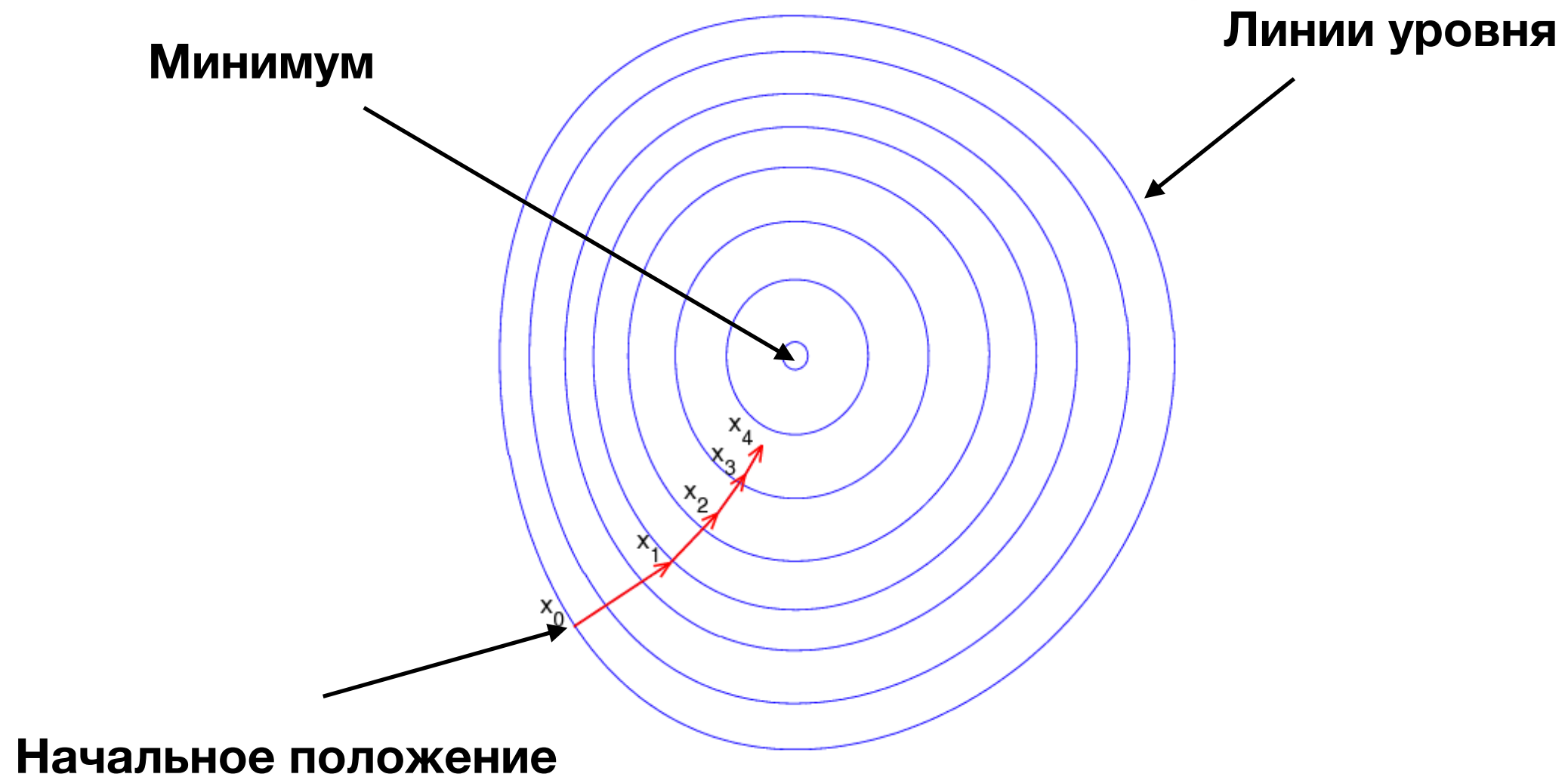


Градиентный спуск и стохастическая оптимизация

Студент 172 группы, Чернышев Вадим

Градиентный спуск. Алгоритм.



Градиентный спуск. Формула.

$$w^{(k)} = w^{(k-1)} - \eta_k \nabla Q(w^{(k-1)})$$

$Q(w^{(k)})$ - значение функционала ошибки для набора параметров

$\nabla Q(w^{(k)})$ - вектор, своим направлением указывающий направление наибольшего возрастания

η_k - длина шага, которая нужна для контроля скорости движения

Градиентный спуск. Градиент.

Градиент

$$\nabla f(x_1, \dots, x_d) = \left(\frac{\partial f}{\partial x_j} \right)_{j=1}^d$$

Формула подсчета градиента для вектора единичной длины

$$\frac{\partial f}{\partial v} = \langle \nabla f, v \rangle = ||\nabla f|| \cos \phi$$

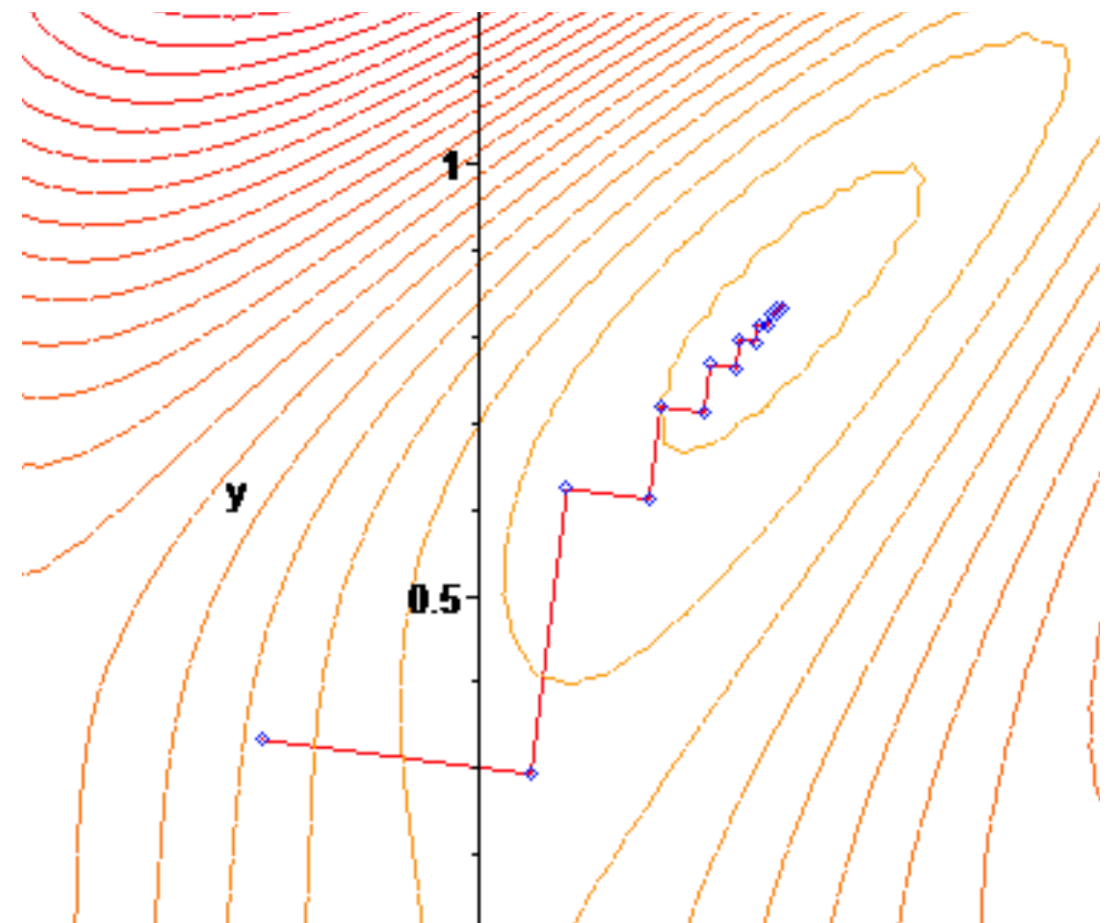
Минимум

$$\cos \phi = -1 \Rightarrow \phi = 180$$

Максимум

$$\cos \phi = 1 \Rightarrow \phi = 0$$

С помощью ряда Тейлора можно доказать, что градиент ортогонален линиям уровня



Стохастический градиентный спуск SGD

Обычный вид функционала

$$Q(w) = \frac{1}{l} \sum_{i=1}^l q_i(w)$$

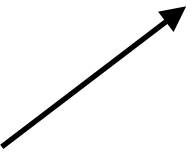
Крутой SGD функционал

$$Q(w) = q_{i_k}(w)$$

Обычный градиентный спуск

$$w^{(k)} = w^{(k-1)} - \eta_k \nabla Q(w^{(k-1)})$$

Считается
медленно



SGD

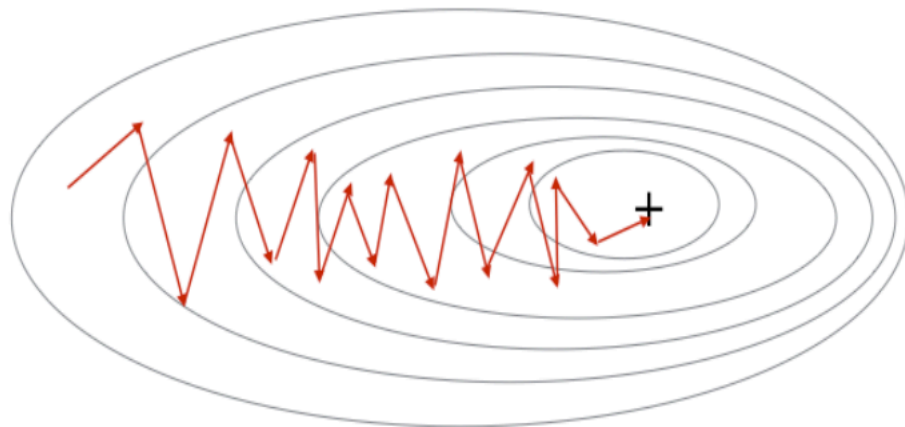
$$w^k = w^{(k-1)} - \eta_k \nabla q_{i_k}(w^{(k-1)})$$

Считается
быстро

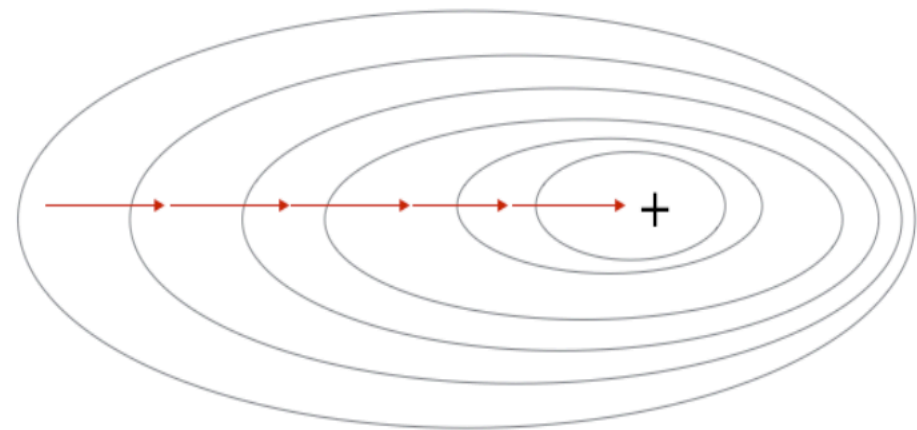


Стохастический градиентный спуск SGD

Stochastic Gradient Descent



Gradient Descent



Средний стохастический градиент SAG

1. Найдем полный градиент для начальной точки

$$z_i^{(0)} = \nabla q_i(w^{(0)}), i = 1, \dots, l$$

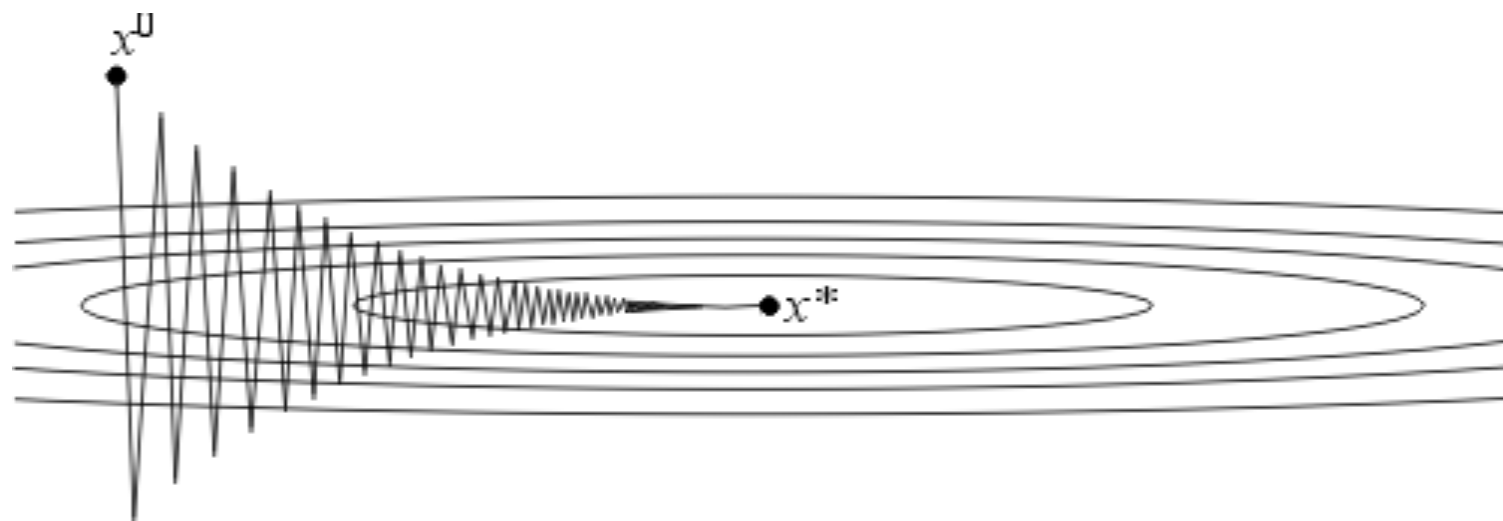
**2. На каждом следующем шаге
будем обновлять только один случайный z**

3. В качестве итогового градиента берем сумму:

$$\frac{1}{l} \sum_{i=1}^l z_i^k$$

Метод импульса Momentum

Движение при обычном градиентном спуске



Метод импульса Momentum

Введем вектор инерции:

$$h_0 = 0$$

$$h_k = \alpha h_{k-1} + \eta_k \nabla Q(w^{k-1})$$

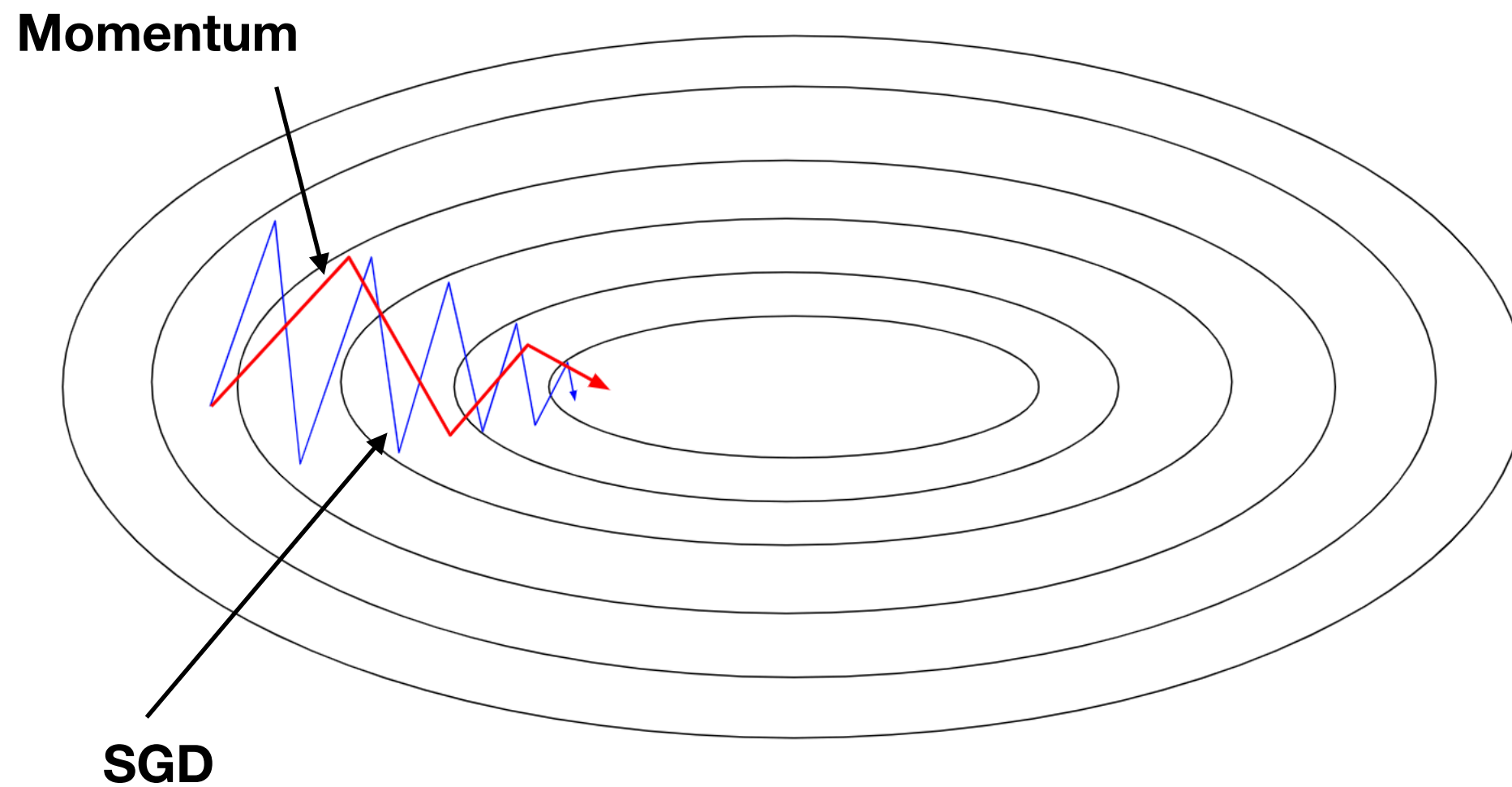
**Параметр, определяющий
скорость затухания**

**Можно использовать
аппроксимацию**

Шаг градиентного спуска:

$$w^k = w^{k-1} - h_k$$

Метод импульса Momentum

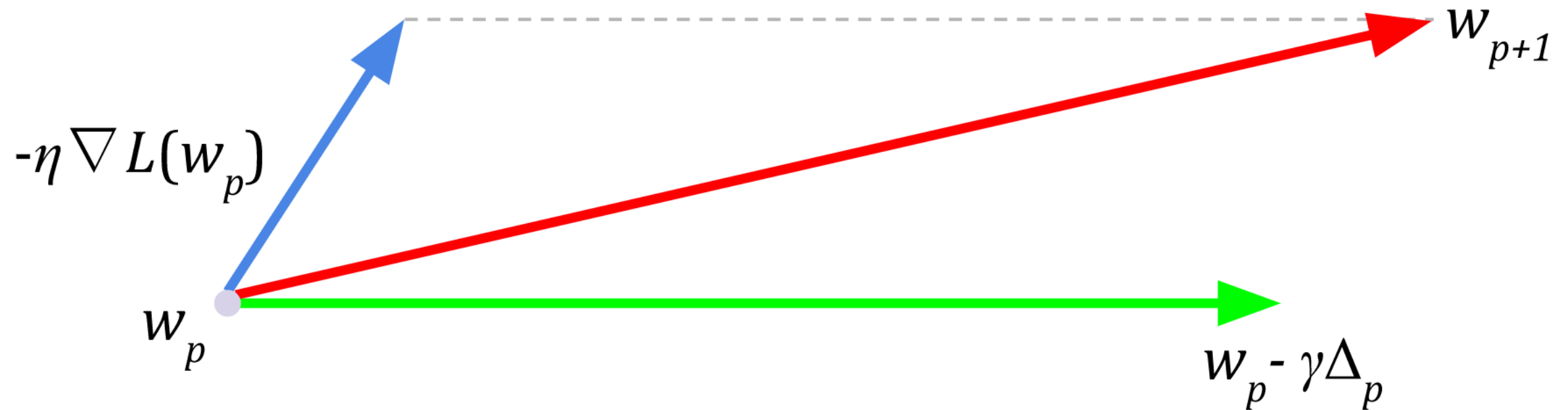


Ускоренные градиенты Нестерова

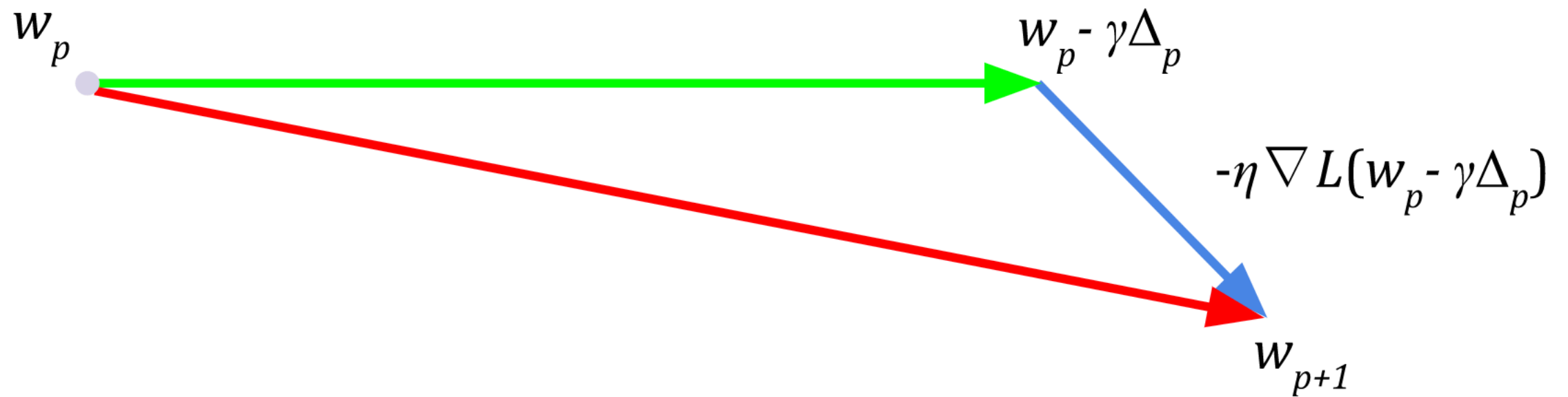
Метод импульса

$$h_k = \alpha h_{k-1} + \eta_k \nabla Q(w^{k-1})$$

$$w^k = w^{k-1} - h_k$$



Ускоренные градиенты Нестерова

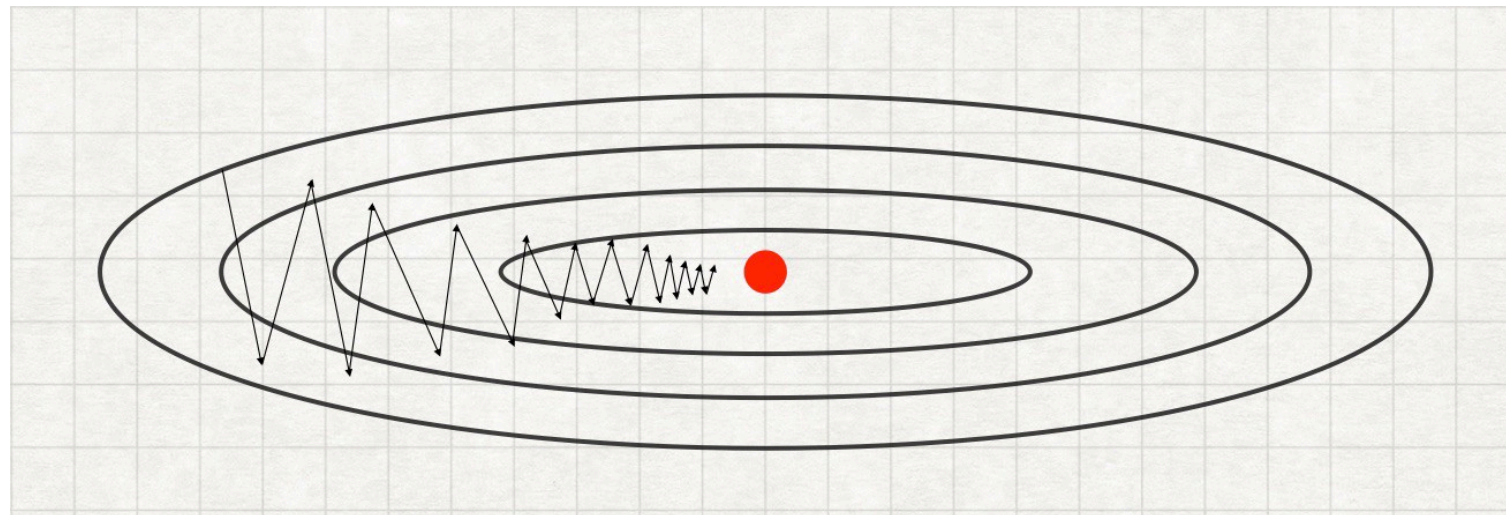


$$h_p = \alpha * h_{p-1} + \eta * \nabla Q(w_{p-1} - \alpha h_{p-1})$$

$$w_p = w_{p-1} - h_p$$

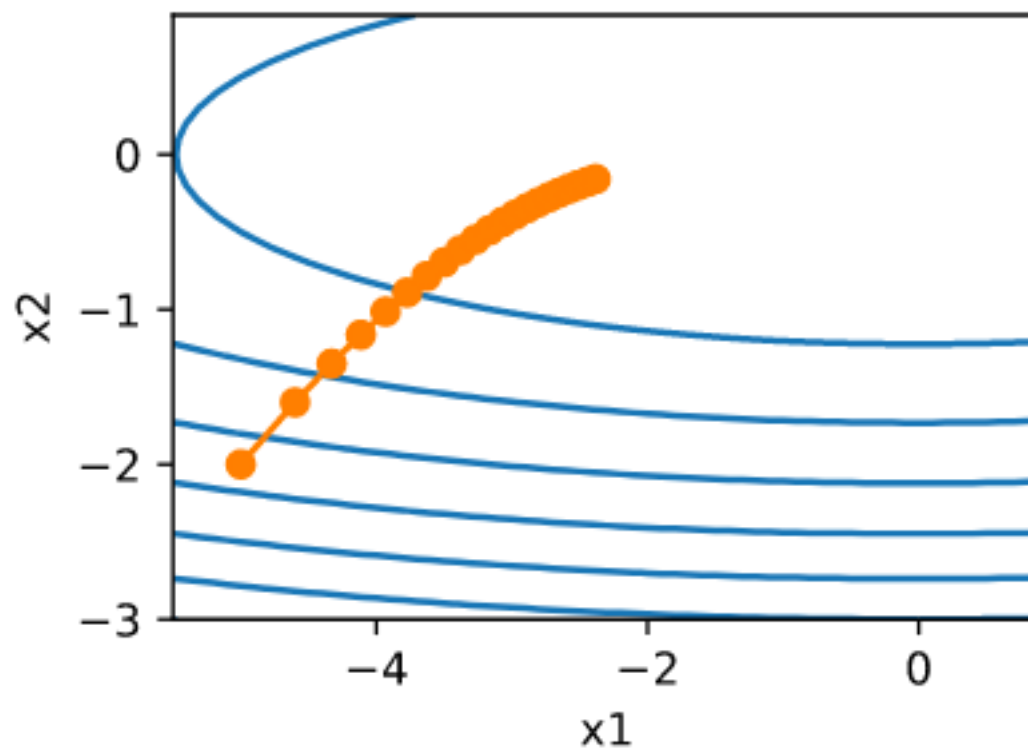
AdaGrad

Движение при обычном градиентном спуске



AdaGrad

Движение при обычном градиентном спуске

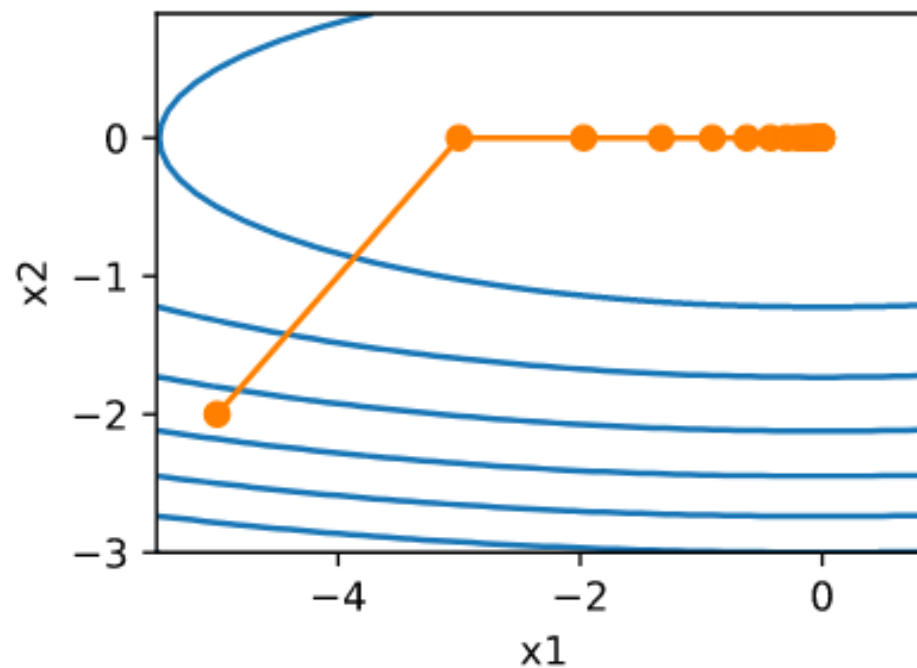


AdaGrad

**Сделаем свою длину шага для
каждой компоненты вектора параметров**

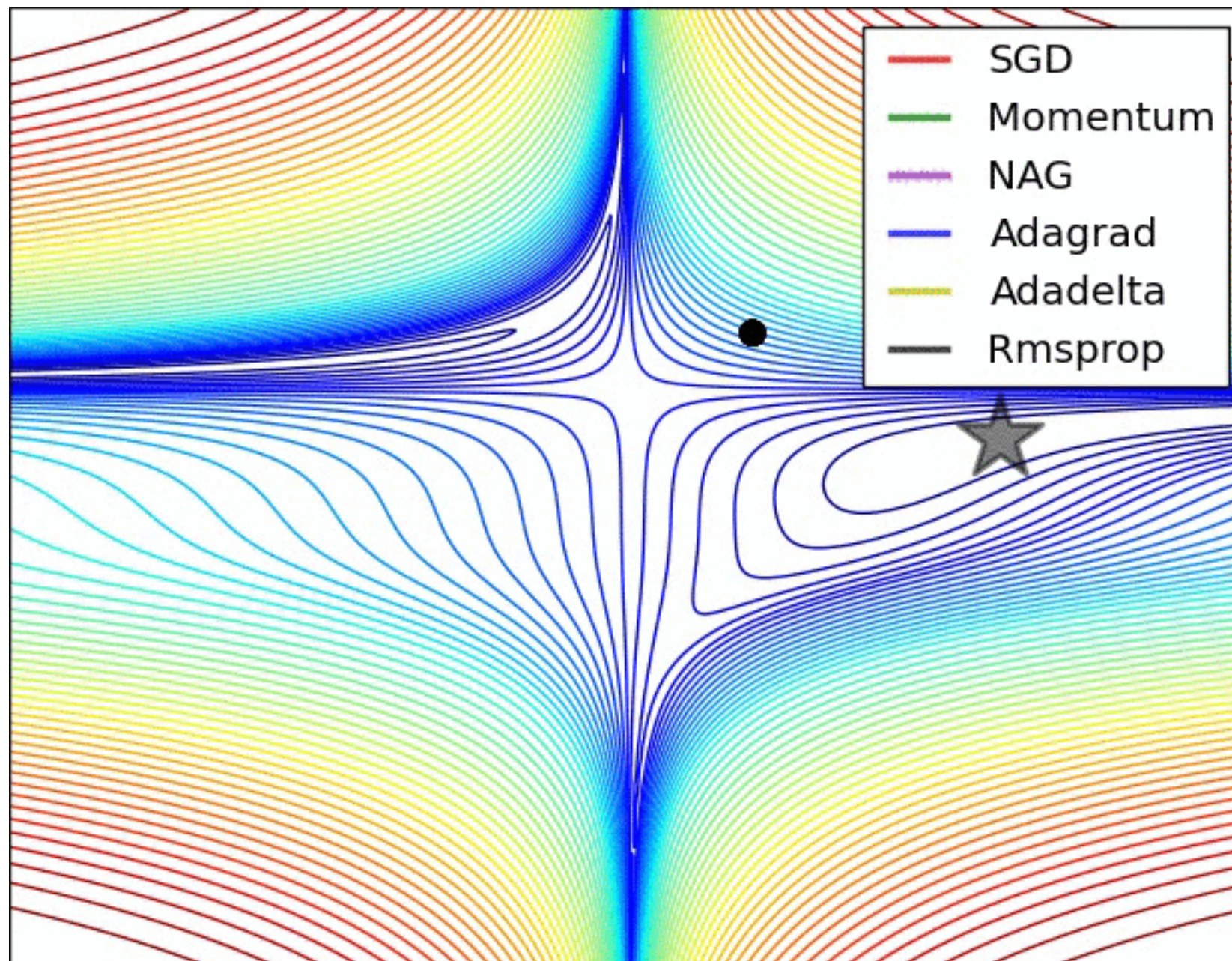
$$G_{k,j} = G_{k-1,j} + (\nabla_w Q(w^{k-1}))_j^2$$

$$w_j^k = w_j^{k-1} - \frac{\eta_t}{\sqrt{G_{k,j} + \varepsilon}} (\nabla_w Q(w^{k-1}))_j$$



RMSprop

$$G_{k,j} = \alpha G_{k-1,j} + (1 - \alpha)(\nabla_w Q(w^{k-1}))_j^2$$



Adam

Найдем первый и второй момент градиентов

$$m_p = \beta_1 m_{p-1} + (1 - \beta_1) \nabla Q(w_{p-1})$$

$$v_p = \beta_2 v_{p-1} + (1 - \beta_2) (\nabla Q(w_{p-1}))^2$$

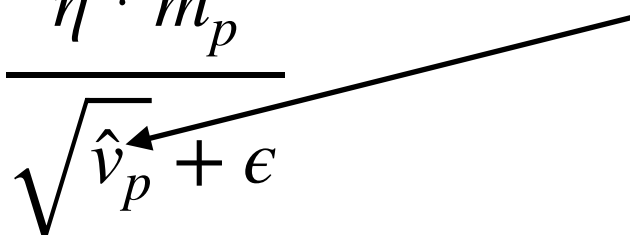
Так как изначальные значения моментов равны 0, а коэффициенты близки к 1, то увеличим наши моменты:

$$\hat{m}_p = \frac{m_p}{1 - \beta_1^p} \qquad \hat{v}_p = \frac{v_p}{1 - \beta_2^p}$$

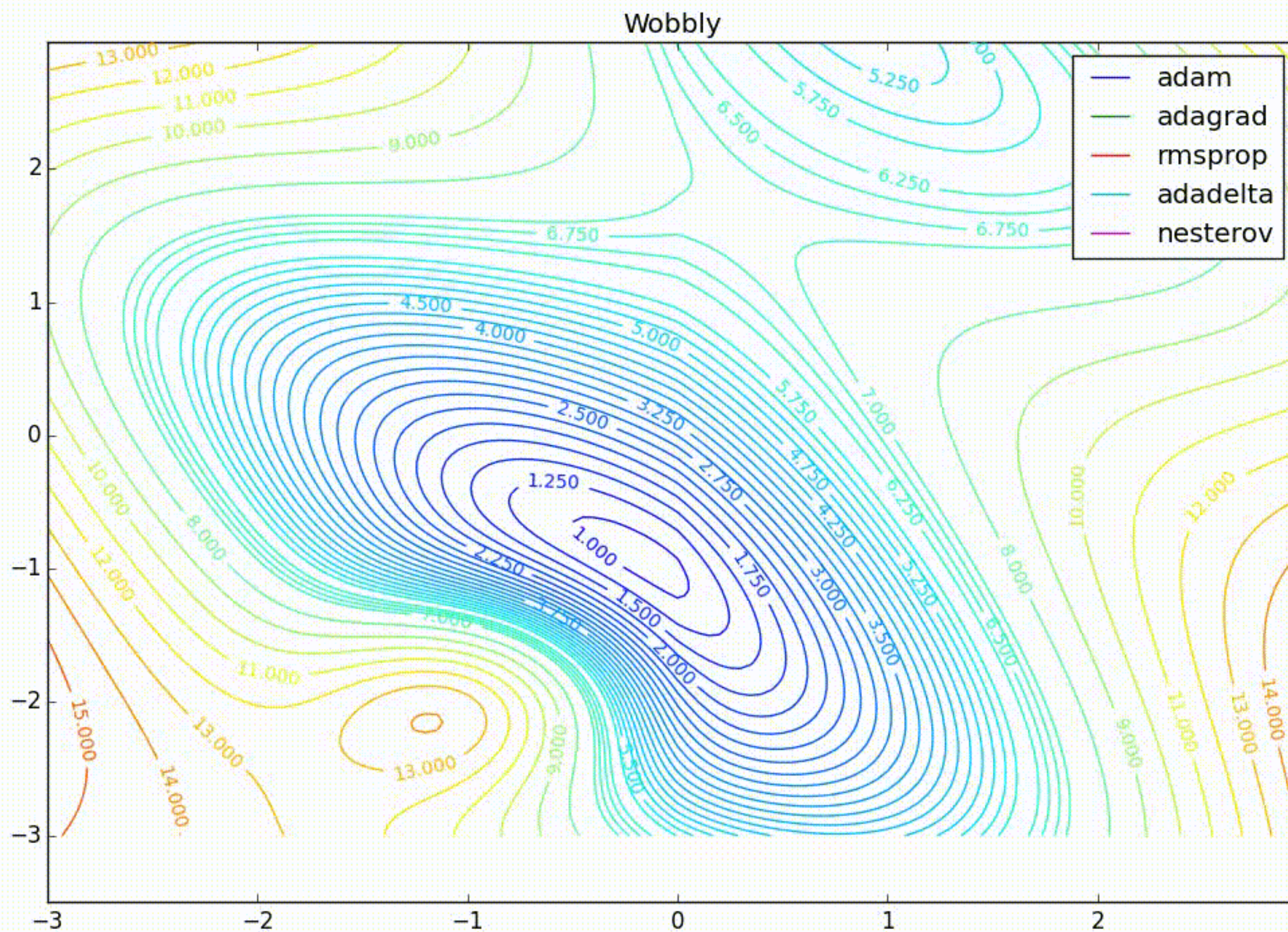
Итог:

$$w_p = w_{p-1} - \frac{\eta * \hat{m}_p}{\sqrt{\hat{v}_p + \epsilon}}$$

Регулирует шаг



Сравнение. Вывод.



Вопросы:

1. Чем SAG лучше и хуже SGD?
2. Формула для итерации в momentum.
3. Когда нужно использовать AdaGrad и RMSprop?

ИСТОЧНИКИ:

- <https://github.com/esokolov/ml-course-hse/blob/master/2019-fall/lecture-notes/lecture02-linregr.pdf>
- <https://ru.wikipedia.org/wiki/Градиент>
- <http://runder.io/optimizing-gradient-descent/index.html#rmsprop>
- <http://www.machinelearning.ru/wiki/index.php?title=Изображение:Grad3.PNG>
- https://d2l.ai/chapter_optimization/adagrad.html
- <https://habr.com/ru/post/318970/>
- <https://medium.com/@congyuzhou/gradient-descent-with-momentum-e3354d7d280d>
- https://datascience-enthusiast.com/DL/Optimization_methods.html
 - https://vbystricky.github.io/2018/03/optimization_grad_desc.html#%D1%83%D1%81%D0%BA%D0%BE%D1%80%D0%B5%D0%BD%D0%BD%D1%8B%D0%B5-%D0%B3%D1%80%D0%B0%D0%B4%D0%B8%D0%B5%D0%BD%D1%82%D1%8B-%D0%BD%D0%B5%D1%81%D1%82%D0%B5%D1%80%D0%BE%D0%B2%D0%B0-nesterov-accelerated-gradient