Нейронные сети в машинном обучении

Лекция №4
Методы оптимизации

Евгений Богатырев

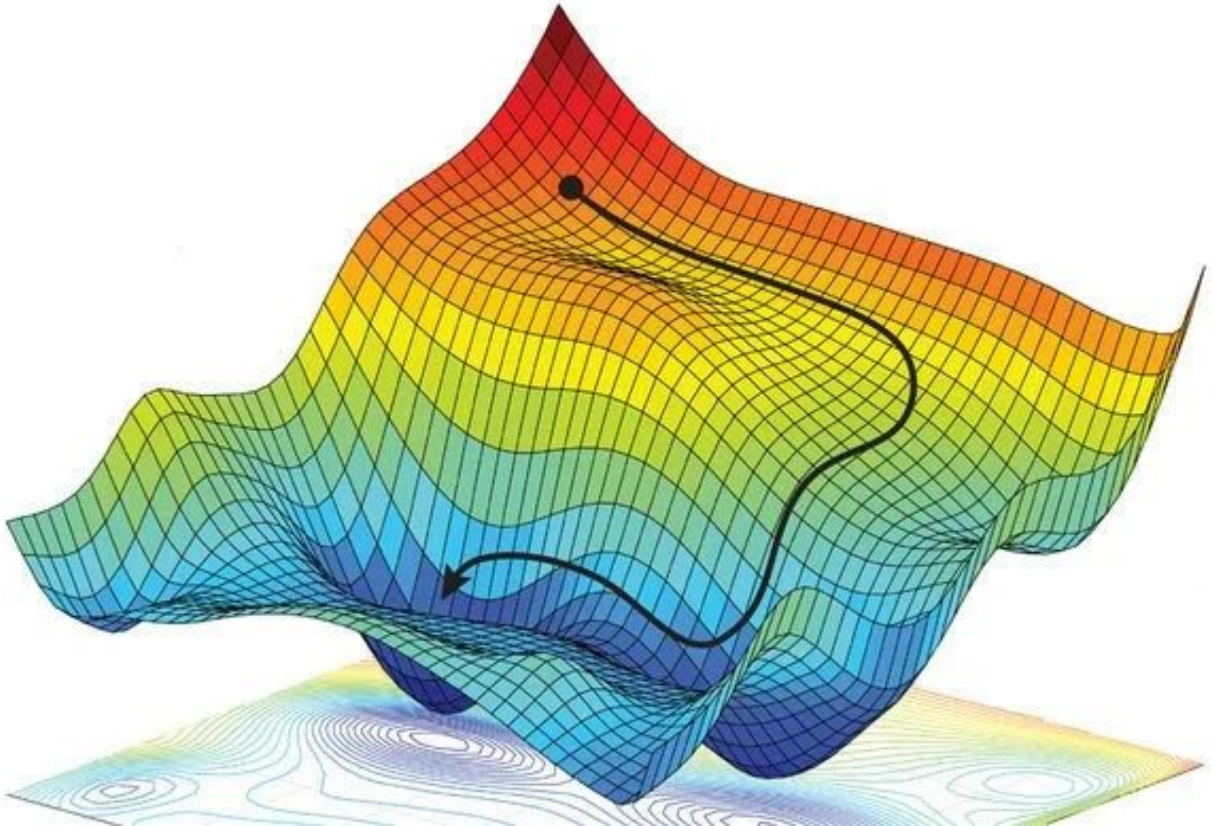
education

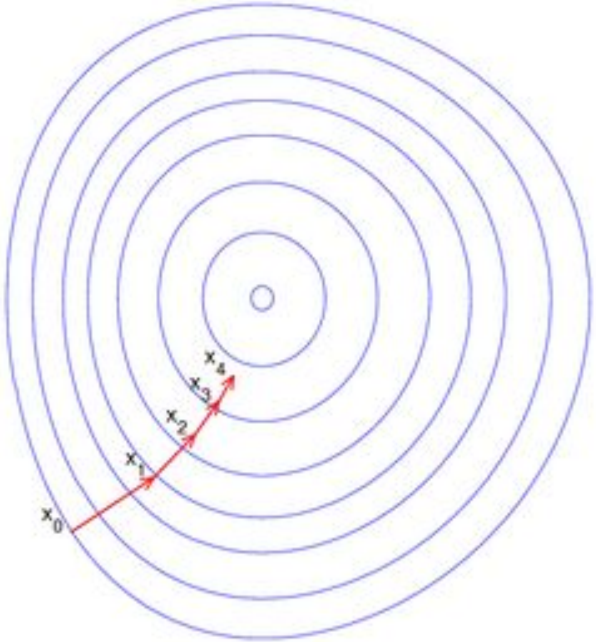
Не забывайте отмечаться
и оставлять отзывы!

Содержание

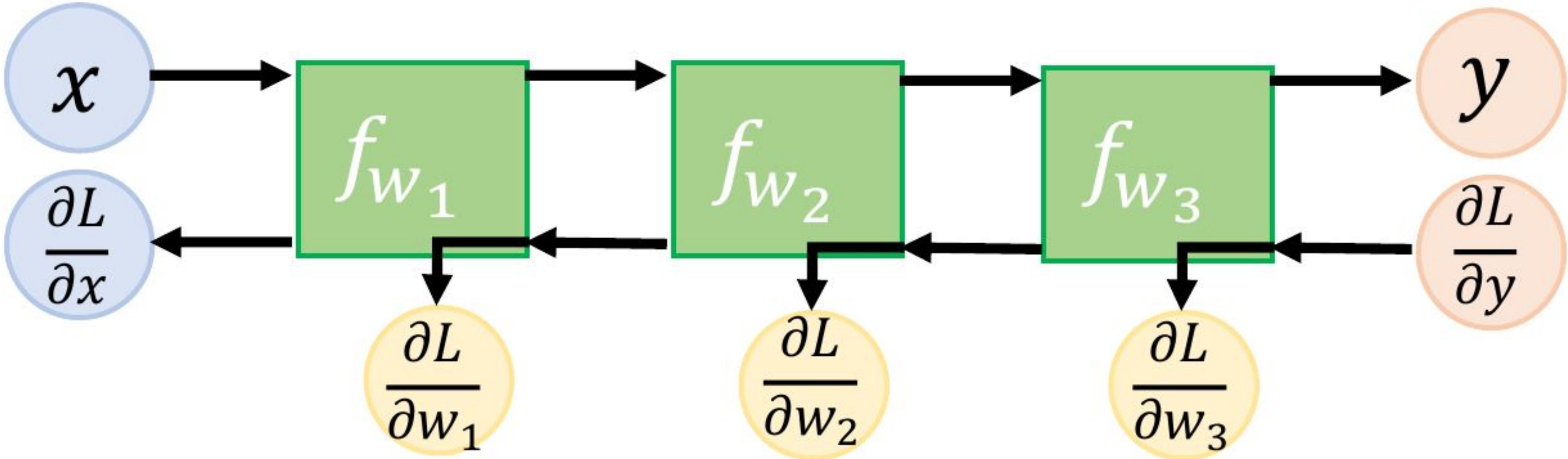
Постановка задачи

1. Надо найти оптимальные веса, которые минимизируют целевую функцию: $\theta_* = \min_\theta J(\theta)$
2. В любой точке можем вычислить градиент: $\nabla_\theta J(\theta)$





Back propagation

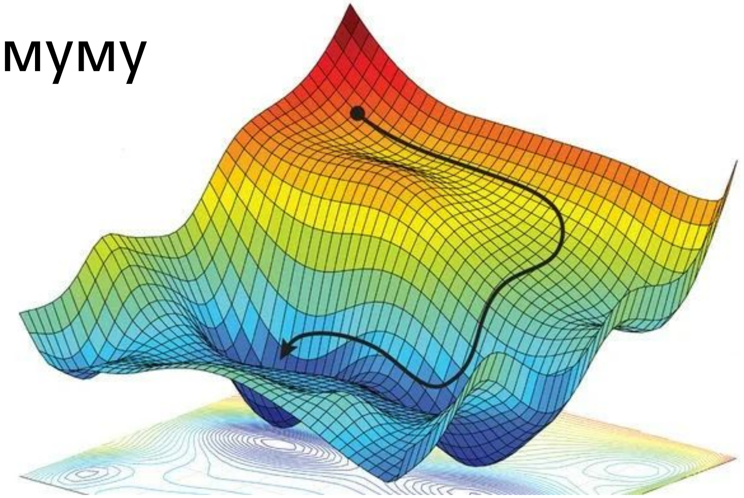


Gradient Descent

Градиентный спуск:

$$\theta_t = \theta_{t-1} - \eta \sum_{i=1}^{N} \nabla_\theta J_i(\theta_{t-1})$$

- Требуется обработать весь датасет для одного шага
- Нет режима online обучения
- Гарантируется сходимость к (локальному) минимуму при правильном выборе шага



education

Stochastic Gradient Descent (SGD)

Полная функция потерь:

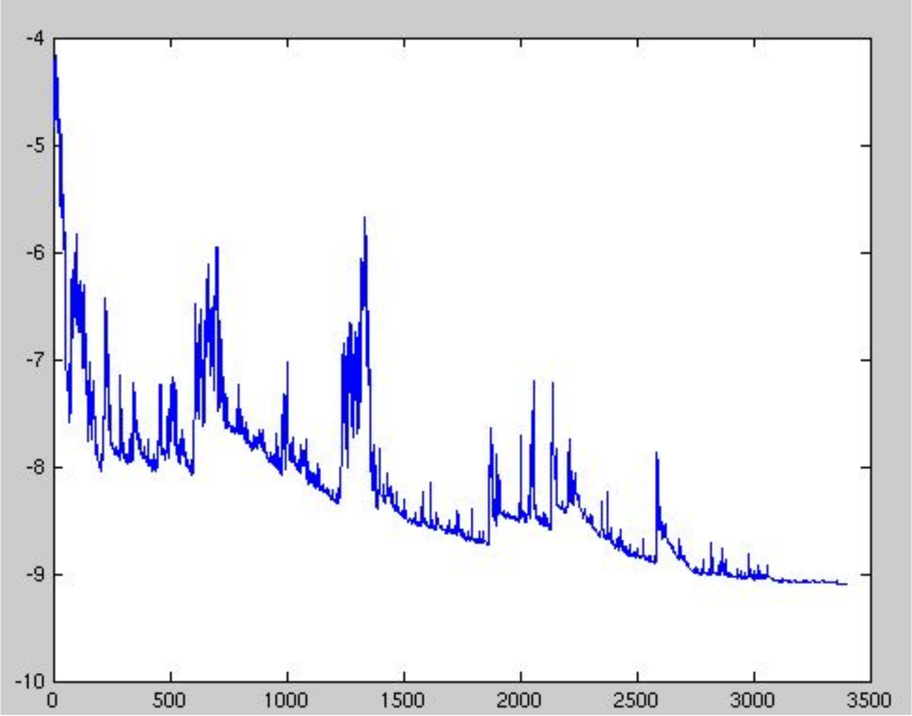
$$J(\theta) = \sum_{i=1}^{N} J_i(\theta)$$

Gradient Descent:

$$\theta_t = \theta_{t-1} - \eta \sum_{i=1}^{N} \nabla_\theta J_i(\theta_{t-1})$$

(Mini-Batch) Stochastic Gradient Descent:

$$\theta_t = \theta_{t-1} - \eta \sum_{i_1,\ldots,i_k} \nabla_\theta J_i(\theta_{t-1})$$

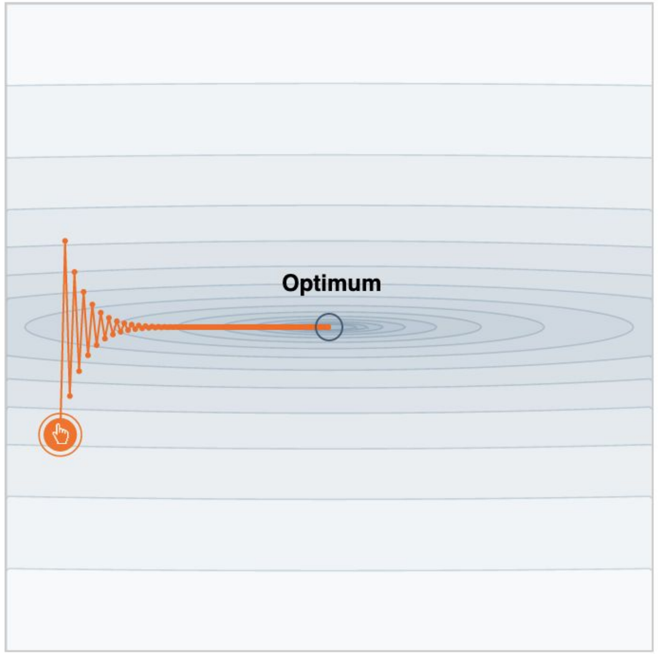


SGD с batch_size=1

https://upload.wikimedia.org/wikipedia/commons/f/f3/Stogra.png

education

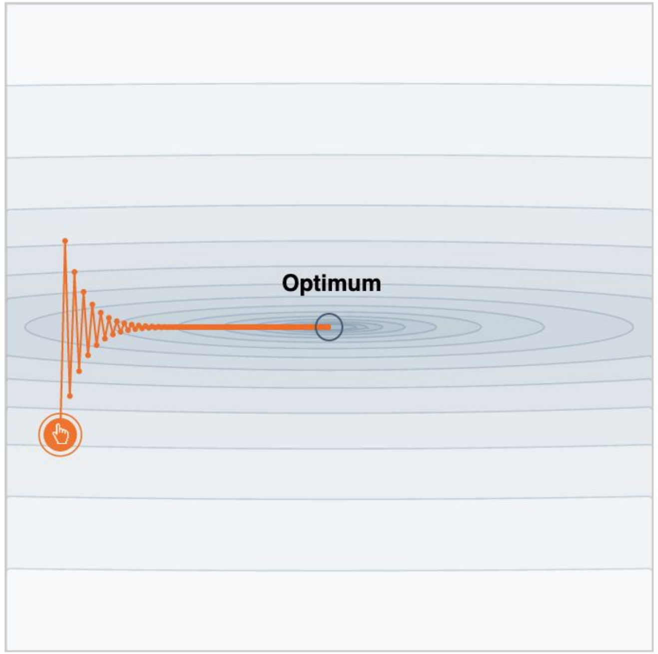
Проблемы SGD



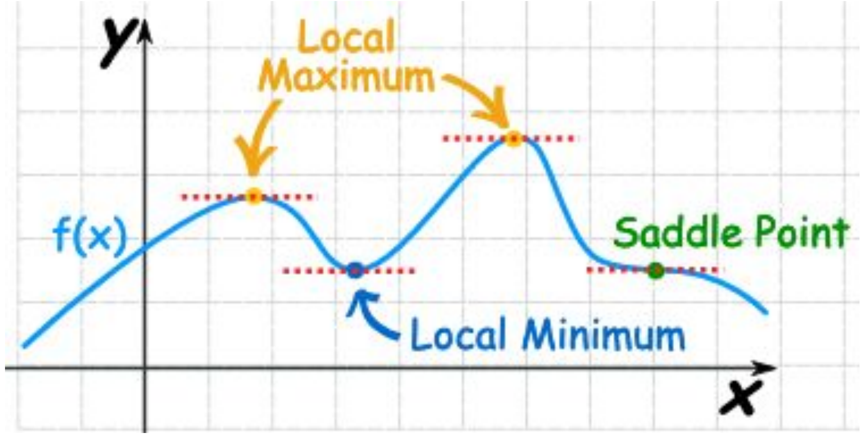
Разное поведение
у весов

https://distill.pub/2017/momentum/

Проблемы SGD



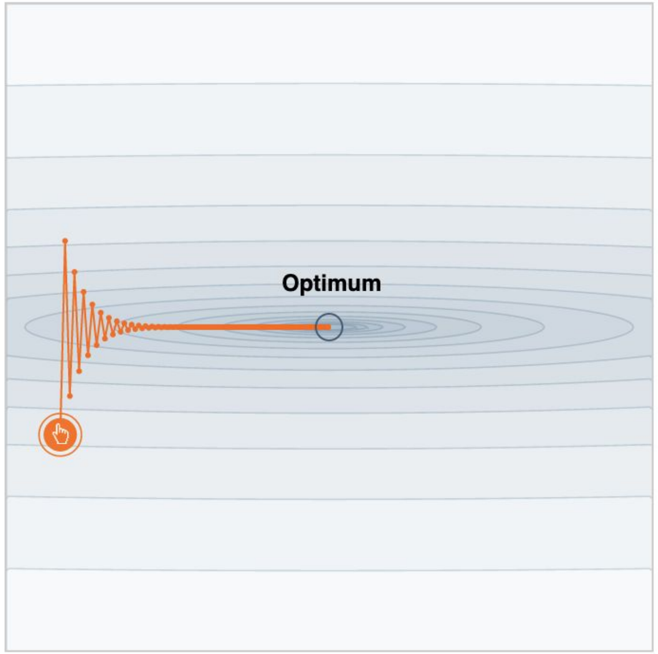
Разное поведение
у весов



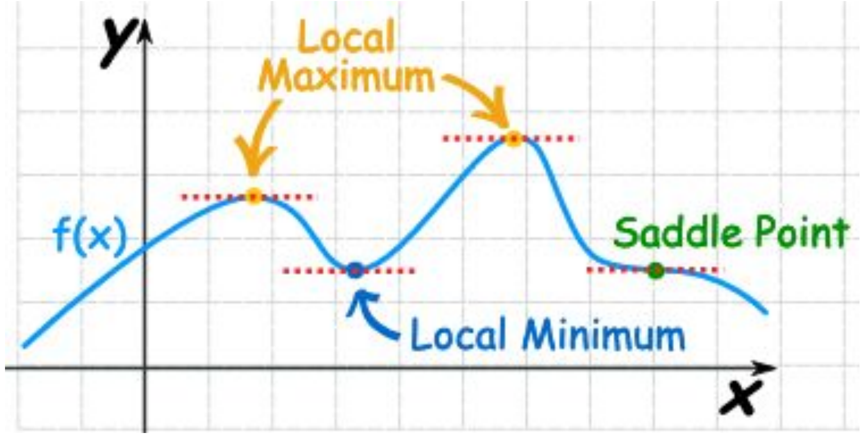
Локальные минимумы и
седловые точки

https://distill.pub/2017/momentum/
https://www.mathsisfun.com/calculus/maxima-minima.html

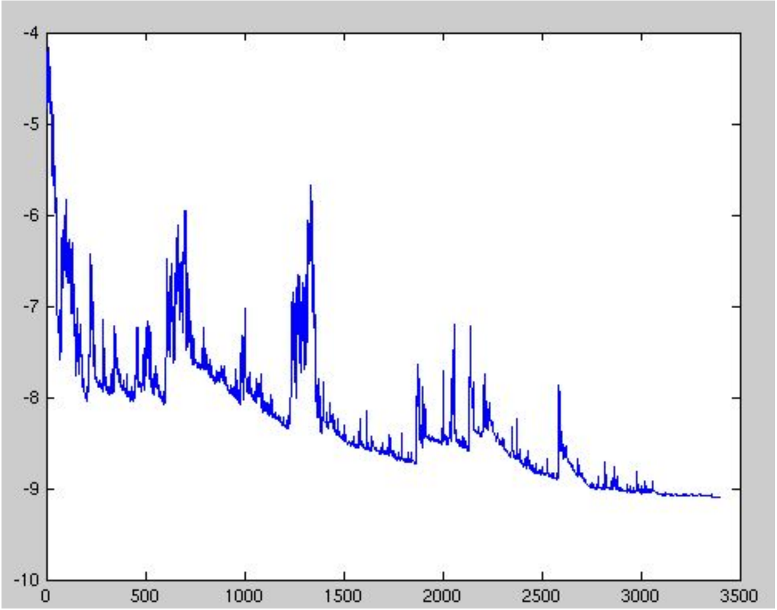
Проблемы SGD



Разное поведение
у весов



Локальные минимумы и
седловые точки



Шумная функция
потерь

https://distill.pub/2017/momentum/
https://www.mathsisfun.com/calculus/maxima-minima.html
https://upload.wikimedia.org/wikipedia/commons/f/f3/Stogra.png

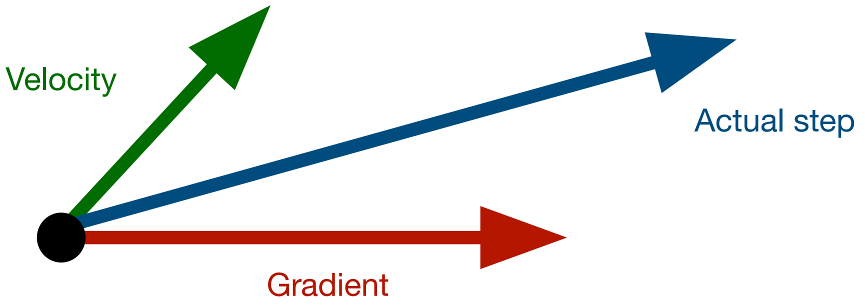
SGD + Momentum

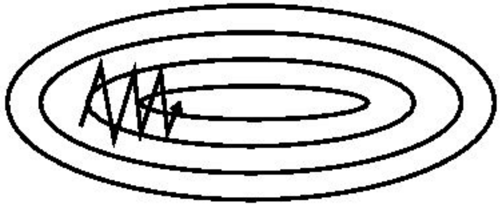
Попробуем сохранять «инерцию»:

$$\nu_t = \gamma \nu_{t-1} + \eta_t \nabla_\theta J(\theta_{t-1})$$

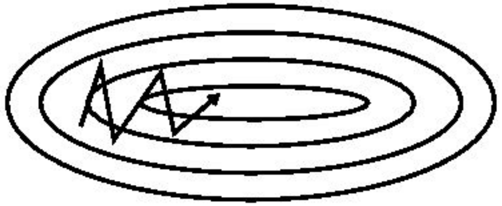
$$\theta_t = \theta_{t-1} - \nu_t$$

Рекомендовано брать: $\gamma = 0.9$





SGD



SGD with momentum

https://www.ruder.io/optimizing-gradient-descent/

SGD + Momentum

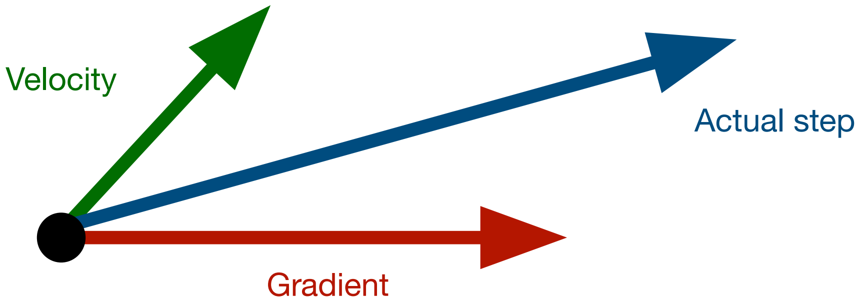
Попробуем сохранять «инерцию»:

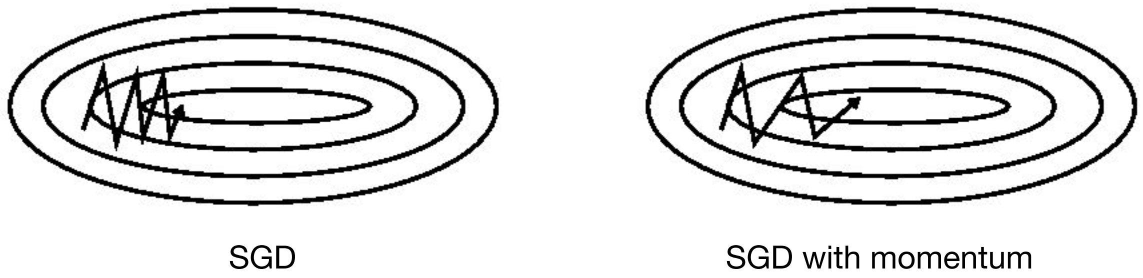
$$\nu_t = \gamma \nu_{t-1} + \eta_t \nabla_\theta J(\theta_{t-1})$$

$$\theta_t = \theta_{t-1} - \nu_t$$

Рекомендовано брать: $\gamma = 0.9$

Какая проблема этого метода?





https://www.ruder.io/optimizing-gradient-descent/

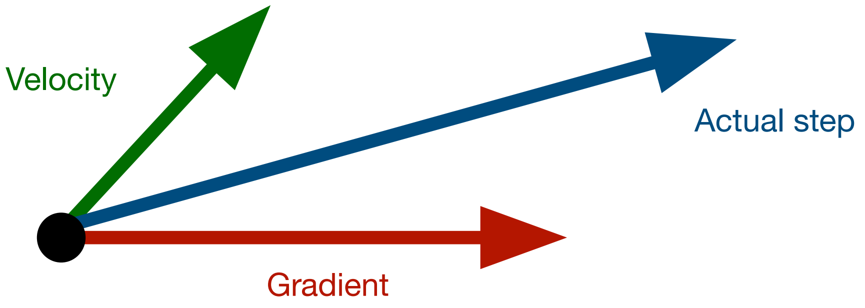
SGD + Momentum

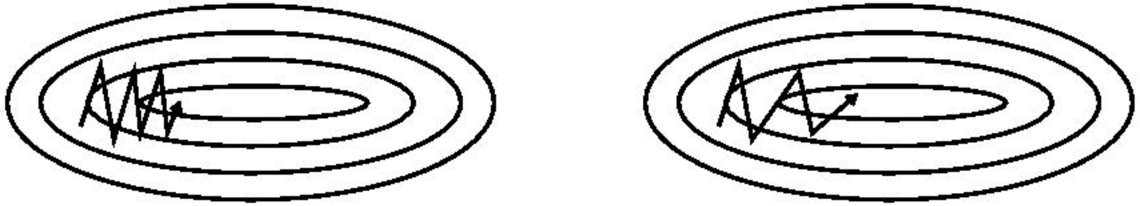
Попробуем сохранять «инерцию»:

$$\nu_t = \gamma \nu_{t-1} + \eta_t \nabla_\theta J(\theta_{t-1})$$

$$\theta_t = \theta_{t-1} - \nu_t$$

Рекомендовано брать: $\gamma = 0.9$





SGD

SGD with momentum

Какая проблема этого метода?

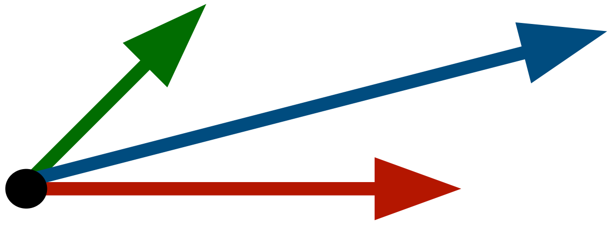
→ Может перескакивать локальные минимумы

https://www.ruder.io/optimizing-gradient-descent/

Nesterov accelerated gradient

Сдвигаемся на момент до вычисления градиента:

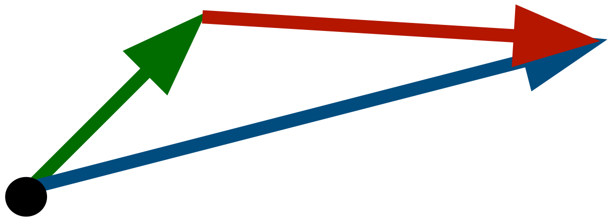
$$\nu_t = \gamma \nu_{t-1} + \eta_t \nabla_\theta J(\theta_{t-1} - \gamma \nu_{t-1})$$



simple momentum

Вычисление градиента в новой точке дает возможность скорректировать направление движения:

$$\theta_t = \theta_{t-1} - \nu_t$$



Nesterov momentum

Nesterov accelerated gradient

$$\nu_t = \gamma\nu_{t-1} + \eta_t \nabla_\theta J(\theta_{t-1} - \gamma\nu_{t-1})$$

$$\theta_t = \theta_{t-1} - \nu_t$$

Сделаем замену переменных так, чтобы не приходилось считать градиент в смещенной точке:

$$\theta_t^{prev} = \theta_t - \gamma\nu_t$$

$$\nu_t = \gamma\nu_{t-1} + \eta_t \nabla J(\theta_{t-1}^{prev})$$

Nesterov accelerated gradient

$$\theta_t = \theta_{t-1} - \nu_t$$

$$\theta_t^{prev} = \theta_t - \gamma\nu_t$$

$$\theta_t^{prev} + \gamma\nu_t = \theta_{t-1}^{prev} + \gamma\nu_{t-1} - \nu_t$$

Nesterov accelerated gradient

$$\theta_t = \theta_{t-1} - \nu_t \qquad\qquad \theta_t^{prev} = \theta_t - \gamma\nu_t$$

$$\theta_t^{prev} + \gamma\nu_t = \theta_{t-1}^{prev} + \gamma\nu_{t-1} - \nu_t$$

$$\theta_t^{prev} - \theta_{t-1}^{prev} = -\gamma\nu_t + \gamma\nu_{t-1} - \nu_t =$$

$$= -\nu_t(1+\gamma) + \gamma\nu_{t-1}$$

Nesterov accelerated gradient

$$\theta_t = \theta_{t-1} - \nu_t \qquad\qquad \theta_t^{prev} = \theta_t - \gamma\nu_t$$

$$\theta_t^{prev} + \gamma\nu_t = \theta_{t-1}^{prev} + \gamma\nu_{t-1} - \nu_t$$

$$\theta_t^{prev} - \theta_{t-1}^{prev} = -\gamma\nu_t + \gamma\nu_{t-1} - \nu_t =$$

$$= -\nu_t(1+\gamma) + \gamma\nu_{t-1}$$

$$\boxed{\theta_t^{prev} - = \nu_t(1+\gamma) - \gamma\nu_{t-1}}$$

AdaGrad

Градиент на шаге t:

$$g_t = \nabla_\theta J(\theta)$$

Сумма квадратов градиентов:

$$G_t = \sum_{k=0}^{t} g_t^2$$

Обновление весов:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \epsilon}} \cdot g_t$$

Стандартные значения:

$$\eta = 0.01,\ \epsilon = 10^{-8}$$

AdaGrad

Градиент на шаге t:

$$g_t = \nabla_\theta J(\theta)$$

Сумма квадратов градиентов:

$$G_t = \sum_{k=0}^{t} g_t^2$$

Обновление весов:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \epsilon}} \cdot g_t$$

Стандартные значения:

$$\eta = 0.01,\ \epsilon = 10^{-8}$$

Мотивация: уменьшаем learning rate для активных параметров

Какая проблема этого метода?

AdaGrad

Градиент на шаге t:

$$g_t = \nabla_\theta J(\theta)$$

Сумма квадратов градиентов:

$$G_t = \sum_{k=0}^{t} g_t^2$$

Обновление весов:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \epsilon}} \cdot g_t$$

Стандартные значения:

$$\eta = 0.01,\ \epsilon = 10^{-8}$$

Мотивация: уменьшаем learning rate для активных параметров

Какая проблема этого метода?

→ Не убывает => затухание обновлений

RMSProp

Будем использовать только последние несколько значений для подсчета

Экспоненциальное среднее:

$$G_t = \gamma G_{t-1} + (1-\gamma) g_t^2$$

Обновление весов:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \epsilon}} \cdot g_t$$

https://www.cs.toronto.edu/~hinton/coursera/lecture6/lec6.pdf

AdaDelta

RMSProp + избавимся от learning rate

$$\Delta\theta_t = \frac{RMS[\Delta\theta]_{t-1}}{RMS[g]_t} g_t$$

Обновление весов:

$$\theta_t = \theta_{t-1} - \Delta\theta_t$$

https://www.cs.toronto.edu/~hinton/coursera/lecture6/lec6.pdf

AdaDelta

Шаги вычисления AdaDelta:

$$\Delta\theta = -\frac{RMS[\Delta\theta]_{t-1}}{RMS[g]_t} g_t$$

$$\theta_{t+1} = \theta_t - \frac{RMS[\Delta\theta]_{t-1}}{RMS[g]_t} g_t$$

$$E[\Delta\theta^2]_t = \gamma E[\Delta\theta^2]_{t-1} + (1-\gamma)\Delta\theta_t^2$$

$$RMS[\Delta\theta]_t = \sqrt{E[\Delta\theta^2]_t + \epsilon}$$

https://www.cs.toronto.edu/~hinton/coursera/lecture6/lec6.pdf

AdaDelta

Шаги вычисления AdaDelta:

$$\Delta\theta = -\frac{RMS[\Delta\theta]_{t-1}}{RMS[g]_t} g_t$$

$$\theta_{t+1} = \theta_t - \frac{RMS[\Delta\theta]_{t-1}}{RMS[g]_t} g_t$$

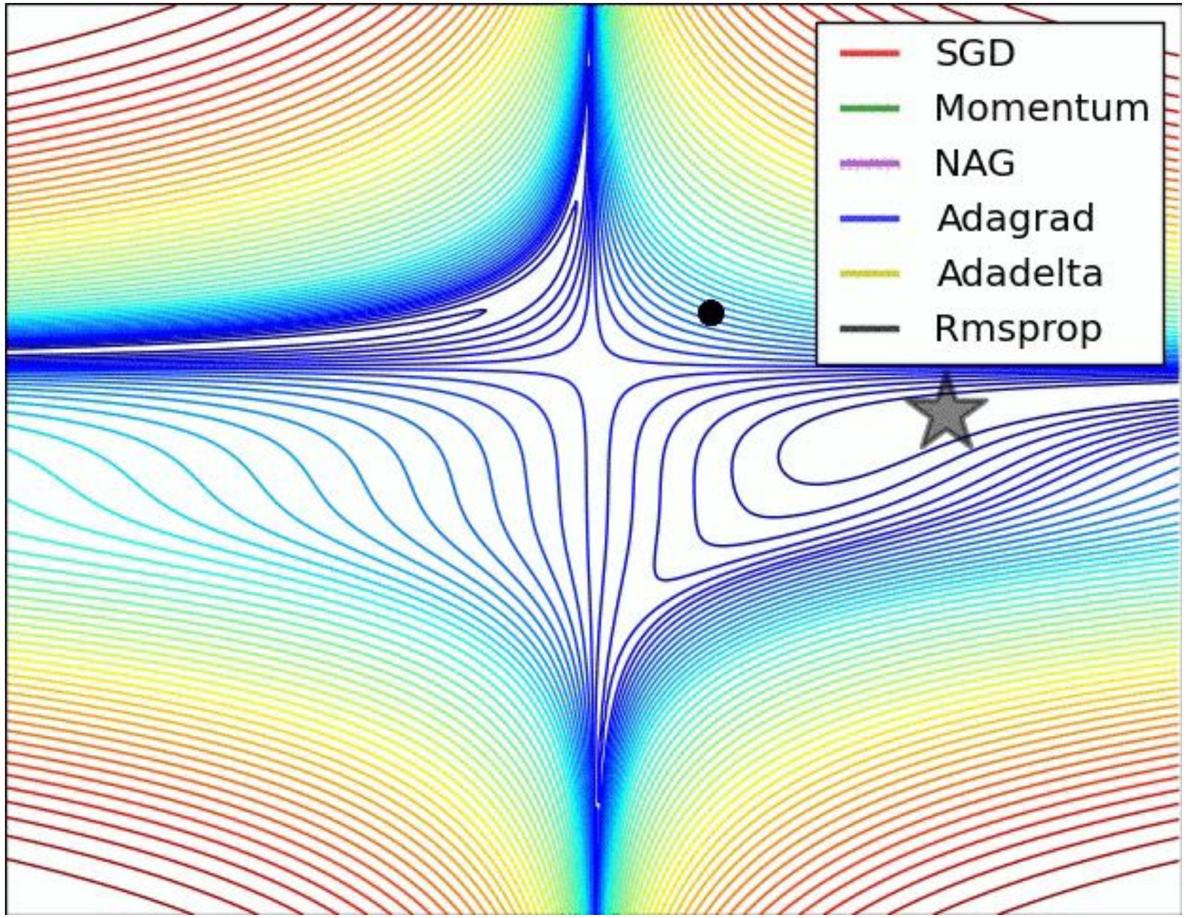
$$E[\Delta\theta^2]_t = \gamma E[\Delta\theta^2]_{t-1} + (1-\gamma)\Delta\theta_t^2$$

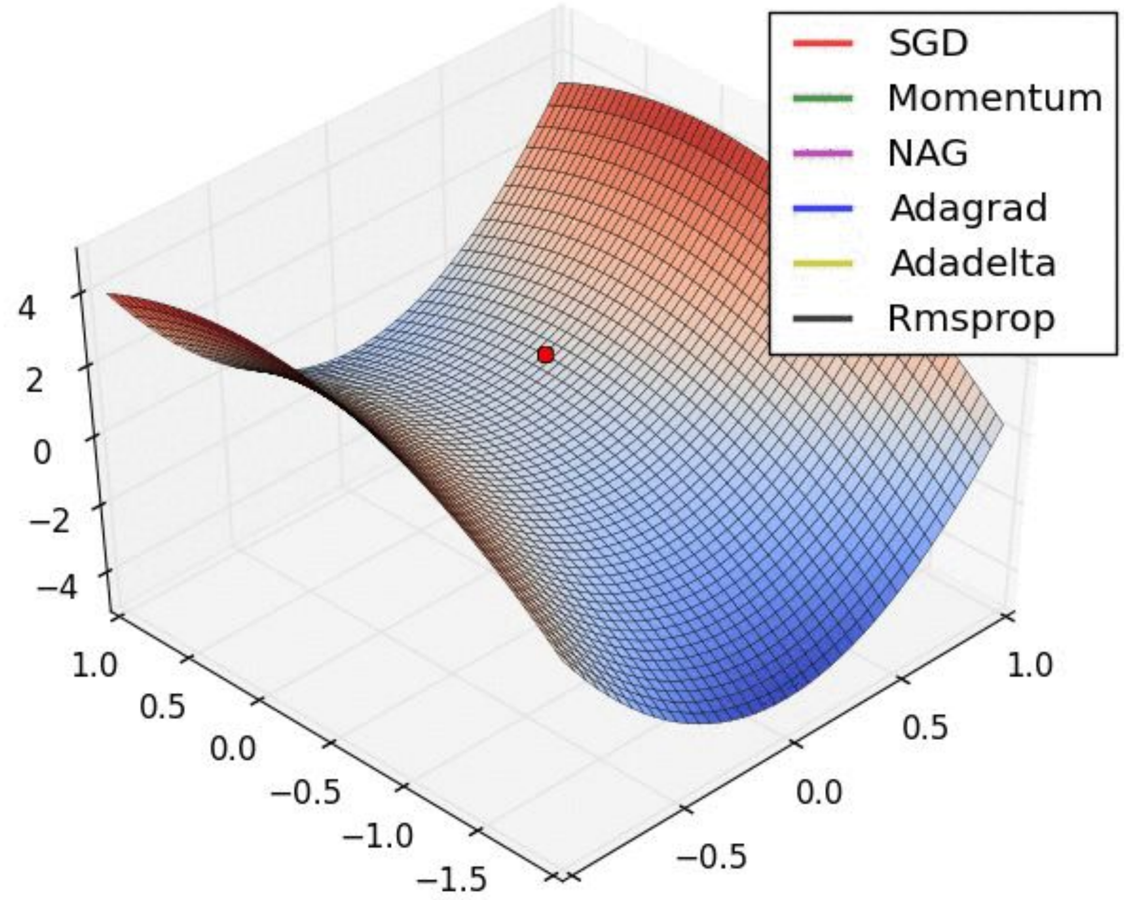
$$RMS[\Delta\theta]_t = \sqrt{E[\Delta\theta^2]_t + \epsilon}$$

$RMS[\Delta\theta]$ нужно инициализировать ненулевым значением!

https://www.cs.toronto.edu/~hinton/coursera/lecture6/lec6.pdf

Сравнение методов





Adaptive Moment Estimation (Adam)

Объединим лучшие идеи

Момент и затухание:

$$\begin{cases} m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \\ \nu_t = \beta_2 \nu_{t-1} + (1 - \beta_2) g_t^2 \end{cases}$$

Обновление весов:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{\nu}_t} + \epsilon} \hat{m}_t$$

education

Adaptive Moment Estimation (Adam)

Объединим лучшие идеи

Момент и затухание:

$$\begin{cases} m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \\ \nu_t = \beta_2 \nu_{t-1} + (1-\beta_2) g_t^2 \end{cases}$$

Обновление весов:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{\nu}_t} + \epsilon} \hat{m}_t$$

Инициализируем средние нулями, поэтому долгий “разгон”

Хотим несмещенность:

$$\mathbb{E}[m_t] = \mathbb{E}[g_t] \quad \mathbb{E}[v_t] = \mathbb{E}[g_t^2]$$

education

Adaptive Moment Estimation (Adam)

Объединим лучшие идеи

Момент и затухание:

$$\begin{cases} m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \\ \nu_t = \beta_2 \nu_{t-1} + (1-\beta_2) g_t^2 \end{cases}$$

Обновление весов:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{\nu}_t + \epsilon}} \hat{m}_t$$

Инициализируем средние нулями, поэтому долгий «разгон»

Хотим несмещенность:

$$\mathbb{E}[m_t] = \mathbb{E}[g_t] \; \mathbb{E}[v_t] = \mathbb{E}[g_t^2]$$

Делаем поправку

$$\begin{cases} \hat{m}_t = \frac{m_t}{1-\beta_1^t} \\ \hat{\nu}_t = \frac{\nu_t}{1-\beta_2^t} \end{cases}$$

Критерии остановки

Когда остановить обучение?

Критерии остановки

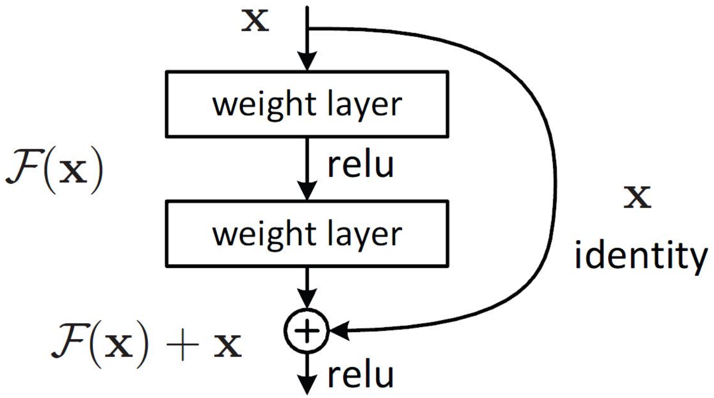
Когда остановить обучение?

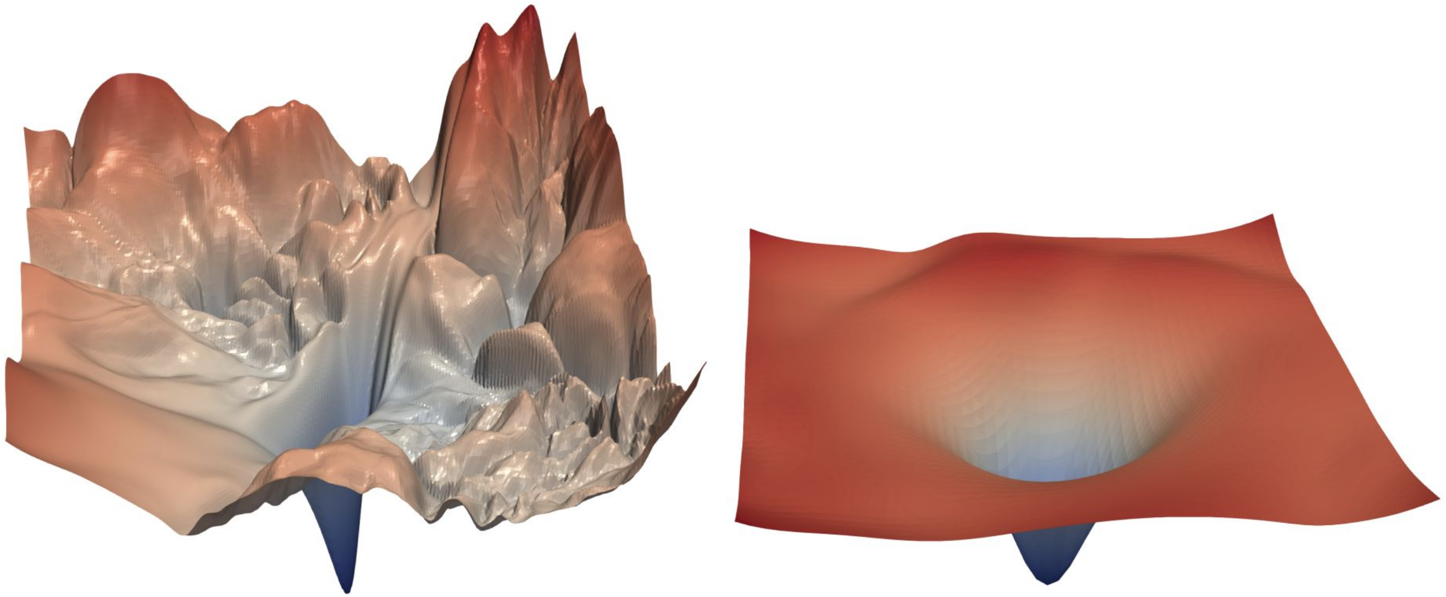
- Превышен лимит по числу итераций или времени
- Качество на валидации начало ухудшаться
- $J(\theta_t) - J(\theta_*) \leq \epsilon$
- $J(\theta_t) \leq \epsilon J(\theta_0)$
- $\|\nabla J(\theta_t)\| \leq \epsilon \|\nabla J(\theta_0)\|$

education

Ландшафт функции потерь

В следующей лекции…





(a) without skip connections

(b) with skip connections

Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The proposed filter normalization scheme is used to enable comparisons of sharpness/flatness between the two figures.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

https://arxiv.org/pdf/1712.09913.pdf,
https://www.youtube.com/watch?v=78vq6kgsTa8

Резюме

В этой лекции мы изучили:

1. В чем заключается задача оптимизации и что такое градиентный спуск
2. Как устроены различные методы оптимизации, в чем их достоинства и недостатки
3. Какие существуют критерии остановки обучения

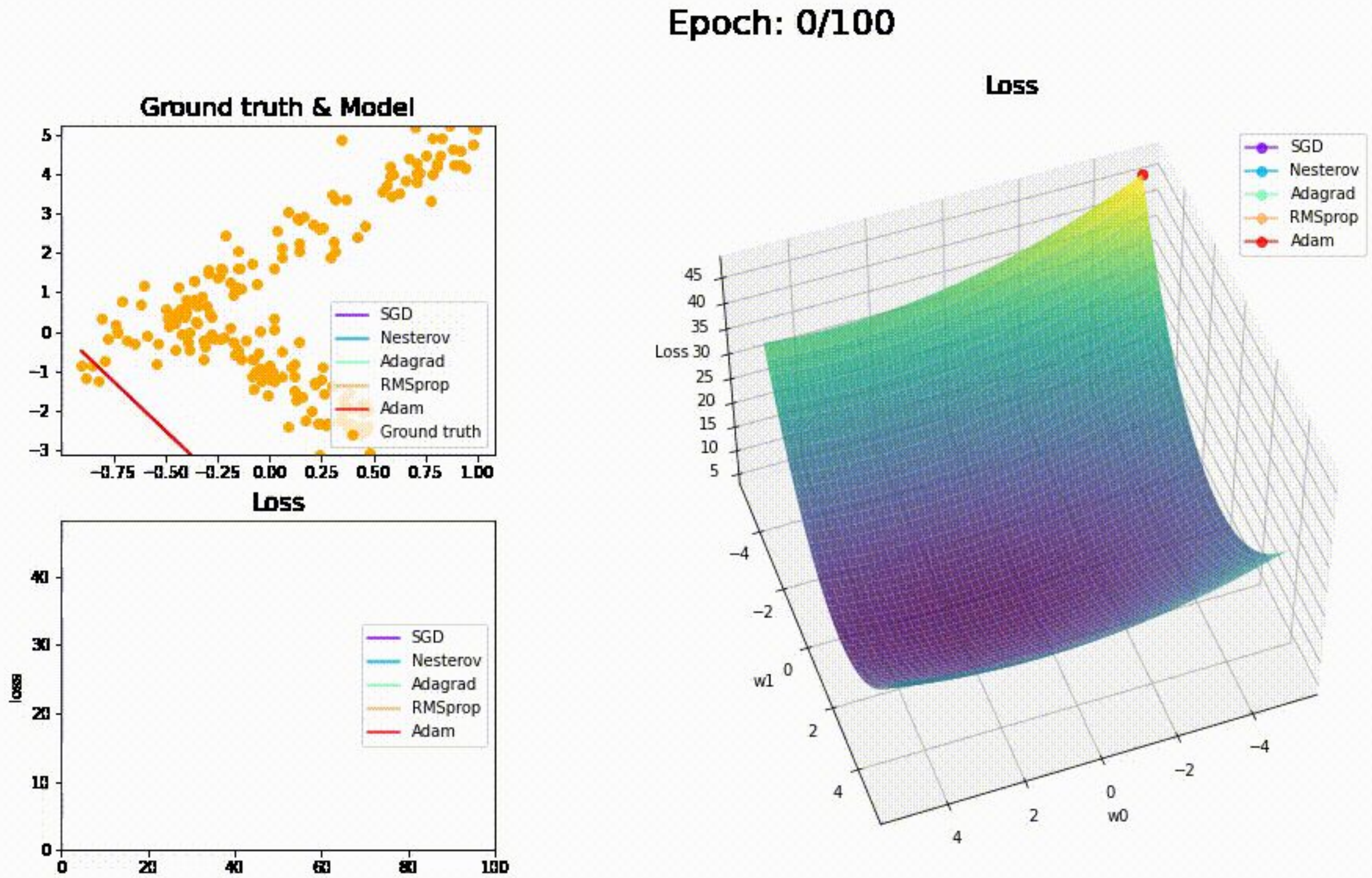
Рекомендованные источники

1. H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, “Visualizing the loss landscape of neural nets,” in *Advances in neural information processing systems*, 2018, p. 31. Available online
2. Why Momentum Really Works – https://distill.pub/2017/momentum/
3. An overview of gradient descent optimization algorithms – https://www.ruder.io/optimizing-gradient-descent/
4. Neural Networks for Machine Learning – https://www.cs.toronto.edu/~hinton/coursera/lecture6/lec6.pdf

Спасибо!

На семинаре будем сравнивать
различные методы оптимизации

VK education


Epoch: 0/100
Ground truth & Model
SGD
Nesterov
Adagrad
RMSprop
Adam
Ground truth
Loss
loss
Loss
Loss
w1
w0