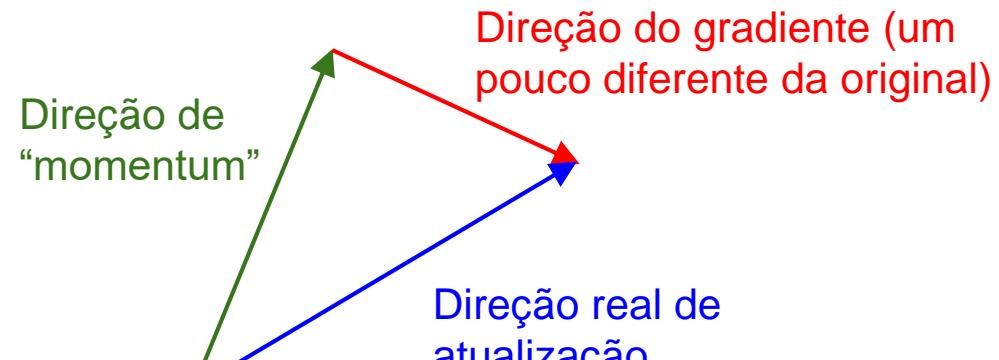
# Atualização pelo “*Momentum*” de Nesterov

Atualização pelo “*Momentum*”



$$\begin{aligned} p^{(t+1)} &= \mu p^{(t)} - \alpha \nabla_W L(W^{(t)}) \\ W^{(t+1)} &= W^{(t)} + p^{(t+1)} \end{aligned}$$

Atualização pelo “*Momentum*” de Nesterov



Nesterov: a única diferença ...

$$\begin{aligned} p^{(t+1)} &= \mu p^{(t)} - \alpha \nabla_W L(W^{(t)} + \boxed{\mu p^{(t)}}) \\ W^{(t+1)} &= W^{(t)} + p^{(t+1)} \end{aligned}$$

# SGD × “*Momentum*” de Nesterov

Um SGD cuidadosamente implementado possui uma taxa de convergência de  $O\left(\frac{1}{t}\right)$ , isto é, seu erro é limitado por uma constante vezes  $\frac{1}{t}$ , em que  $t$  é o número de passos

O “momentum” de Nesterov apresenta uma taxa de convergência de  $O\left(\frac{1}{t^2}\right)$ , que é mais rápida

OBS: Ambos os limites se aplicam a problemas de otimização convexa

Veja p.ex.

Attouch; Peypouquet, “*The rate of convergence of Nesterov’s accelerated forward-backward method is actually faster than  $1/k^2$* ” arXiv 1510.08740, 2015.

# Atualização pelo “*Momentum*” de Nesterov

$$\begin{aligned} p^{(t+1)} &= \mu p^{(t)} - \alpha \nabla_W L(W^{(t)} + \mu p^{(t)}) \\ W^{(t+1)} &= W^{(t)} + p^{(t+1)} \end{aligned}$$

# Atualização pelo “*Momentum*” de Nesterov

$$\begin{aligned} p^{(t+1)} &= \mu p^{(t)} - \alpha \nabla_W L(W^{(t)} + \mu p^{(t)}) \\ W^{(t+1)} &= W^{(t)} + p^{(t+1)} \end{aligned}$$

Pequeno inconveniente...  
Geralmente, temos:

$$W^{(t)}, g^{(t)} = \nabla_W L(W^{(t)})$$

# Atualização pelo “*Momentum*” de Nesterov

$$\begin{aligned} p^{(t+1)} &= \mu p^{(t)} - \alpha \nabla_W L(W^{(t)} + \mu p^{(t)}) \\ W^{(t+1)} &= W^{(t)} + p^{(t+1)} \end{aligned}$$

Pequeno inconveniente...  
Geralmente, temos:

$$W^{(t)}, g^{(t)} = \nabla_W L(W^{(t)})$$

---

Transformação e reorganização de variáveis :

$$X^{(t)} = W^{(t)} + \mu p^{(t)}$$

# Atualização pelo “*Momentum*” de Nesterov

$$\begin{aligned} p^{(t+1)} &= \mu p^{(t)} - \alpha \nabla_W L(W^{(t)} + \mu p^{(t)}) \\ W^{(t+1)} &= W^{(t)} + p^{(t+1)} \end{aligned}$$

Pequeno inconveniente...  
Geralmente, temos:

$$W^{(t)}, g^{(t)} = \nabla_W L(W^{(t)})$$

---

Transformação e reorganização de variáveis :

$$X^{(t)} = W^{(t)} + \mu p^{(t)}$$

$$W^{(t)} = X^{(t)} - \mu p^{(t)}$$

# Atualização pelo “*Momentum*” de Nesterov

$$\begin{aligned} p^{(t+1)} &= \mu p^{(t)} - \alpha \nabla_W L(W^{(t)} + \mu p^{(t)}) \\ W^{(t+1)} &= W^{(t)} + p^{(t+1)} \end{aligned}$$

Pequeno inconveniente...  
Geralmente, temos:

$$W^{(t)}, g^{(t)} = \nabla_W L(W^{(t)})$$

---

Transformação e reorganização de variáveis :

$$X^{(t)} = W^{(t)} + \mu p^{(t)}$$

Substituir todos  $W$ s por  $X$ s, e rearranjar:

$$W^{(t)} = X^{(t)} - \mu p^{(t)}$$

$$\begin{aligned} p^{(t+1)} &= \mu p^{(t)} - \alpha \nabla_W L(X^{(t)}) \\ X^{(t+1)} &= X^{(t)} - \mu p^{(t)} + (1 + \mu) p^{(t+1)} \end{aligned}$$

# Atualização pelo “*Momentum*” de Nesterov

$$\begin{aligned} p^{(t+1)} &= \mu p^{(t)} - \alpha \nabla_W L(W^{(t)} + \mu p^{(t)}) \\ W^{(t+1)} &= W^{(t)} + p^{(t+1)} \end{aligned}$$

Pequeno inconveniente...  
Geralmente, temos:

$$W^{(t)}, g^{(t)} = \nabla_W L(W^{(t)})$$

---

Transformação e reorganização de variáveis :

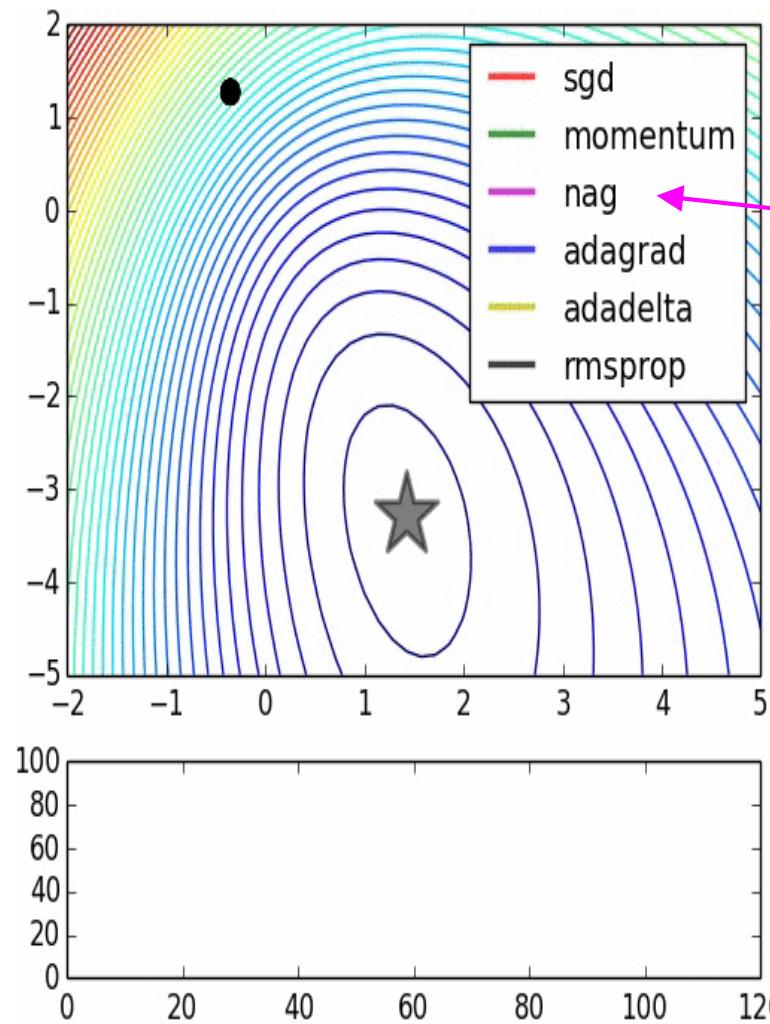
$$X^{(t)} = W^{(t)} + \mu p^{(t)}$$

Substituir todos  $W$ s por  $X$ s, e rearranjar:

$$\begin{aligned} p^{(t+1)} &= \mu p^{(t)} - \alpha \nabla_W L(X^{(t)}) \\ X^{(t+1)} &= X^{(t)} - \mu p^{(t)} + (1 + \mu) p^{(t+1)} \end{aligned}$$

```
# Nesterov momentum update rewrite
v_prev = v
v = mu * v - learning_rate * dx
x += -mu * v_prev + (1 + mu) * v
```

# Atualização pelo “*Momentum*” de Nesterov



NAG = Nesterov  
Accelerated Gradient