ATUALIZAÇÃO DE PESOS

MOMENTUM

Zenilton K. G. Patrocínio Jr
zenilton@pucminas.br

SGD – Treinamento de Rede Neural

SGD – *loop* principal:

```
while True:
    data_batch = dataset.sample_data_batch()
    loss = network.forward(data_batch)
    dx = network.backward()
    x += - learning_rate * dx
```

SGD – Treinamento de Rede Neural

SGD – *loop* principal:

```
while True:
    data_batch = dataset.sample_data_batch()
    loss = network.forward(data_batch)
    dx = network.backward()
    x += - learning_rate * dx
```

SGD – Treinamento de Rede Neural

SGD – *loop* principal:

```
while True:
  data_batch = dataset.sample_data_batch()
  loss = network.forward(data_batch)
  dx = network.backward()
  x += - learning_rate * dx
```

SGD – Treinamento de Rede Neural

SGD – *loop* principal:

```
while True:
    data_batch = dataset.sample_data_batch()
    loss = network.forward(data_batch)
    dx = network.backward()
    x += - learning_rate * dx
```

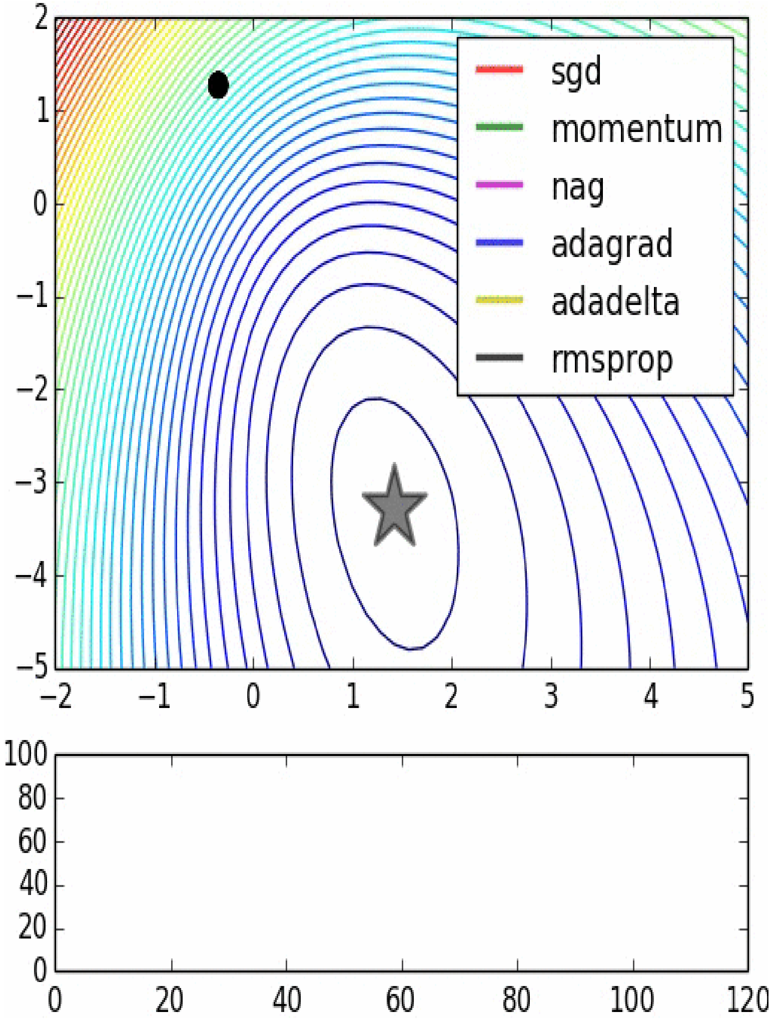
SGD – Treinamento de Rede Neural

SGD – *loop* principal:

```
while True:
    data_batch = dataset.sample_data_batch()
    loss = network.forward(data_batch)
    dx = network.backward()
    x += - learning_rate * dx
```

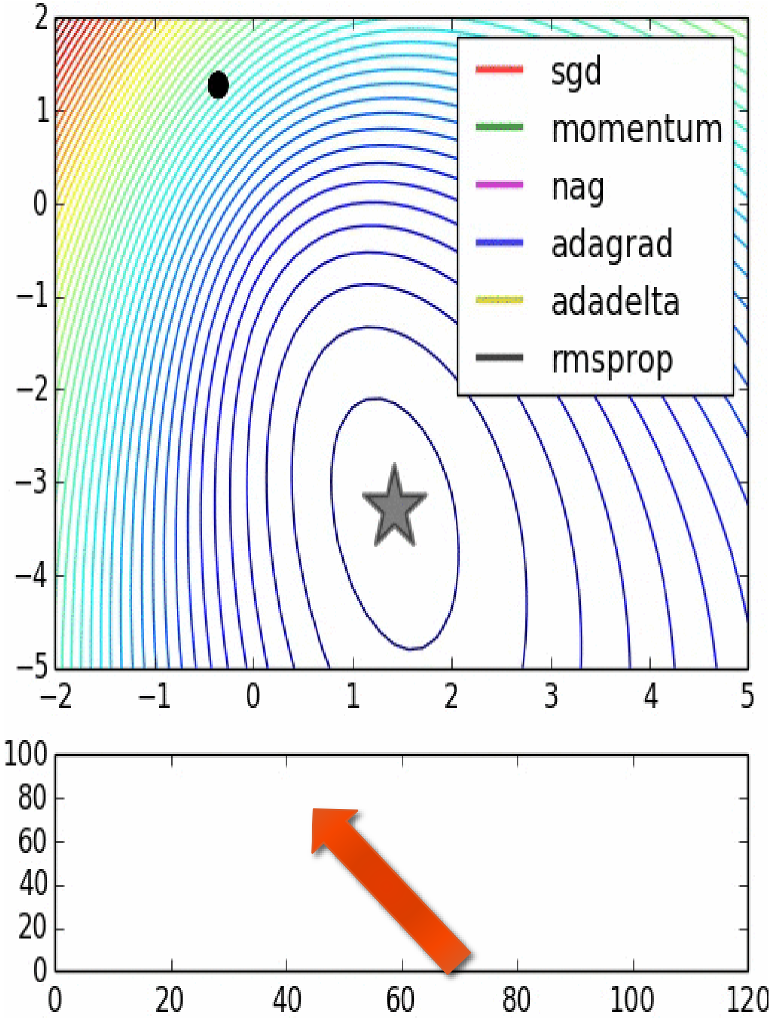
Atualização por meio de descida mais íngreme
(ou gradiente simples)

Vários Esquemas de Atualização de Pesos



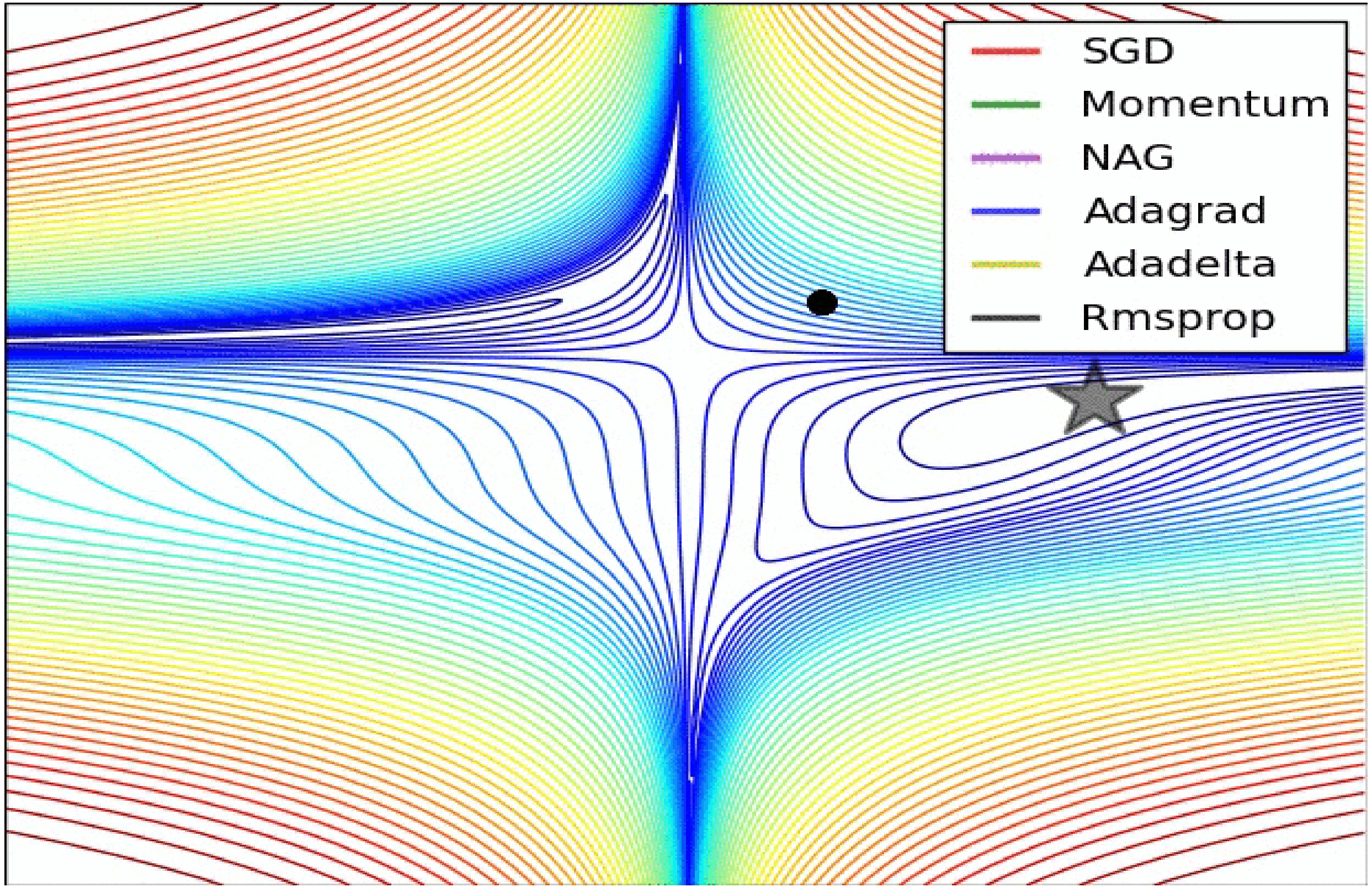
Crédito: Alec Radford, 2015

Vários Esquemas de Atualização de Pesos



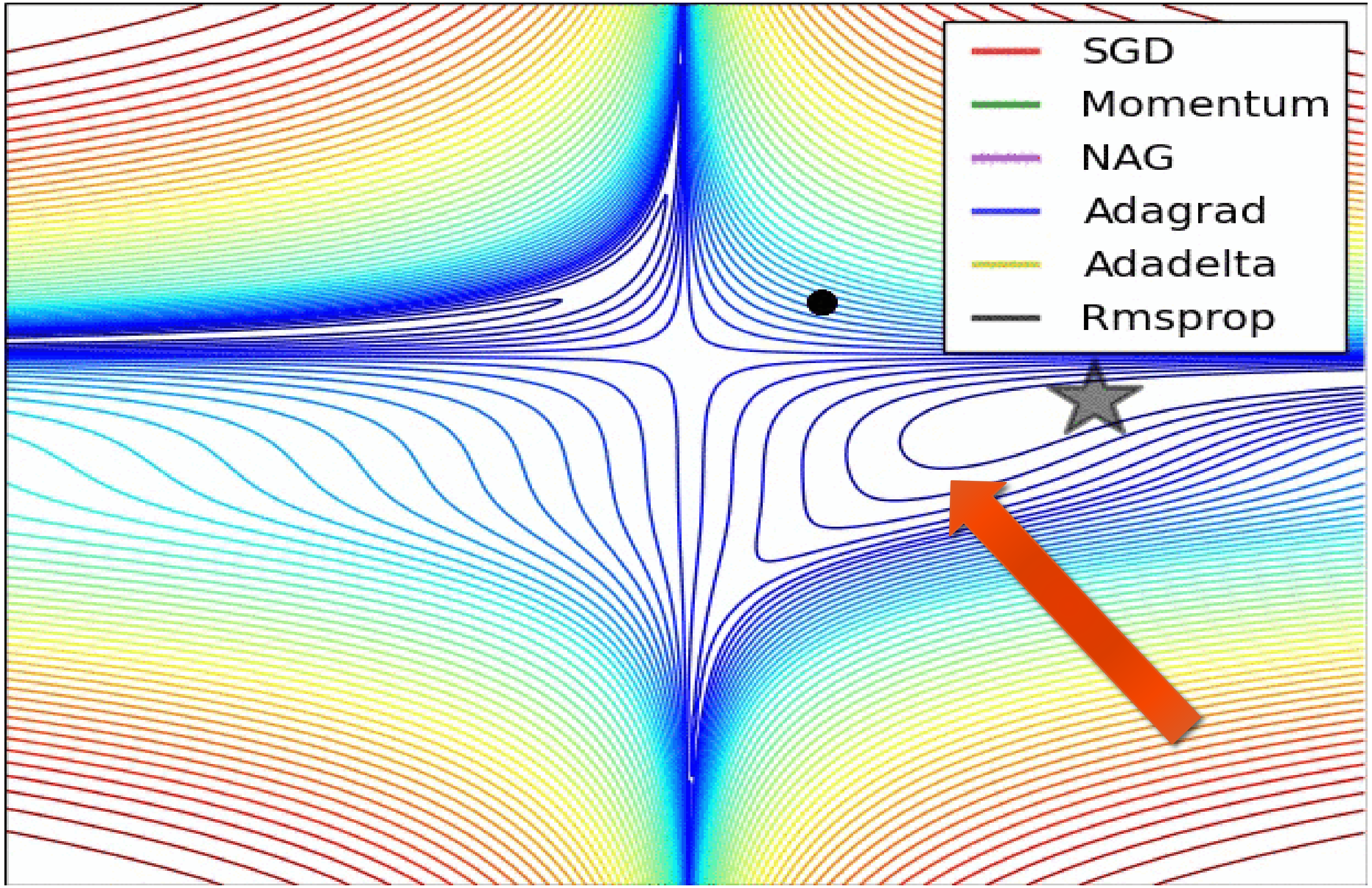
Crédito: Alec Radford, 2015

Vários Esquemas de Atualização de Pesos



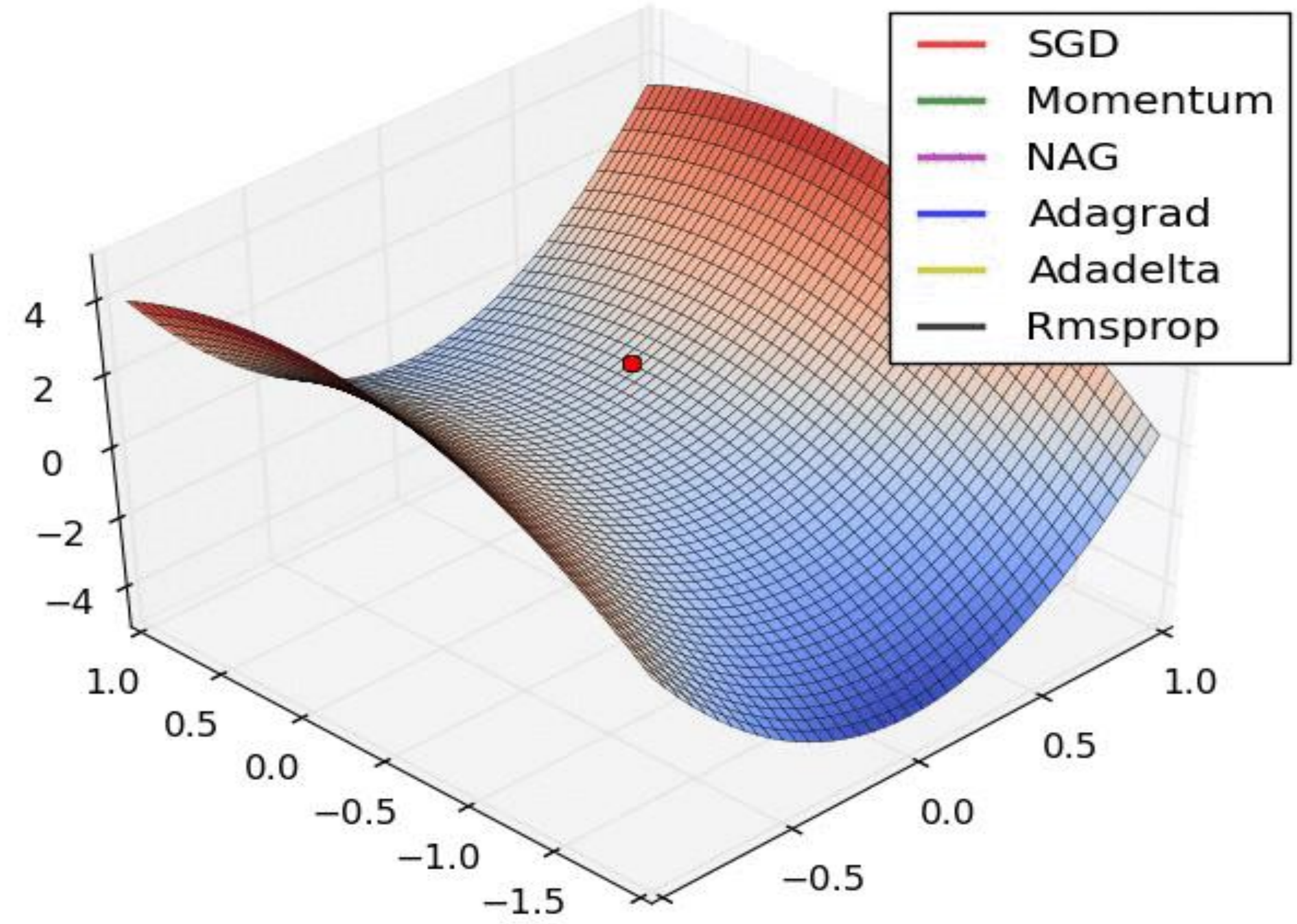
Crédito: Alec Radford, 2015

Vários Esquemas de Atualização de Pesos



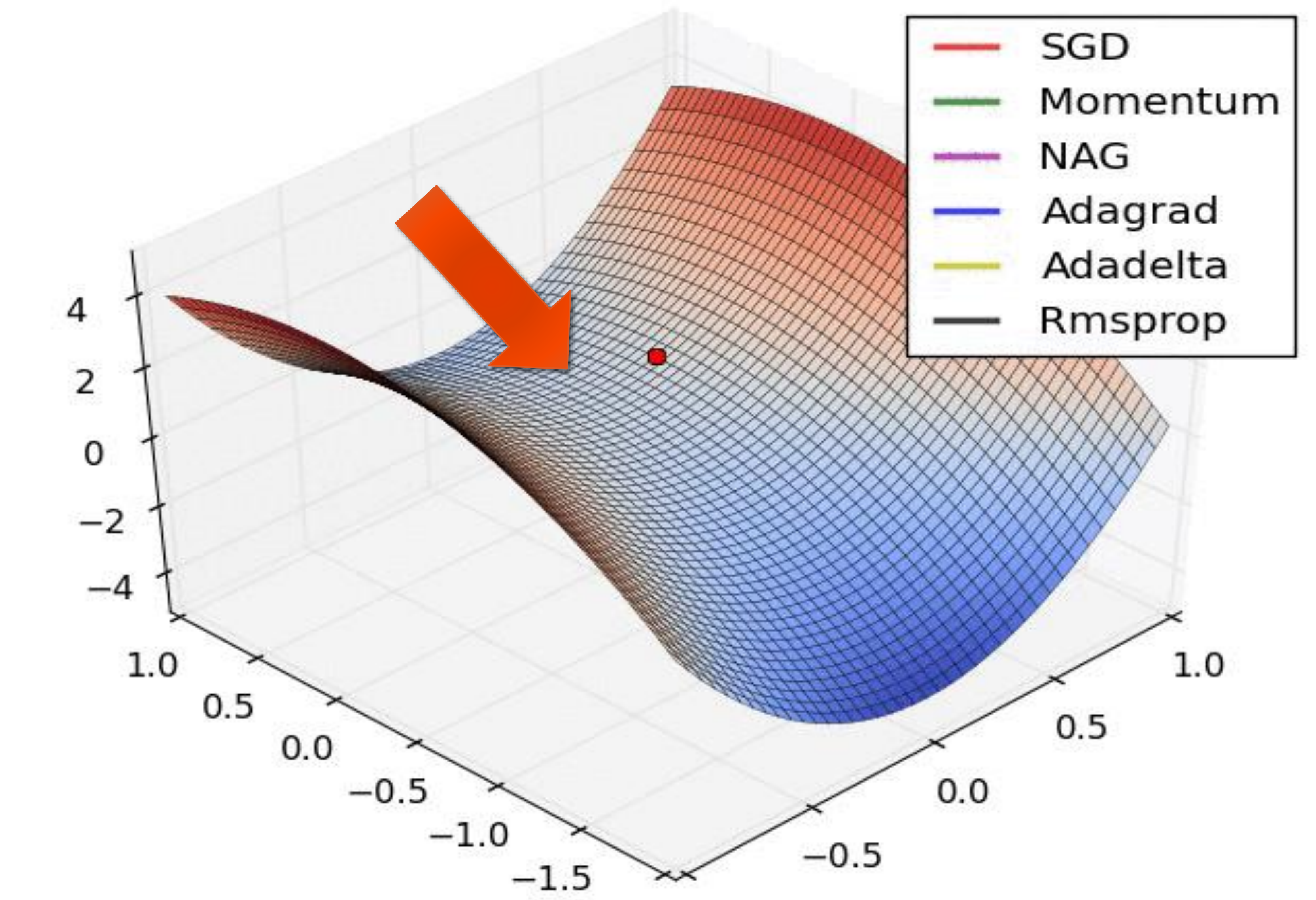
Crédito: Alec Radford, 2015

Vários Esquemas de Atualização de Pesos



Crédito: Alec Radford, 2015

Vários Esquemas de Atualização de Pesos



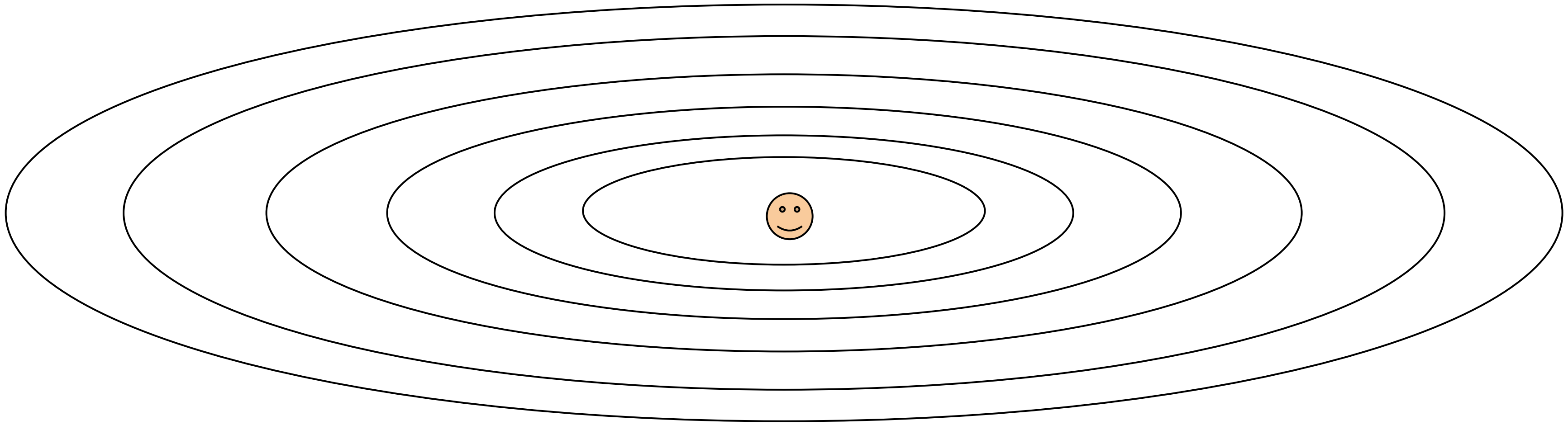
Crédito: Alec Radford, 2015

Problema com SGD

Suponha que a função de perda seja íngreme verticalmente, mas rasa horizontalmente

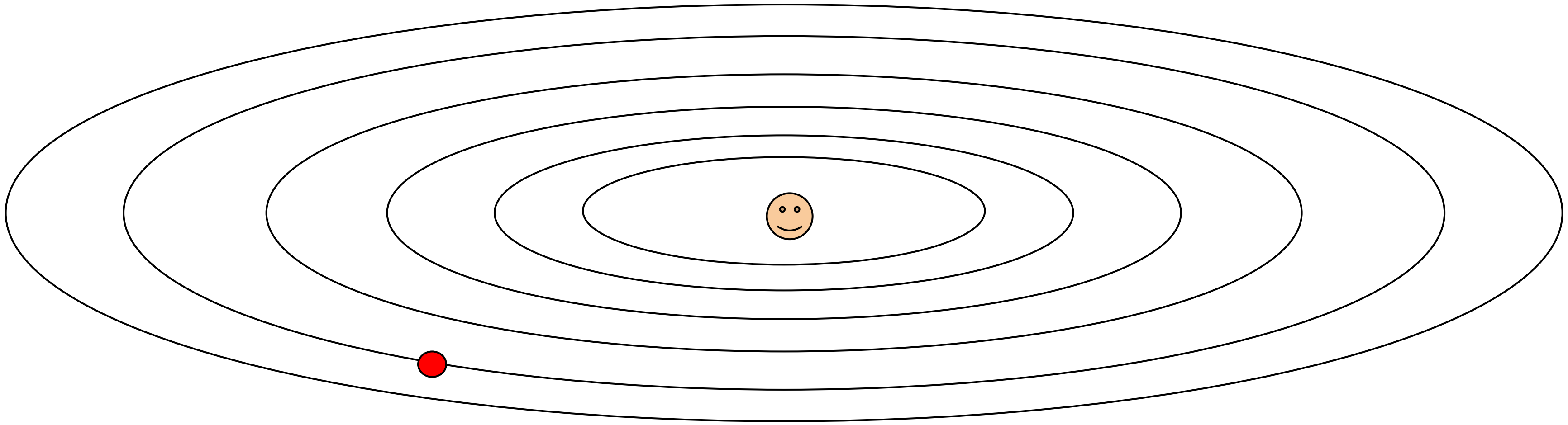
Problema com SGD

Suponha que a função de perda seja íngreme verticalmente, mas rasa horizontalmente



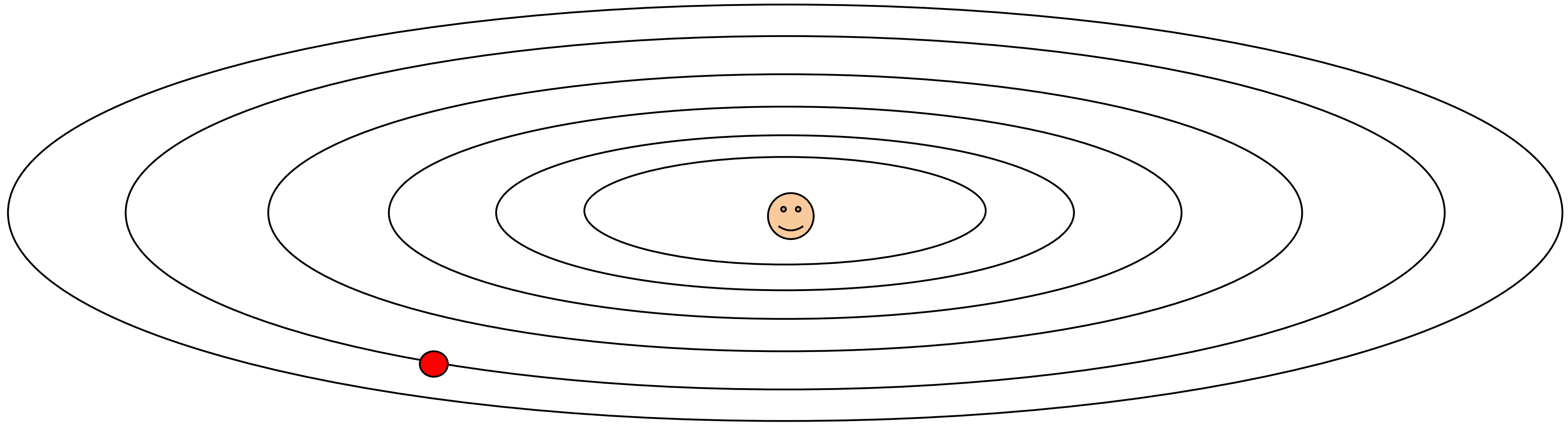
Problema com SGD

Suponha que a função de perda seja íngreme verticalmente, mas rasa horizontalmente



Problema com SGD

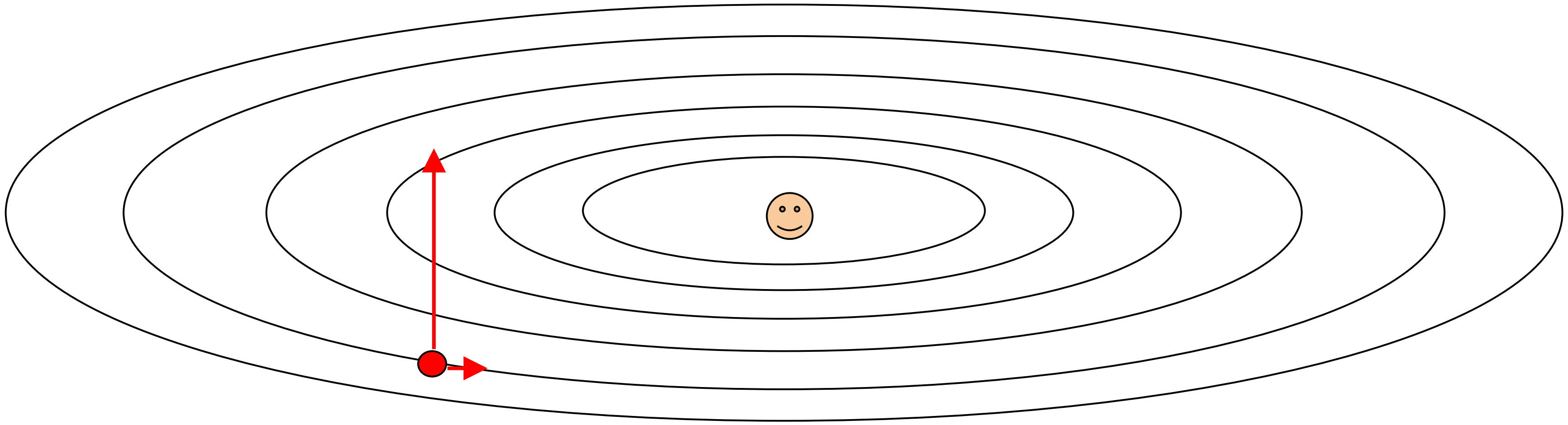
Suponha que a função de perda seja íngreme verticalmente, mas rasa horizontalmente



P: Qual é a trajetória que o SGD usa para alcançar o mínimo?

Problema com SGD

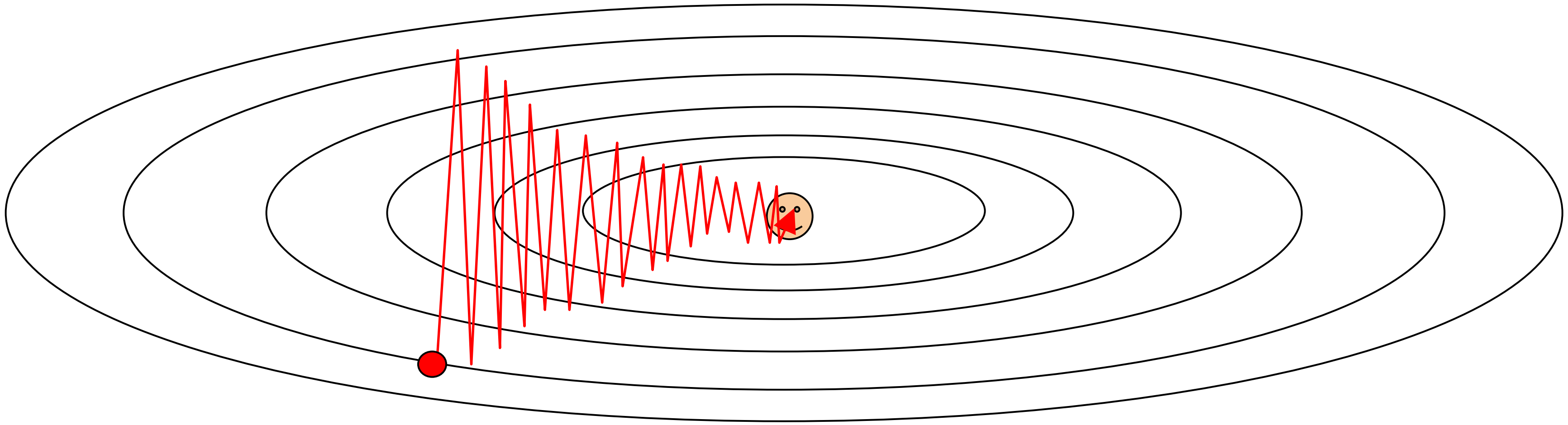
Suponha que a função de perda seja íngreme verticalmente, mas rasa horizontalmente



P: Qual é a trajetória que o SGD usa para alcançar o mínimo?

Problem with SGD

Suponha que a função de perda seja íngreme verticalmente, mas rasa horizontalmente



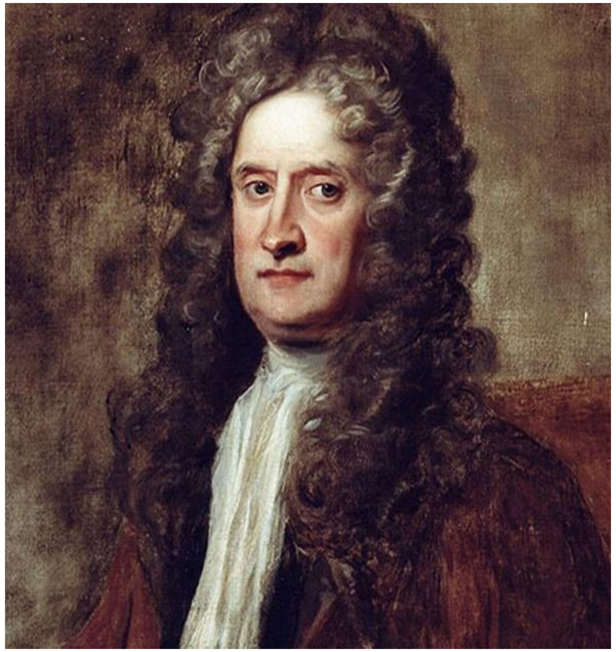
P: Qual é a trajetória que o SGD usa para alcançar o mínimo?

Progresso muito lento na horizontal, “zig-zag” na vertical

SGD com *Momentum*

“Todo corpo persiste em seu estado de repouso ou de se mover uniformemente para a frente, exceto na medida em que é obrigado a mudar de estado por uma força impressa.”

– Isaac Newton

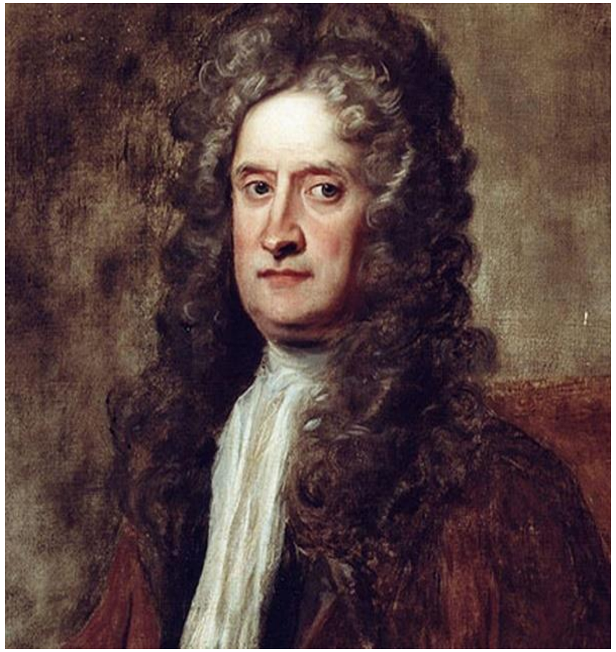


SGD com *Momentum*

“Todo corpo persiste em seu estado de repouso ou de se mover uniformemente para a frente, exceto na medida em que é obrigado a mudar de estado por uma força impressa.”

– Isaac Newton

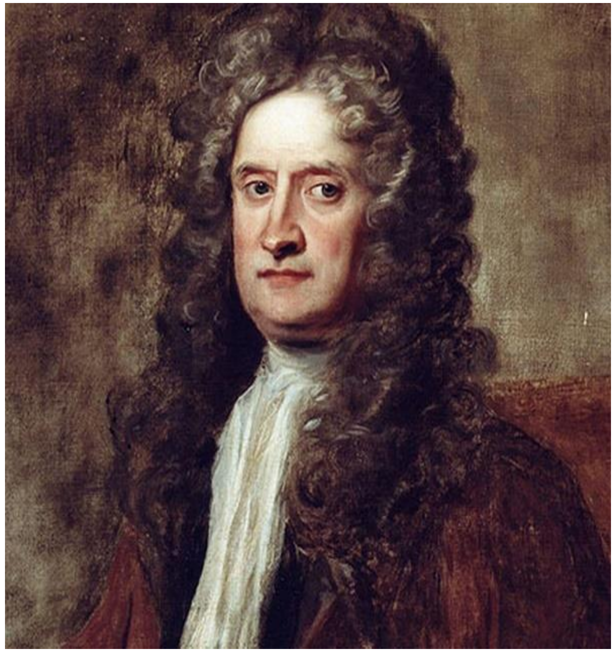
A “memória” do objeto de seu estado de movimento é ***momentum***



SGD com *Momentum*

"Todo corpo persiste em seu estado de repouso ou de se mover uniformemente para a frente, exceto na medida em que é obrigado a mudar de estado por uma força impressa."

– Isaac Newton



A "memória" do objeto de seu estado de movimento é ***momentum***

Como é comumente usado em aprendizagem profunda, o parâmetro "*momentum*" é na verdade a "taxa de decaimento do *momentum* por minibatch"

(E seu análogo físico seria a viscosidade)

SGD com *Momentum*

Regra de atualização do "*momentum*" para o passo t:

$$p^{(t+1)} = \mu p^{(t)} - \alpha g^{(t)}$$

em que

- $p^{(t)}$ representa o "*momentum*",

SGD com *Momentum*

Regra de atualização do "*momentum*" para o passo t:

$$p^{(t+1)} = \mu p^{(t)} - \alpha g^{(t)}$$

em que

- $p^{(t)}$ representa o "*momentum*",
- $\mu \in [0,1]$ é a constante de "*momentum*",

SGD com *Momentum*

Regra de atualização do "*momentum*" para o passo t:

$$p^{(t+1)} = \mu p^{(t)} - \alpha g^{(t)}$$

em que

- $p^{(t)}$ representa o "*momentum*",
- $\mu \in [0,1]$ é a constante de "*momentum*",
- $g^{(t)}$ é o gradiente do *minibatch*, isto é,

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

SGD com *Momentum*

Regra de atualização do "*momentum*" para o passo t:

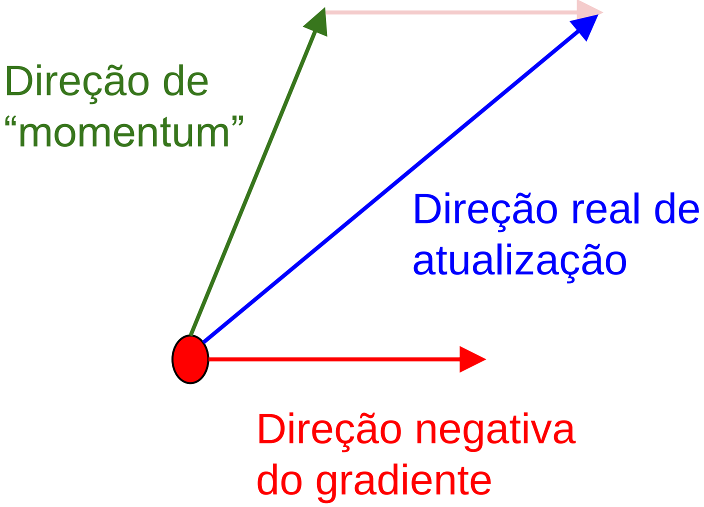
$$p^{(t+1)} = \mu p^{(t)} - \alpha g^{(t)}$$

em que

- $p^{(t)}$ representa o "*momentum*",
- $\mu \in [0,1]$ é a constante de "*momentum*",
- $g^{(t)}$ é o gradiente do *minibatch*, isto é,

  $g^{(t)} = \nabla_W L(W^{(t)})$, e
- α é a taxa de aprendizado

Atualização pelo "*Momentum*"



SGD com *Momentum*

Regra de atualização do "*momentum*" para o passo t:

$$p^{(t+1)} = \mu p^{(t)} - \alpha g^{(t)}$$

em que

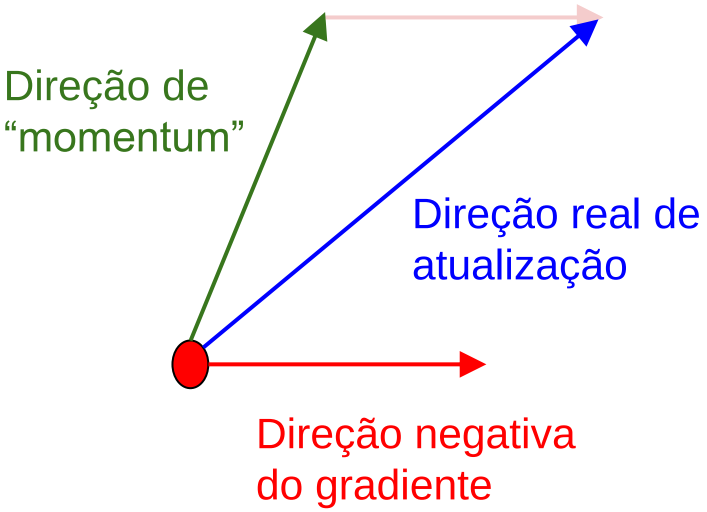
- $p^{(t)}$ representa o "*momentum*",
- $\mu \in [0,1]$ é a constante de "*momentum*",
- $g^{(t)}$ é o gradiente do *minibatch*, isto é,

  $g^{(t)} = \nabla_W L(W^{(t)})$, e
- α é a taxa de aprendizado

A atualização de pesos é dados por:

$$W^{(t+1)} = W^{(t)} + p^{(t+1)}$$

Atualização pelo "*Momentum*"



Atualização pelo “*Momentum*”

```
# Gradient descent update
x += - learning_rate * dx
```

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

Atualização pelo "*Momentum*"

```
# Gradient descent update
x += - learning_rate * dx
```

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

- Interpretação física como uma bola descendo pela função de perda + atrito (coeficiente μ)
- μ = geralmente ~0,5; 0,9 ou 0,99 (às vezes ajustado ao longo do tempo, p.ex. de 0,5 $\rightarrow$ 0,99)

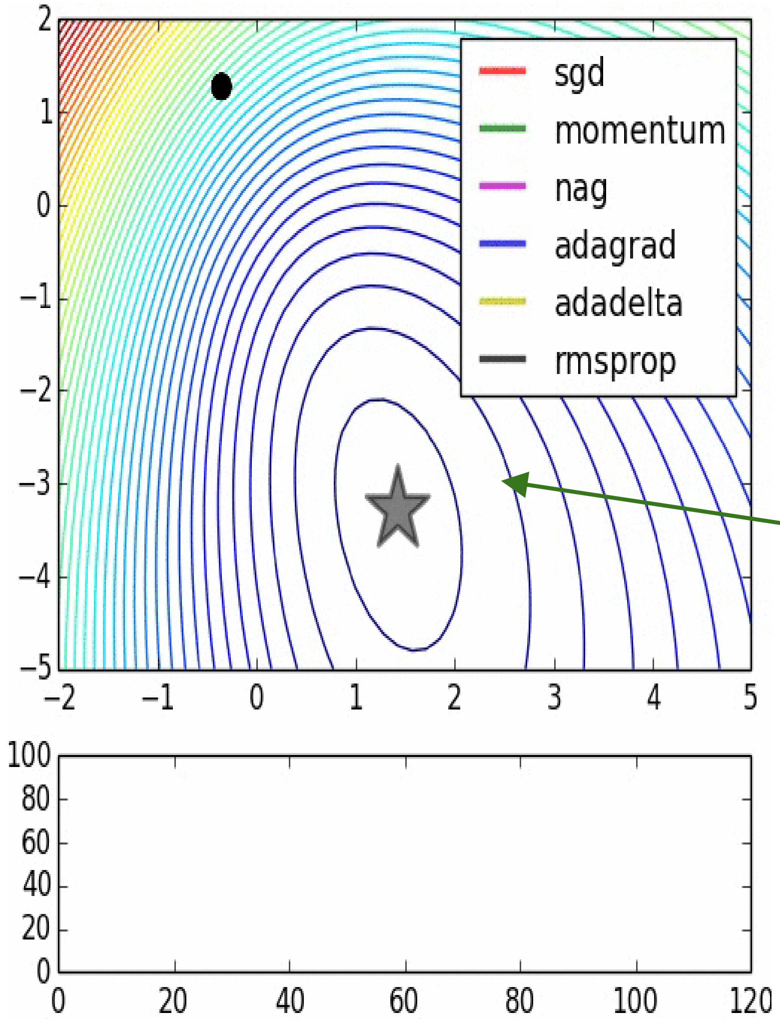
Atualização pelo “*Momentum*”

```
# Gradient descent update
x += - learning_rate * dx
```

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

- Interpretação física como uma bola descendo pela função de perda + atrito (coeficiente μ)
- μ = geralmente ~0,5; 0,9 ou 0,99 (às vezes ajustado ao longo do tempo, p.ex. de 0,5 $\rightarrow$ 0,99)
- Permite que a velocidade “se acumule” nas direções rasas
- Reduz a velocidade nas direções íngremes devido à mudança rápida de sinal

SGD × SGD+*Momentum*



Observe que o “momentum” ultrapassar a meta, mas, em geral, chega ao mínimo muito mais rápido