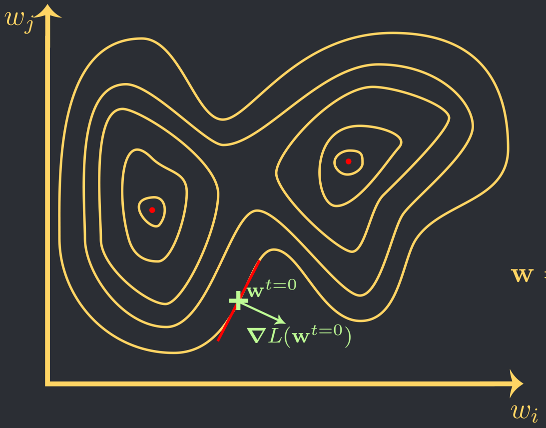
Часть 4: Методы оптимизации

Романов Михаил, Игорь Слинько

Градиентный спуск

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix}$$

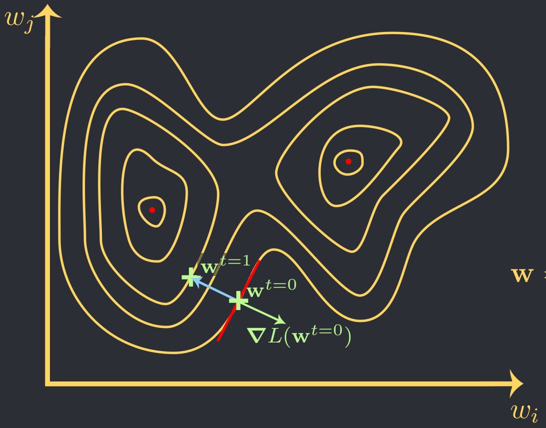
Градиентный спуск



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

Градиентный спуск

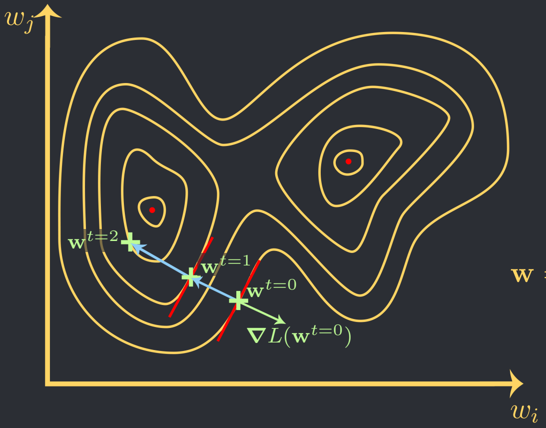
$$\mathbf{w}^{t=0}$$
$$\mathbf{w}^{t=1} = \mathbf{w}^{t=0} - \alpha \boldsymbol{\nabla} L(\mathbf{w}^{t=0})$$



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

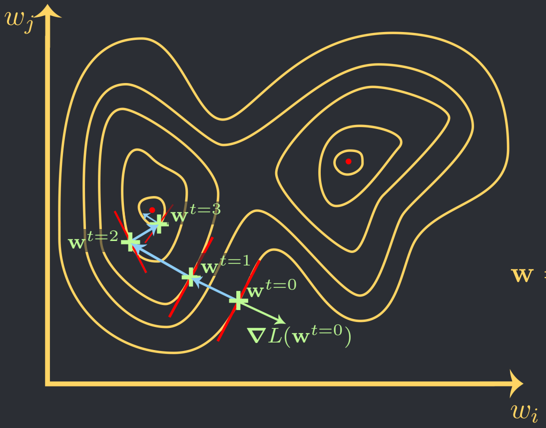
Градиентный спуск

$$\mathbf{w}^{t=0}$$
$$\mathbf{w}^{t=1} = \mathbf{w}^{t=0} - \alpha \nabla L(\mathbf{w}^{t=0})$$
$$\mathbf{w}^{t=2} = \mathbf{w}^{t=1} - \alpha \nabla L(\mathbf{w}^{t=1})$$



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

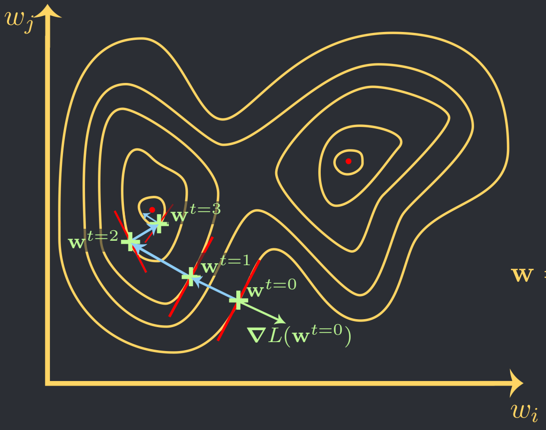
Градиентный спуск



$$\mathbf{w}^{t=0}$$
$$\mathbf{w}^{t=1} = \mathbf{w}^{t=0} - \alpha \nabla L(\mathbf{w}^{t=0})$$
$$\mathbf{w}^{t=2} = \mathbf{w}^{t=1} - \alpha \nabla L(\mathbf{w}^{t=1})$$
$$\mathbf{w}^{t=3} = \mathbf{w}^{t=2} - \alpha \nabla L(\mathbf{w}^{t=2})$$

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

Градиентный спуск



$$
\begin{aligned}
&\mathbf{w}^{t=0}\\
&\mathbf{w}^{t=1} = \mathbf{w}^{t=0} - \alpha \boldsymbol{\nabla} L(\mathbf{w}^{t=0})\\
&\mathbf{w}^{t=2} = \mathbf{w}^{t=1} - \alpha \boldsymbol{\nabla} L(\mathbf{w}^{t=1})\\
&\mathbf{w}^{t=3} = \mathbf{w}^{t=2} - \alpha \boldsymbol{\nabla} L(\mathbf{w}^{t=2})\\
&\dots\\
&\mathbf{w}^{T} = \mathbf{w}^{T-1} - \alpha \boldsymbol{\nabla} L(\mathbf{w}^{T-1})
\end{aligned}
$$

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

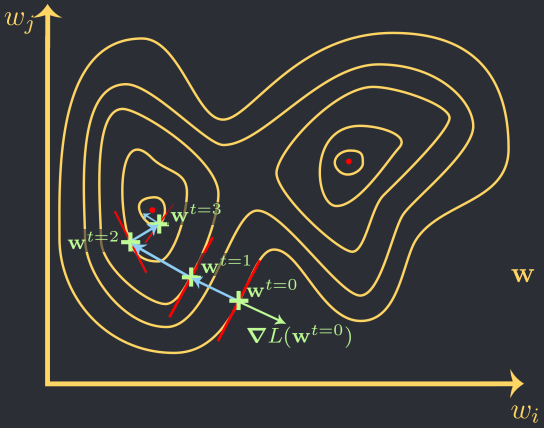
SGD

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

Лосс функция – сумма лоссов на нескольких примерах

Градиент вычислен на одном примере

$$\mathbf{w}^{t=0}$$
$$\mathbf{w}^{t=1} = \mathbf{w}^{t=0} - \alpha \boldsymbol{\nabla} L_{s_0}(\mathbf{w}^{t=0})$$
$$\mathbf{w}^{t=2} = \mathbf{w}^{t=1} - \alpha \boldsymbol{\nabla} L_{s_1}(\mathbf{w}^{t=1})$$
$$\mathbf{w}^{t=3} = \mathbf{w}^{t=2} - \alpha \boldsymbol{\nabla} L_{s_2}(\mathbf{w}^{t=2})$$
$$\ldots$$
$$\mathbf{w}^{T} = \mathbf{w}^{T-1} - \alpha \boldsymbol{\nabla} L_{s_{T-1}}(\mathbf{w}^{T-1})$$

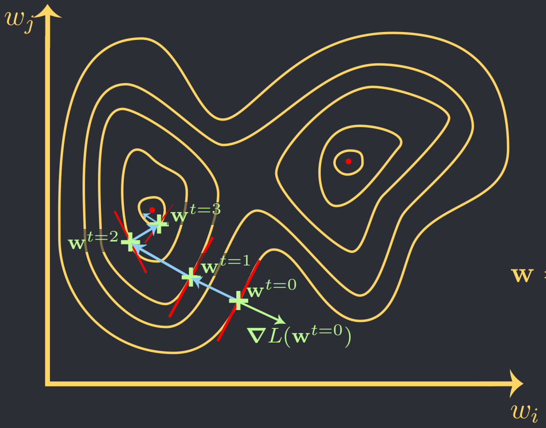


$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} \quad \boldsymbol{\nabla} L = \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_n} \end{bmatrix}$$

Батчи

$$L_b = L_1 + L_{10} + L_3 + ... + L_8$$

Батч b



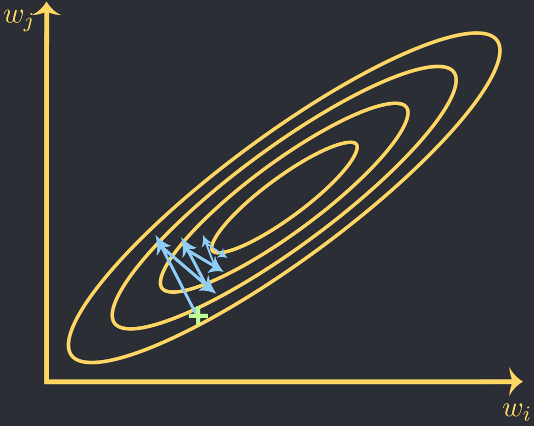
Градиент вычислен на одном батче

$$\begin{aligned}
&\mathbf{w}^{t=0}\\
&\mathbf{w}^{t=1} = \mathbf{w}^{t=0} - \alpha \nabla L_{b_0}(\mathbf{w}^{t=0})\\
&\mathbf{w}^{t=2} = \mathbf{w}^{t=1} - \alpha \nabla L_{b_1}(\mathbf{w}^{t=1})\\
&\mathbf{w}^{t=3} = \mathbf{w}^{t=2} - \alpha \nabla L_{b_2}(\mathbf{w}^{t=2})\\
&\dots\\
&\mathbf{w}^{T} = \mathbf{w}^{T-1} - \alpha \nabla L_{b_{T-1}}(\mathbf{w}^{T-1})
\end{aligned}$$

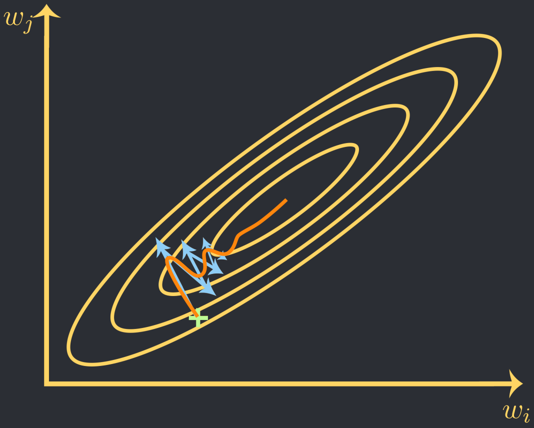
$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} \quad \nabla L = \begin{bmatrix} \frac{\partial L}{\partial w_0} \\ \frac{\partial L}{\partial w_1} \\ \vdots \\ \frac{\partial L}{\partial w_n} \end{bmatrix}$$

Хьюстон, у нас ... проблемы!

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

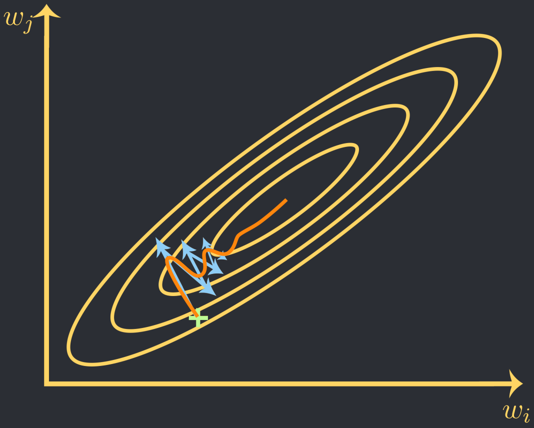


Как катится шар



Как катится шар

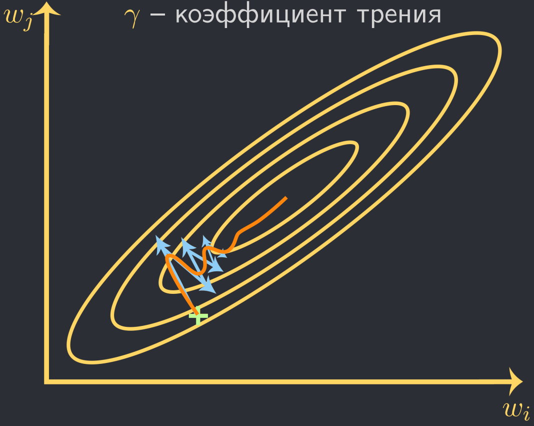
$$\begin{cases} \dfrac{\partial \mathbf{v}}{\partial t} = \dfrac{1}{m}(\mathbf{F} + \mathbf{F}_{\text{тр}}) \\ \dfrac{\partial \mathbf{x}}{\partial t} = \mathbf{v} \end{cases}$$



Как катится шар

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

γ – коэффициент трения



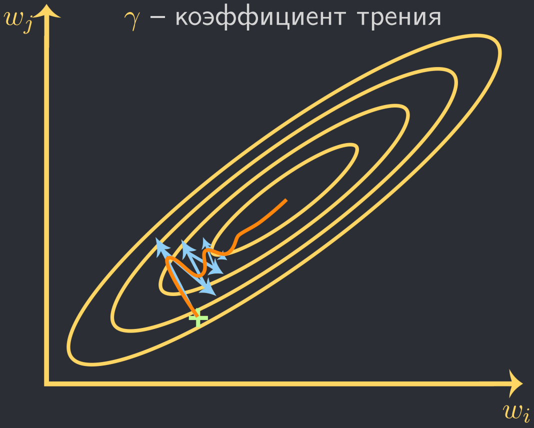
Как катится шар

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

$$\frac{\partial \mathbf{v}}{\partial t} = \frac{\mathbf{v}^{t+1} - \mathbf{v}^{t}}{\Delta t}$$

$$\frac{\partial \mathbf{x}}{\partial t} = \frac{\mathbf{x}^{t+1} - \mathbf{x}^{t}}{\Delta t}$$

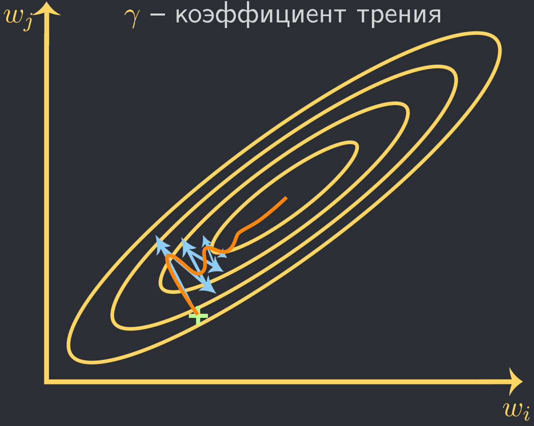
γ – коэффициент трения



Как катится шар

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

γ – коэффициент трения



$$\frac{\partial \mathbf{v}}{\partial t} = \frac{\mathbf{v}^{t+1} - \mathbf{v}^t}{\Delta t}$$

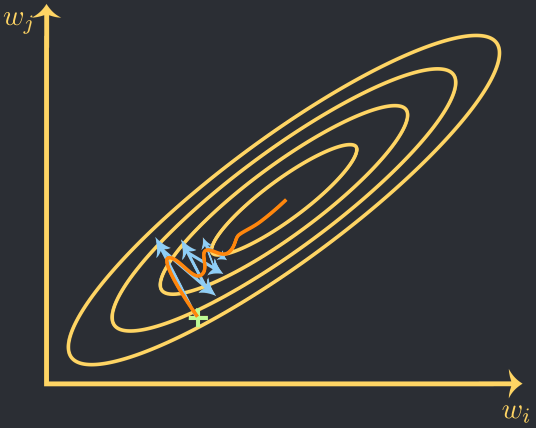
$$\frac{\partial \mathbf{x}}{\partial t} = \frac{\mathbf{x}^{t+1} - \mathbf{x}^t}{\Delta t}$$

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

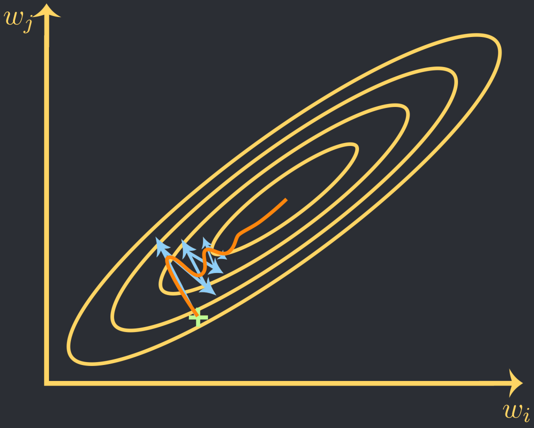
Заглядывание вперед

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

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



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 $sign(\nabla L)$
- 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

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

i – индекс веса

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

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

i – индекс веса

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

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

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

i – индекс веса

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

RMSprop

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

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

RMSprop

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

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

RMSprop

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

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

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

RMSprop

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

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

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

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

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!

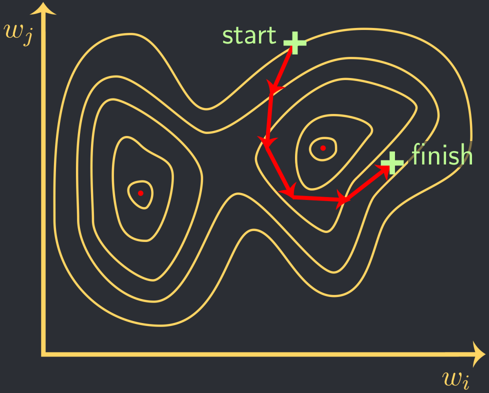
$$\mathbf{w}^{t+1} = \mathbf{w}^t - \alpha \frac{EMA^t_{\beta_1}(\boldsymbol{\nabla} L)}{\sqrt{EMA^t_{\beta_2}(\boldsymbol{\nabla} L^2)} + \varepsilon}$$

$$EMA^t_{\beta_1}(\boldsymbol{\nabla} L) = \beta_1 \boldsymbol{\nabla} L^t + (1 - \beta_1) EMA^{t-1}_{\beta_1}(\boldsymbol{\nabla} L)$$

$$EMA^t_{\beta_2}(\boldsymbol{\nabla} L^2) = \beta_2 \left(\boldsymbol{\nabla} L^t\right)^2 + (1 - \beta_2) EMA^{t-1}_{\beta_2}(\boldsymbol{\nabla} L^2)$$

Lookahead

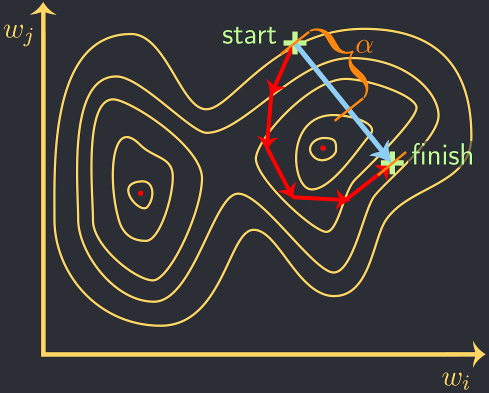
Lookahead



Идея:

1) Сделать несколько шагов k некоторым алгоритмом A

Lookahead



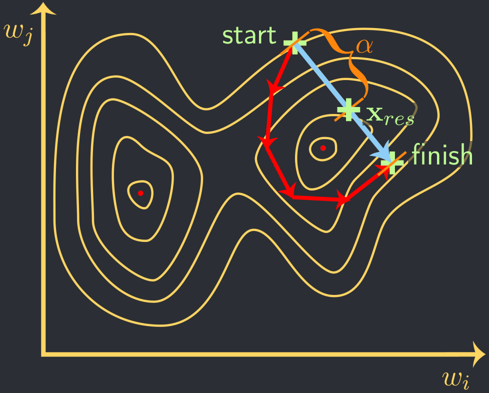
Идея:

1) Сделать несколько шагов k некоторым алгоритмом A

2) А потом выберем некоторую точку между точкой старта и финиша:

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

Lookahead



Идея:

1) Сделать несколько шагов k некоторым алгоритмом A

2) А потом выберем некоторую точку между точкой старта и финиша:

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

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

Lookahead

1) α можно выбрать фиксированным $\alpha \in (0, 1)$
2) Или каждый раз на шаге 2 подбиать α, таким образом, чтобы минимизировать приближение функции:

$$\hat{L}(\alpha) = a\alpha^2 + b\alpha + c,$$

в исходной функции $L(\mathbf{x})$ потерь считаем, что

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

получая зависимость $L(\alpha)$, a, b и c подобраны из условий:

$$\hat{L}\left(\alpha = 0\right) = L\left(\alpha = 0\right)$$
$$\hat{L}\left(\alpha = 1\right) = L\left(\alpha = 1\right)$$
$$\frac{\partial \hat{L}}{\partial \alpha}\left(\alpha = 1\right) = \frac{\partial L}{\partial \alpha}\left(\alpha = 1\right)$$

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