

## Часть 4: Методы оптимизации

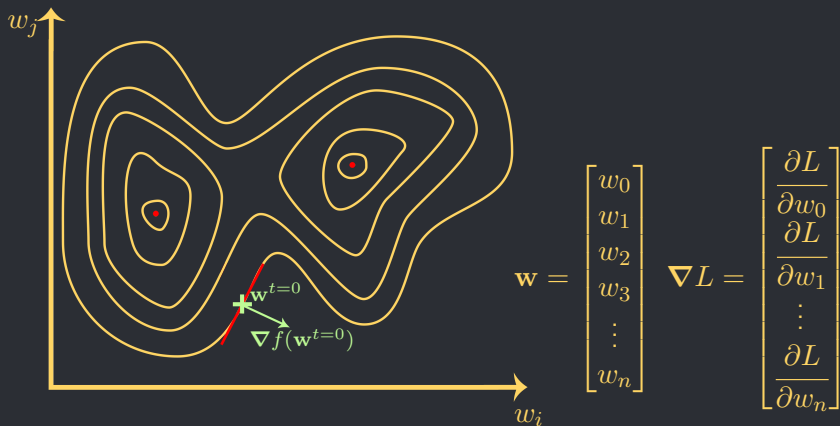
---

Романов Михаил, Игорь Слинько

# Градиентный спуск



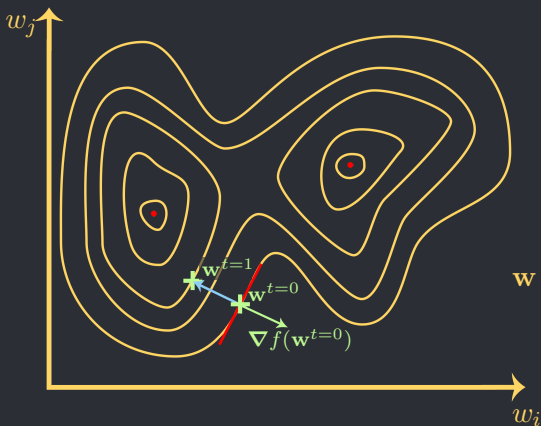
# Градиентный спуск



# Градиентный спуск

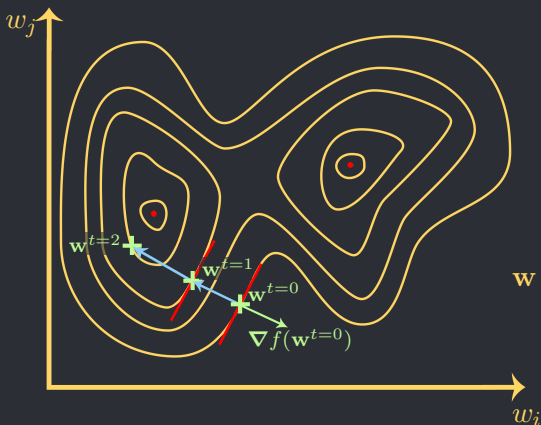
$$\mathbf{w}^{t=0}$$

$$\mathbf{w}^{t=1} = \mathbf{w}^{t=0} - \alpha \nabla L(\mathbf{w}^{t=0})$$



$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} \quad \nabla L = \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_n} \end{bmatrix}$$

# Градиентный спуск



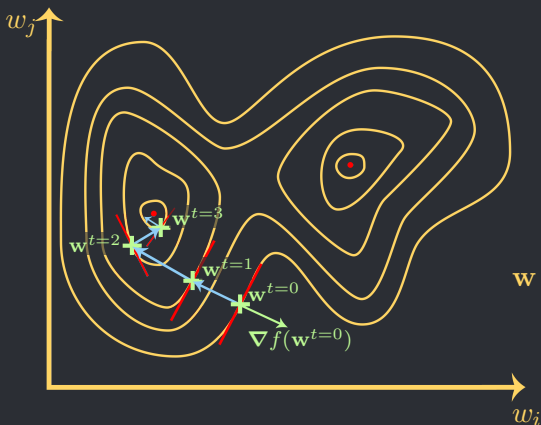
$$\mathbf{w}^{t=0}$$

$$\mathbf{w}^{t=1} = \mathbf{w}^{t=0} - \alpha \nabla L(\mathbf{w}^{t=0})$$

$$\mathbf{w}^{t=2} = \mathbf{w}^{t=1} - \alpha \nabla L(\mathbf{w}^{t=1})$$

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} \quad \nabla L = \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_n} \end{bmatrix}$$

# Градиентный спуск



$$\mathbf{w}^{t=0}$$

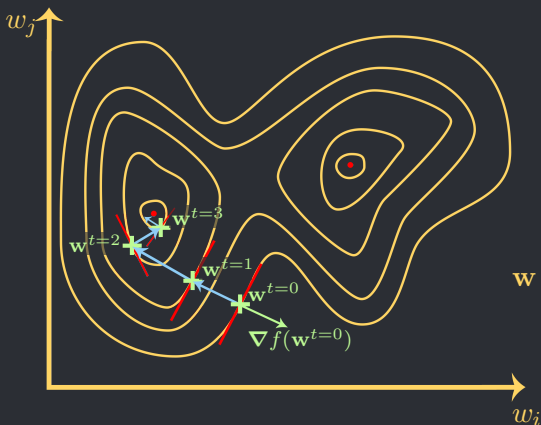
$$\mathbf{w}^{t=1} = \mathbf{w}^{t=0} - \alpha \nabla L(\mathbf{w}^{t=0})$$

$$\mathbf{w}^{t=2} = \mathbf{w}^{t=1} - \alpha \nabla L(\mathbf{w}^{t=1})$$

$$\mathbf{w}^{t=3} = \mathbf{w}^{t=2} - \alpha \nabla L(\mathbf{w}^{t=2})$$

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} \quad \nabla L = \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_n} \end{bmatrix}$$

# Градиентный спуск



$$\mathbf{w}^{t=0}$$

$$\mathbf{w}^{t=1} = \mathbf{w}^{t=0} - \alpha \nabla L(\mathbf{w}^{t=0})$$

$$\mathbf{w}^{t=2} = \mathbf{w}^{t=1} - \alpha \nabla L(\mathbf{w}^{t=1})$$

$$\mathbf{w}^{t=3} = \mathbf{w}^{t=2} - \alpha \nabla L(\mathbf{w}^{t=2})$$

$$\dots$$

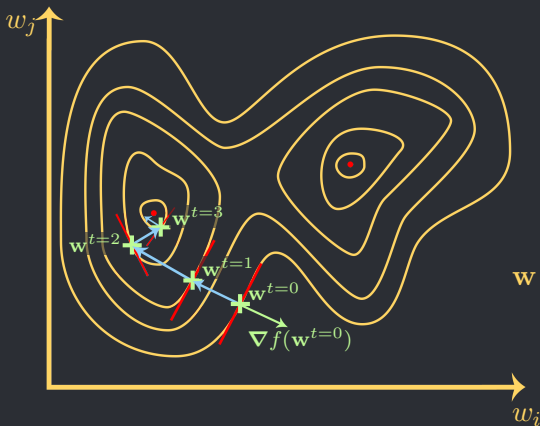
$$\mathbf{w}^T = \mathbf{w}^{T-1} - \alpha \nabla L(\mathbf{w}^{T-1})$$

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} \quad \nabla L = \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_n} \end{bmatrix}$$

# SGD

$$L = L_1 + L_2 + L_3 + \dots + L_S$$

Лосс функция – сумма лоссов  
на нескольких примерах



Градиент вычислен  
на одном примере

$$\mathbf{w}^{t=0}$$

$$\mathbf{w}^{t=1} = \mathbf{w}^{t=0} - \alpha \nabla L_{s_0}(\mathbf{w}^{t=0})$$

$$\mathbf{w}^{t=2} = \mathbf{w}^{t=1} - \alpha \nabla L_{s_1}(\mathbf{w}^{t=1})$$

$$\mathbf{w}^{t=3} = \mathbf{w}^{t=2} - \alpha \nabla L_{s_2}(\mathbf{w}^{t=2})$$

...

$$\mathbf{w}^T = \mathbf{w}^{T-1} - \alpha \nabla L_{s_{T-1}}(\mathbf{w}^{T-1})$$

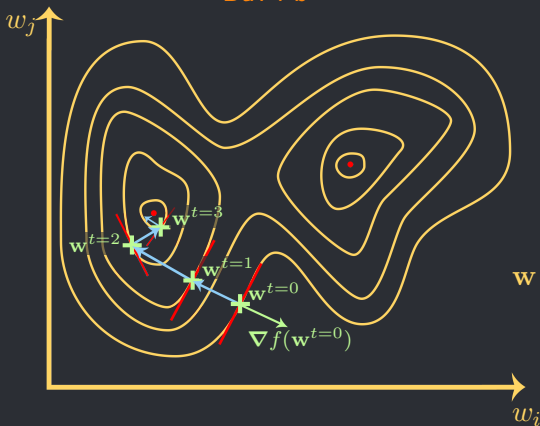
$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} \quad \nabla L = \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_n} \end{bmatrix}$$



## Батчи

$$L_b = L_1 + L_{10} + L_3 + \dots + L_8$$

Батч b



Градиент вычислен  
на одном батче

$$\mathbf{w}^{t=0}$$

$$\mathbf{w}^{t=1} = \mathbf{w}^{t=0} - \alpha \nabla L_{b_0}(\mathbf{w}^{t=0})$$

$$\mathbf{w}^{t=2} = \mathbf{w}^{t=1} - \alpha \nabla L_{b_1}(\mathbf{w}^{t=1})$$

$$\mathbf{w}^{t=3} = \mathbf{w}^{t=2} - \alpha \nabla L_{b_2}(\mathbf{w}^{t=2})$$

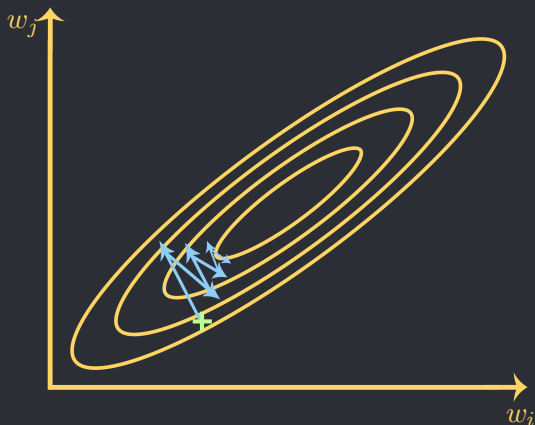
...

$$\mathbf{w}^T = \mathbf{w}^{T-1} - \alpha \nabla L_{b_{T-1}}(\mathbf{w}^{T-1})$$

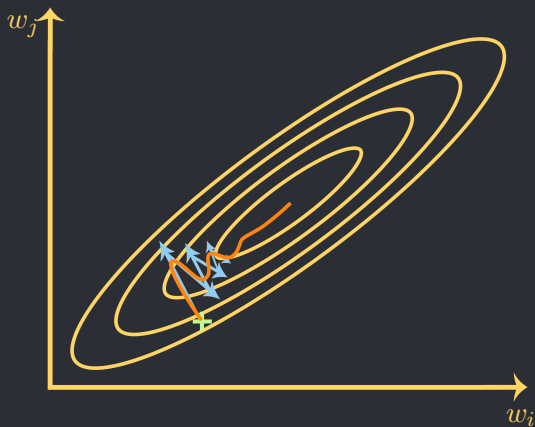
$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} \quad \nabla L = \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_n} \end{bmatrix}$$

# Хьюстон, у нас ... проблемы!

Очень медленно, нужно  
сделать много шагов чтобы  
сойтись к чему-то приличному

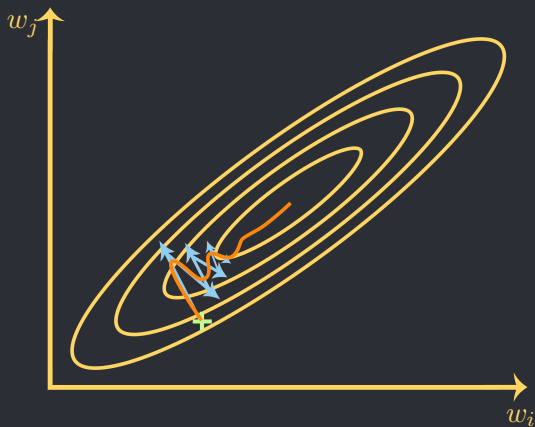


# Как катится шар



## Как катится шар

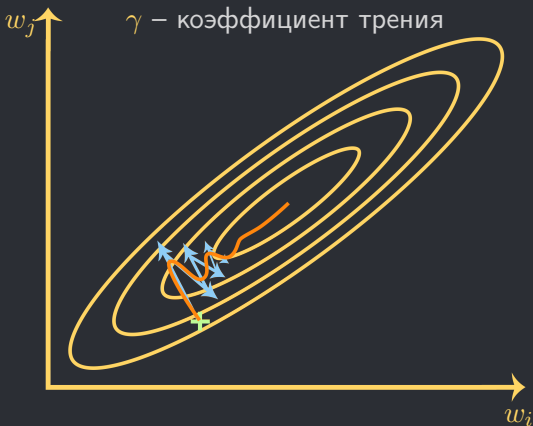
$$\begin{cases} \frac{\partial \mathbf{v}}{\partial t} = \frac{1}{m} (\mathbf{F} + \mathbf{F}_{\text{тр}}) \\ \frac{\partial \mathbf{x}}{\partial t} = \mathbf{v} \end{cases}$$



## Как катится шар

$$\begin{cases} \frac{\partial \mathbf{v}}{\partial t} = \frac{1}{m} (\mathbf{F} + \mathbf{F}_{\text{тр}}) = -\frac{1}{m} \nabla L - \frac{1}{m} \gamma \mathbf{v} \\ \frac{\partial \mathbf{x}}{\partial t} = \mathbf{v} \end{cases}$$

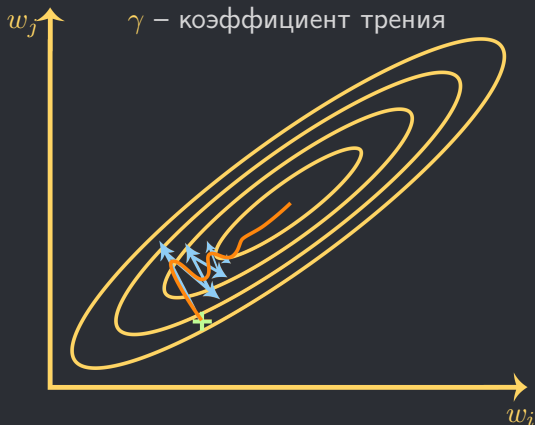
$\gamma$  – коэффициент трения



## Как катится шар

$$\begin{cases} \frac{\partial \mathbf{v}}{\partial t} = \frac{1}{m} (\mathbf{F} + \mathbf{F}_{\text{тр}}) = -\frac{1}{m} \nabla L - \frac{1}{m} \gamma \mathbf{v} \\ \frac{\partial \mathbf{x}}{\partial t} = \mathbf{v} \end{cases}$$

$\gamma$  – коэффициент трения

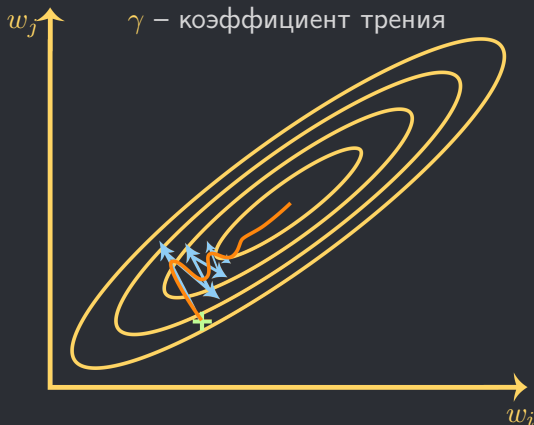


$$\begin{aligned} \frac{\partial \mathbf{v}}{\partial t} &= \frac{\mathbf{v}^{t+1} + \mathbf{v}^t}{\Delta t} \\ \frac{\partial \mathbf{x}}{\partial t} &= \frac{\mathbf{x}^{t+1} + \mathbf{x}^t}{\Delta t} \end{aligned}$$

## Как катится шар

$$\begin{cases} \frac{\partial \mathbf{v}}{\partial t} = \frac{1}{m} (\mathbf{F} + \mathbf{F}_{\text{тр}}) = -\frac{1}{m} \nabla L - \frac{1}{m} \gamma \mathbf{v} \\ \frac{\partial \mathbf{x}}{\partial t} = \mathbf{v} \end{cases}$$

$\gamma$  – коэффициент трения



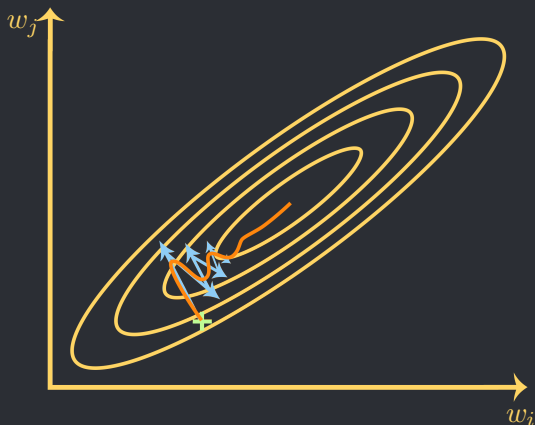
$$\begin{aligned} \frac{\partial \mathbf{v}}{\partial t} &= \frac{\mathbf{v}^{t+1} - \mathbf{v}^t}{\Delta t} \\ \frac{\partial \mathbf{x}}{\partial t} &= \frac{\mathbf{x}^{t+1} - \mathbf{x}^t}{\Delta t} \end{aligned}$$

$$\begin{cases} \mathbf{v}^{t+1} = -\alpha \nabla L(\mathbf{w}^t) - \beta \mathbf{v}^t \\ \mathbf{w}^{t+1} = \mathbf{w}^t + \mathbf{v}^t \end{cases}$$

## Заглядывание вперед

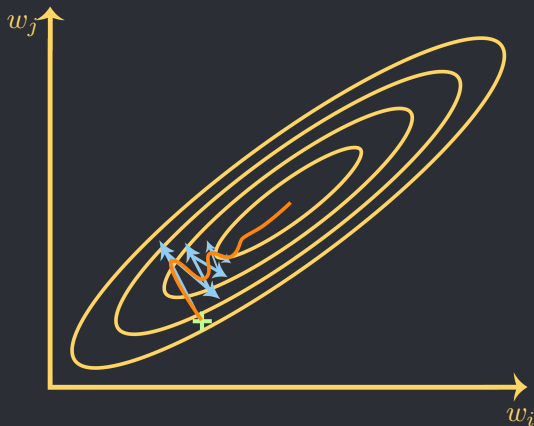
$$\begin{cases} \frac{\partial \mathbf{v}}{\partial t} = \frac{1}{m} (\mathbf{F} + \mathbf{F}_{\text{TP}}) = -\frac{1}{m} \nabla L - \frac{1}{m} \gamma \mathbf{v} \\ \frac{\partial \mathbf{x}}{\partial t} = \mathbf{v} \end{cases}$$

$$\begin{cases} \mathbf{v}^{t+1} = -\alpha \nabla L(\mathbf{w}^t + \mathbf{v}^t) - \beta \mathbf{v}^t \\ \mathbf{w}^{t+1} = \mathbf{w}^t + \mathbf{v}^{t+1} \end{cases}$$





## Похожий вариант



$$\mathbf{w}^{t+1} = \mathbf{w}^t - \alpha EMA_{\beta}^t(\nabla L)$$

$$EMA_{\beta}^t(\nabla L) = (1 - \beta) \nabla L^t + \beta EMA_{\beta}^{t-1}(\nabla L)$$

# RProp

- Ранее у нас были одинаковые learning rate для всех параметров
- А что если сделать индивидуальные?
- Будем учитывать только знаки градиента  $\text{sign}(\nabla L)$
- Для начала выберем некоторый learning rate  $\alpha^{t=0}$  для всех весов

## RProp

- Ранее у нас были одинаковые learning rate для всех параметров
- А что если сделать индивидуальные?
- Будем учитывать только знаки градиента  $sign(\nabla L)$
- Для начала выберем некоторый learning rate  $\alpha^{t=0}$  для всех весов

$$\mathbf{w}_i^t = \mathbf{w}_i^{t-1} - \alpha^t \cdot sign(\nabla_i L(\mathbf{w}^{t-1}))$$

$i$  — индекс веса

## RProp

- Ранее у нас были одинаковые learning rate для всех параметров
- А что если сделать индивидуальные?
- Будем учитывать только знаки градиента  $\text{sign}(\nabla L)$
- Для начала выберем некоторый learning rate  $\alpha^{t=0}$  для всех весов

$$\mathbf{w}_i^t = \mathbf{w}_i^{t-1} - \alpha^t \cdot \text{sign}(\nabla_i L(\mathbf{w}^{t-1}))$$

$i$  – индекс веса

$$\alpha_i^{t+1} = \begin{cases} \alpha_i^t \cdot 1.2 & \text{if } \text{sign}(\nabla_i L(\mathbf{w}^t) \cdot \nabla_i L(\mathbf{w}^{t-1})) > 0 \\ \alpha_i^t \cdot 0.5 & \text{if } \text{sign}(\nabla_i L(\mathbf{w}^t) \cdot \nabla_i L(\mathbf{w}^{t-1})) < 0 \end{cases}$$

## RProp

- Ранее у нас были одинаковые learning rate для всех параметров
- А что если сделать индивидуальные?
- Будем учитывать только знаки градиента  $\text{sign}(\nabla L)$
- Для начала выберем некоторый learning rate  $\alpha^{t=0}$  для всех весов

$$\mathbf{w}_i^t = \mathbf{w}_i^{t-1} - \alpha^t \cdot \text{sign}(\nabla_i L(\mathbf{w}^{t-1}))$$

$i$  – индекс веса

$$\alpha_i^{t+1} = \begin{cases} \alpha_i^t \cdot 1.2 & \text{if } \text{sign}(\nabla_i L(\mathbf{w}^t) \cdot \nabla_i L(\mathbf{w}^{t-1})) > 0 \\ \alpha_i^t \cdot 0.6 & \text{if } \text{sign}(\nabla_i L(\mathbf{w}^t) \cdot \nabla_i L(\mathbf{w}^{t-1})) \leq 0 \end{cases}$$

## RMSprop

- RProp плохо батчеризуется, и при этом весь какой-то эмпирический
- RMSprop похож на RProp, но выглядит адекватнее

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \alpha \frac{\nabla L}{\|\nabla L\|}$$

## RMSprop

- RProp плохо батчеризуется, и при этом весь какой-то эмпирический
- RMSprop похож на RProp, но выглядит адекватнее

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \alpha \frac{\nabla L(\mathbf{w}^t)}{\sqrt{EMA_{\beta}^t(\nabla L)^2}}$$

## RMSprop

- RProp плохо батчеризуется, и при этом весь какой-то эмпирический
- RMSprop похож на RProp, но выглядит адекватнее

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \alpha \frac{\nabla L(\mathbf{w}^t)}{\sqrt{EMA_{\beta}^t(\nabla L)^2}}$$

$$EMA_{\beta}^t(\nabla L)^2 = (1 - \beta) \nabla L(\mathbf{w}^t)^2 + \beta \cdot EMA_{\beta}^{t-1}(\nabla L)^2$$



## RMSprop

- RProp плохо батчеризуется, и при этом весь какой-то эмпирический
- RMSprop похож на RProp, но выглядит адекватнее

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \alpha \frac{\nabla L(\mathbf{w}^t)}{\sqrt{EMA_{\beta}^t(\nabla L)^2}}$$

$$EMA_{\beta}^t(\nabla L)^2 = (1 - \beta) \nabla L(\mathbf{w}^t)^2 + \beta \cdot EMA_{\beta}^{t-1}(\nabla L)^2$$

$$\nabla L^2 = \underbrace{(\nabla L_0 + \nabla L_1 + \dots + \nabla L_n)}_{\text{градиенты примеров в батче}}^2$$

# Adam

- Вспомним о Шаре!

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \alpha \frac{EMA_{\beta_1}^t(\nabla L)}{\sqrt{EMA_{\beta_2}^t(\nabla L^2) + \varepsilon}}$$

# Adam

- Вспомним о Шаре!

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \alpha \frac{EMA_{\beta_1}^t(\nabla L)}{\sqrt{EMA_{\beta_2}^t(\nabla L^2)} + \varepsilon}$$

$$EMA_{\beta_1}^t(\nabla L) = \beta_1 \nabla L^t + (1 - \beta_1) EMA_{\beta_1}^{t-1}(\nabla L)$$

$$EMA_{\beta_2}^t(\nabla L^2) = \beta_2 (\nabla L^t)^2 + (1 - \beta_2) EMA_{\beta_2}^{t-1}(\nabla L^2)$$

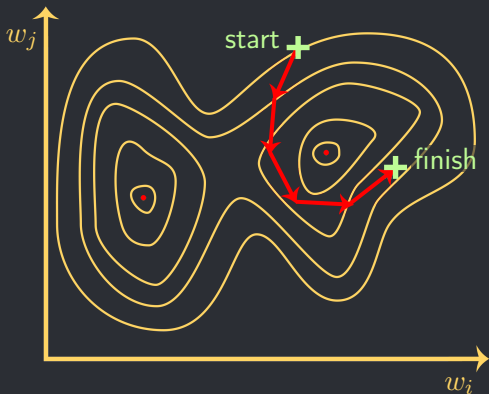
# Lookahead



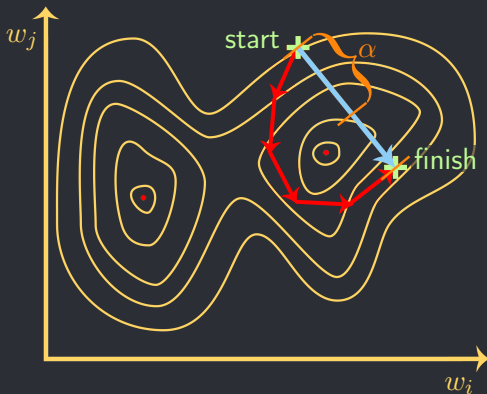
# Lookahead

Идея:

- 1) Сделать несколько шагов  $k$  некоторым алгоритмом  $A$



# Lookahead

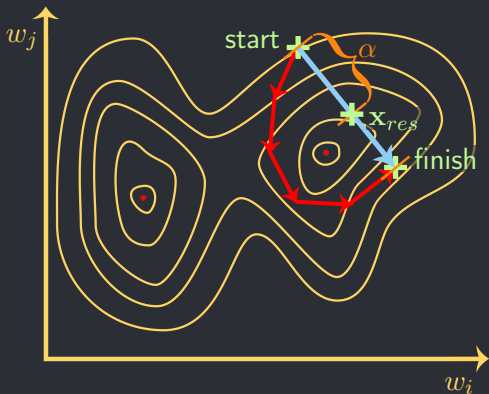


Идея:

- 1) Сделать несколько шагов  $k$  некоторым алгоритмом  $A$
- 2) А потом выберем некоторую точку между точкой старта и финиша:

$$\mathbf{x}_{res} = \alpha \mathbf{x}_{start} + (1 - \alpha) \mathbf{x}_{finish}$$

# Lookahead



Идея:

- 1) Сделать несколько шагов  $k$  некоторым алгоритмом  $A$
- 2) А потом выберем некоторую точку между точкой старта и финиша:

$$\mathbf{x}_{res} = \alpha \mathbf{x}_{start} + (1 - \alpha) \mathbf{x}_{finish}$$

- 3) И в следующий раз шаги будем начинать с точки  $\mathbf{x}_{res}$

# Lookahead

Параметр  $\alpha$  можно выбрать:

- 1) Фиксированным  $\alpha \in (0, 1)$
- 2) Каждый раз на шаге 2 подбирать  $\alpha$ , таким образом, чтобы минимизировать приближение функции:

$$\hat{L} = \alpha x^2 + bx + c,$$

где

$$\mathbf{x}_{res} = \alpha \mathbf{x}_{start} + (1 - \alpha) \mathbf{x}_{end},$$

$a$ ,  $b$  и  $c$  подобраны из условий:

$$\begin{aligned}\hat{L}(\alpha = 0) &= L(\alpha = 0) \\ \hat{L}(\alpha = 1) &= L(\alpha = 1) \\ \frac{\partial \hat{L}}{\partial \alpha}(\alpha = 1) &= \frac{\partial L}{\partial \alpha}(\alpha = 1)\end{aligned}$$



## ИТОГ

- Stochastic Gradient Descent

## ИТОГ

- Stochastic Gradient Descent
- Momentum

## ИТОГ

- Stochastic Gradient Descent
- Momentum
- RProp

## ИТОГ

- Stochastic Gradient Descent
- Momentum
- RProp
- RMSprop

## ИТОГ

- Stochastic Gradient Descent
- Momentum
- RProp
- RMSprop
- Adam

## ИТОГ

- Stochastic Gradient Descent
- Momentum
- RProp
- RMSprop
- Adam
- Lookahead