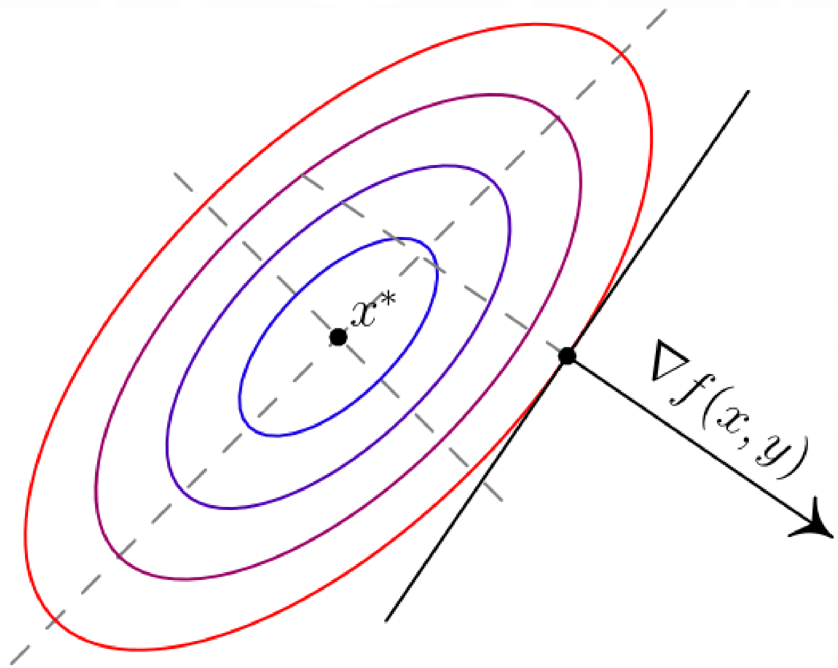
Градиентный спуск

Кантонистова Е.О.

ВШЭ, 2019

ТЕОРЕМА О ГРАДИЕНТЕ

Теорема. Градиент **–** это направление наискорейшего роста функции.



$$f(x,y) = \frac{3}{2}(x-y)^2 + \frac{1}{3}(x+y)^2$$

ГРАДИЕНТНЫЙ СПУСК

Теорема. Градиент **–** это направление наискорейшего роста функции.

Метод градиентного спуска:

- Инициализируем веса $w^{(0)}$.
- На каждом следующем шаге обновляем веса по формуле:

$$\boxed{\boldsymbol{w^{(k)} = w^{(k-1)} - \eta_k \nabla Q\left(w^{(k-1)}\right)}}$$

ГРАДИЕНТНЫЙ СПУСК

Теорема. Градиент **–** это направление наискорейшего роста функции.

Метод градиентного спуска:

- Инициализируем веса $w^{(0)}$.
- На каждом следующем шаге обновляем веса по формуле:

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

Скорость сходимости: $\boldsymbol{Q\left(w^{(k)}\right) - Q(w^*) = O(\frac{1}{k})}$

ГРАДИЕНТНЫЙ СПУСК

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

Градиент функции Q:

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

Градиентный спуск:

$$w^{(k)} = w^{(k-1)} - \eta_k \sum_{i=1}^{l} \nabla q_i(w^{(k-1)})$$

ВАРИАНТЫ ИНИЦИАЛИЗАЦИИ ВЕСОВ

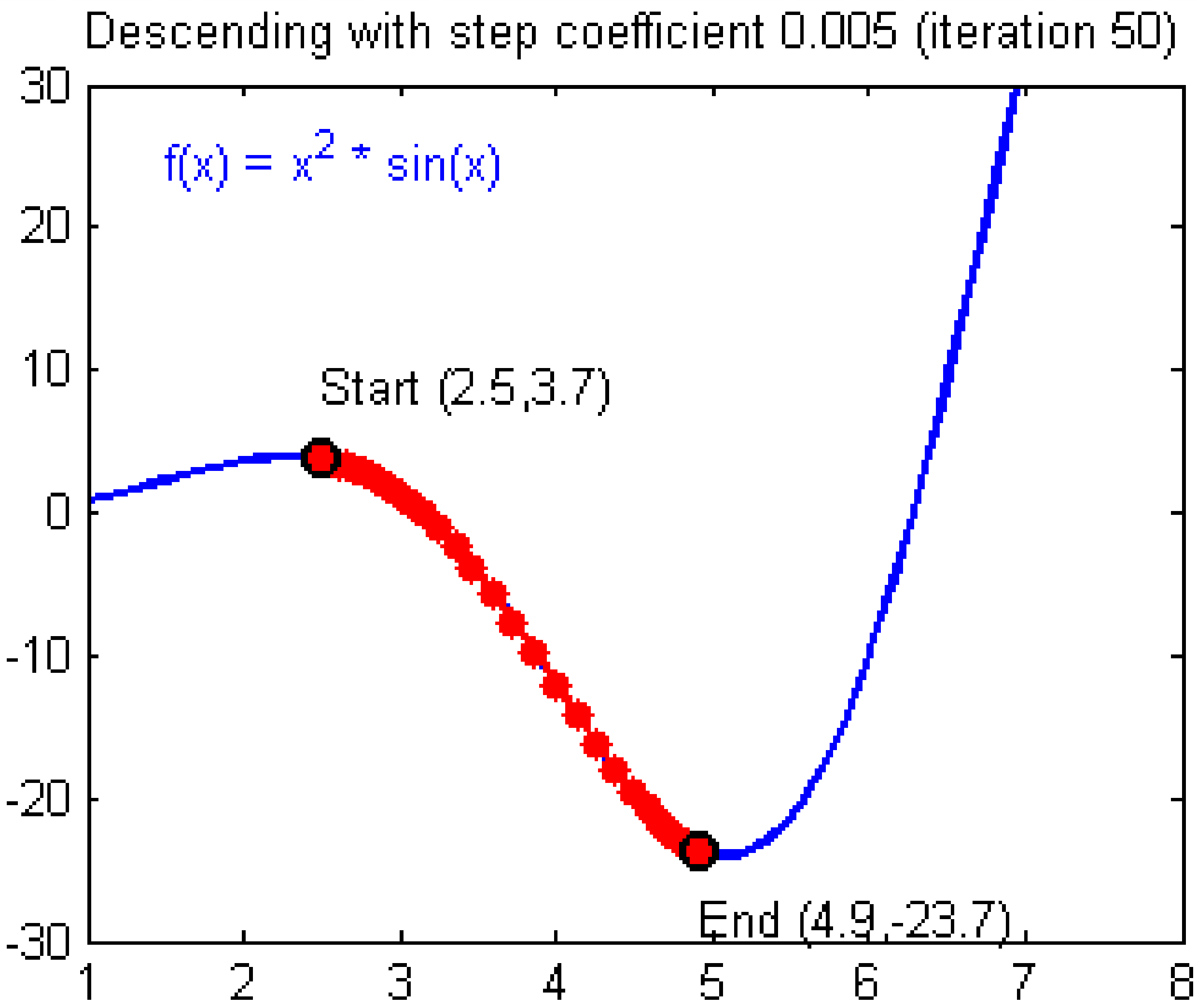
- $w_j = 0, j = 1, \dots, n$
- Небольшие случайные значения:
$$w_j := random(-\varepsilon, \varepsilon)$$
- Обучение по небольшой случайной подвыборке объектов
- Мультистарт: многократный запуск из разных случайных начальных приближений и выбор лучшего решения

КРИТЕРИИ ОСТАНОВА

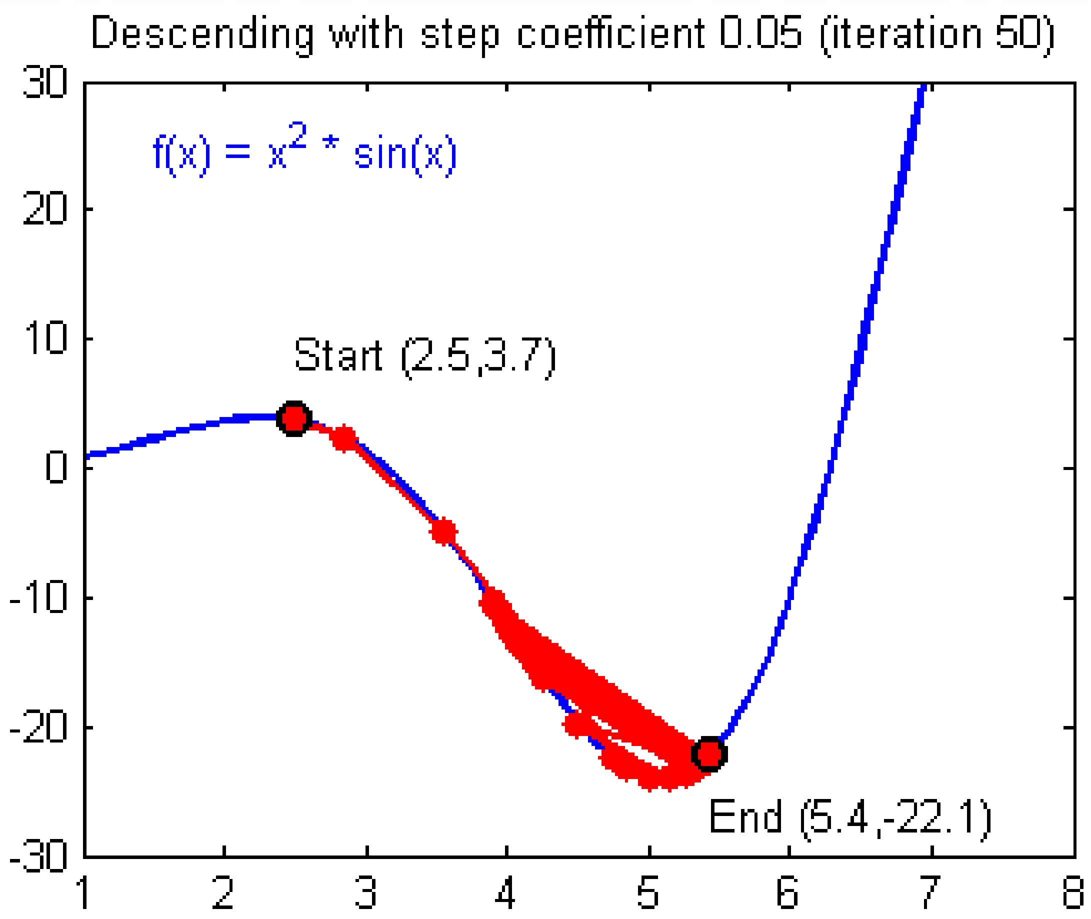
- $\left|\nabla Q\left(w^{(k-1)}\right)\right| < \varepsilon$

- $\Delta w = \left|w^{(k)} - w^{(k-1)}\right| < \varepsilon$

ГРАДИЕНТНЫЙ СПУСК



ПРОБЛЕМА ВЫБОРА ГРАДИЕНТНОГО ШАГА



ГРАДИЕНТНЫЙ ШАГ

- $\eta_k = c$
- $\eta_k = \frac{1}{k}$
- $\eta_k = \lambda \left(\frac{s_0}{s_0+k}\right)^p$, λ, s_0, p - параметры

МЕТОДЫ ОЦЕНИВАНИЯ ГРАДИЕНТА: SGD

1) Stochastic gradient descent (SGD):

- на каждом шаге выбираем один случайный объект и сдвигаемся в сторону антиградиента по этому объекту:

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

МЕТОДЫ ОЦЕНИВАНИЯ ГРАДИЕНТА: SGD

1) Stochastic gradient descent (SGD):

- на каждом шаге выбираем один случайный объект и сдвигаемся в сторону антиградиента по этому объекту:

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

Скорость сходимости: $\mathrm{E}[\boldsymbol{Q}\left(\boldsymbol{w}^{(\boldsymbol{k})}\right) - \boldsymbol{Q}(\boldsymbol{w}^{*})] = \boldsymbol{O}(\frac{\mathbf{1}}{\sqrt{\boldsymbol{k}}})$

МЕТОДЫ ОЦЕНИВАНИЯ ГРАДИЕНТА: SGD

1) Stochastic gradient descent (SGD):

- на каждом шаге выбираем один случайный объект и сдвигаемся в сторону антиградиента по этому объекту:

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

Скорость сходимости: $\mathrm{E}[\boldsymbol{Q}\left(\boldsymbol{w}^{(k)}\right) - \boldsymbol{Q}(\boldsymbol{w}^{*})] = \boldsymbol{O}(\frac{\mathbf{1}}{\sqrt{\boldsymbol{k}}})$

\+ **Менее трудоемкий метод**

\- **Медленнее сходится**

МЕТОДЫ ОЦЕНИВАНИЯ ГРАДИЕНТА: SAG

2) Stochastic average gradient (SAG):

- Инициализируем веса w_j
- Инициализируем вспомогательные переменные $z^{(1)}, z^{(2)}, \dots$:

$$z^{(i)} = \nabla q_i(w)$$

МЕТОДЫ ОЦЕНИВАНИЯ ГРАДИЕНТА: SAG

2) Stochastic average gradient (SAG):

- Инициализируем веса w_j
- Инициализируем вспомогательные переменные $z^{(1)}, z^{(2)}, \dots$:

$$z^{(i)} = \nabla q_i(w)$$

- На каждом шаге выбираем один случайный объект и обновляем градиент по нему:

$$z_i^{(k)} = \begin{cases} \nabla q_i\left(w^{(k-1)}\right), i = i_k \\ z_i^{(k-1)}, \text{иначе} \end{cases}$$

Методы оценивания градиента: SAG

2) Stochastic average gradient (SAG):

- Инициализируем веса w_j
- Инициализируем вспомогательные переменные $z^{(1)}, z^{(2)}, \ldots$:

$$z^{(i)} = \nabla q_i(w)$$

- На каждом шаге выбираем один случайный объект и обновляем градиент по нему:

$$z_i^{(k)} = \begin{cases} \nabla q_i\left(w^{(k-1)}\right), i = i_k \\ z_i^{(k-1)}, \text{иначе} \end{cases}$$

- Формула градиентного шага:

$$\boldsymbol{w^{(k)} = w^{(k-1)} - \eta_k \sum_{i=1}^{l} z_i^{(k)}}$$

МЕТОДЫ ОЦЕНИВАНИЯ ГРАДИЕНТА: SAG

2) Stochastic average gradient (SAG):

- Формула градиентного шага:

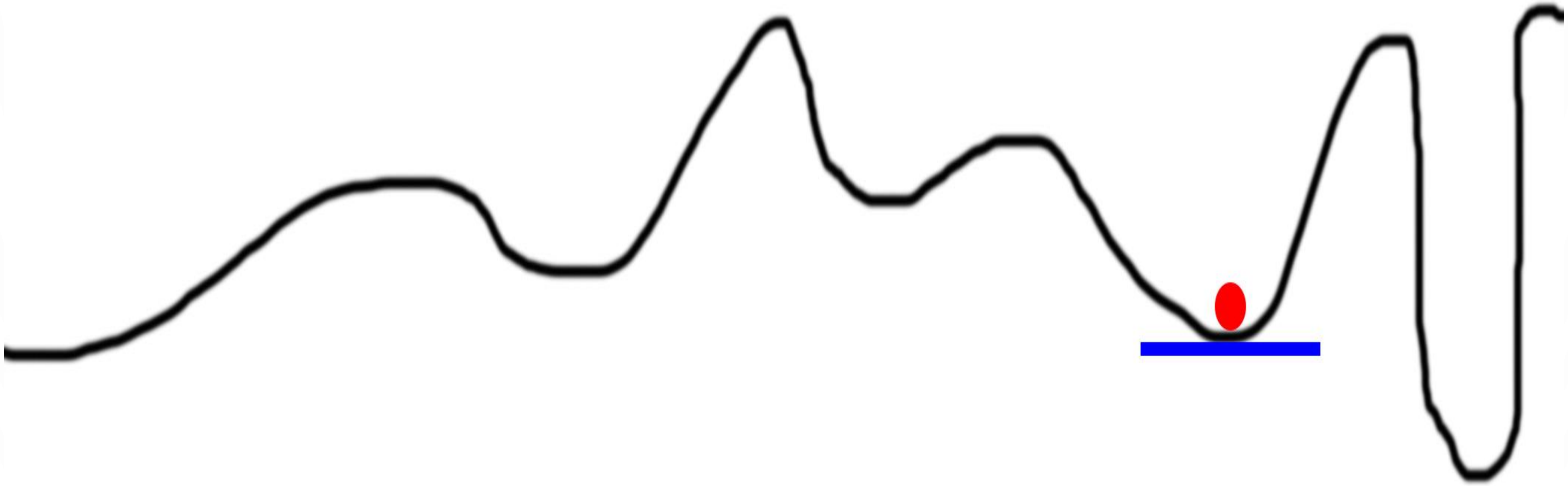
$$w^{(k)} = w^{(k-1)} - \eta_k \sum_{i=1}^{l} z_i^{(k)}$$

Скорость сходимости: $\mathbf{E}[\boldsymbol{Q}\left(\boldsymbol{w}^{(k)}\right) - \boldsymbol{Q}(\boldsymbol{w}^{*})] = \boldsymbol{O}(\frac{\mathbf{1}}{\boldsymbol{k}})$

ПРОБЛЕМЫ ГРАДИЕНТНОГО СПУСКА

- Медленно сходится
- Застревает в локальных минимумах

ПРОБЛЕМА ЗАСТРЕВАНИЯ В LOCMIN



МЕТОД МОМЕНТОВ (MOMENTUM)

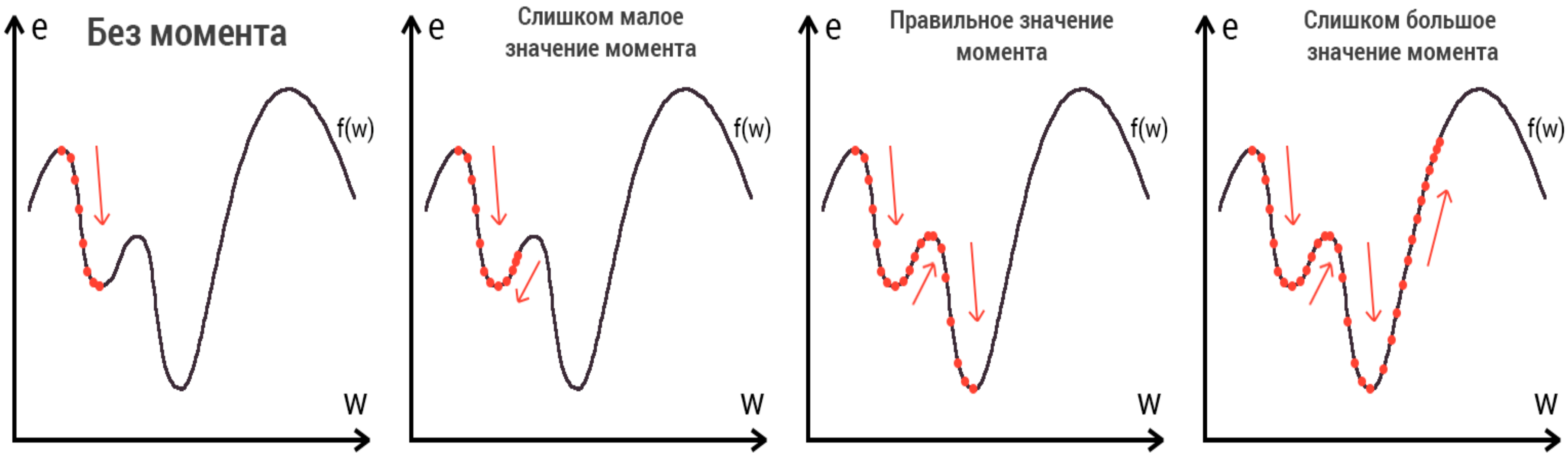
Вектор инерции (*усреднение градиента по предыдущим шагам*):

$$h_0 = 0;$$
$$h_k = \alpha h_{k-1} + \eta_k \nabla_w Q(w^{(k-1)})$$

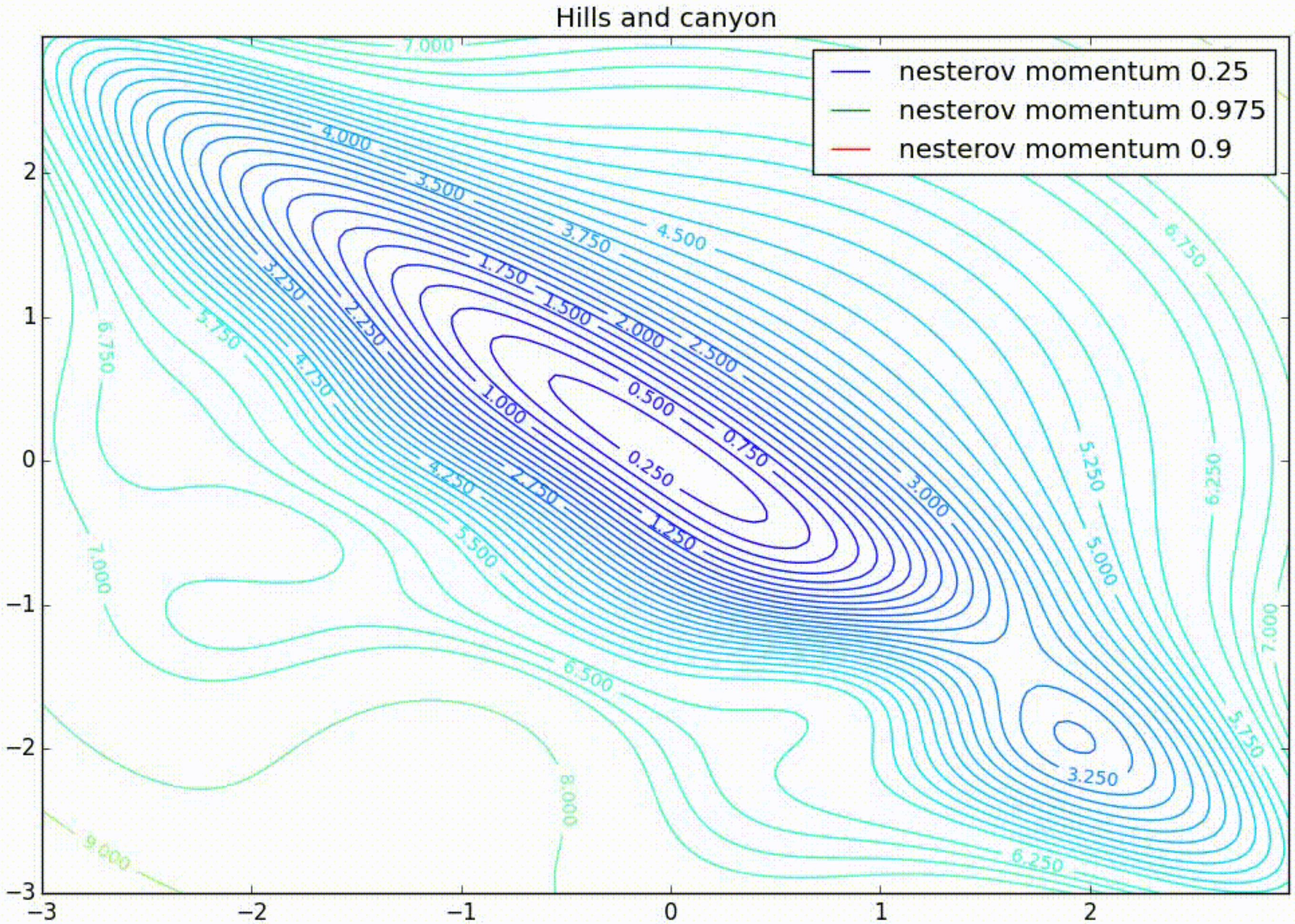
Формула метода моментов:

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

MOMENTUM


Без момента
Слишком малое значение момента
Правильное значение момента
Слишком большое значение момента
e
f(w)
w

MOMENTUM



Hills and canyon
nesterov momentum 0.25
nesterov momentum 0.975
nesterov momentum 0.9

ADAGRAD (ADAPTIVE GRADIENT)

Сумма квадратов обновлений:

$$g_{k-1,j} = (\nabla Q(w^{(k-1)}))_j^2$$

Формулы метода AdaGrad:

- $G_{k,j} = G_{k-1,j} + g_{k-1.j} = G_{k-1,j} + (\nabla Q(w^{(k-1)}))_j^2$
- $\omega_j^{(k)} = \omega_j^{k-1} - \frac{\eta}{\sqrt{G_{k,j}+\varepsilon}} \cdot \left(\nabla Q\left(w^{(k-1)}\right)\right)_j$

ADAGRAD (ADAPTIVE GRADIENT)

Сумма квадратов обновлений:

$$g_{k-1,j} = (\nabla Q(w^{(k-1)}))_j^2$$

Формулы метода AdaGrad:

- $G_{k,j} = G_{k-1,j} + g_{k-1.j}$
- $\omega_j^{(k)} = \omega_j^{k-1} - \frac{\eta}{\sqrt{G_{k,j}+\varepsilon}} \cdot \left(\nabla Q\left(w^{(k-1)}\right)\right)_j$

+ Автоматическое затухание скорости обучения

- G_{kj} монотонно возрастают, поэтому шаги укорачиваются, и мы можем не успеть дойти до минимума

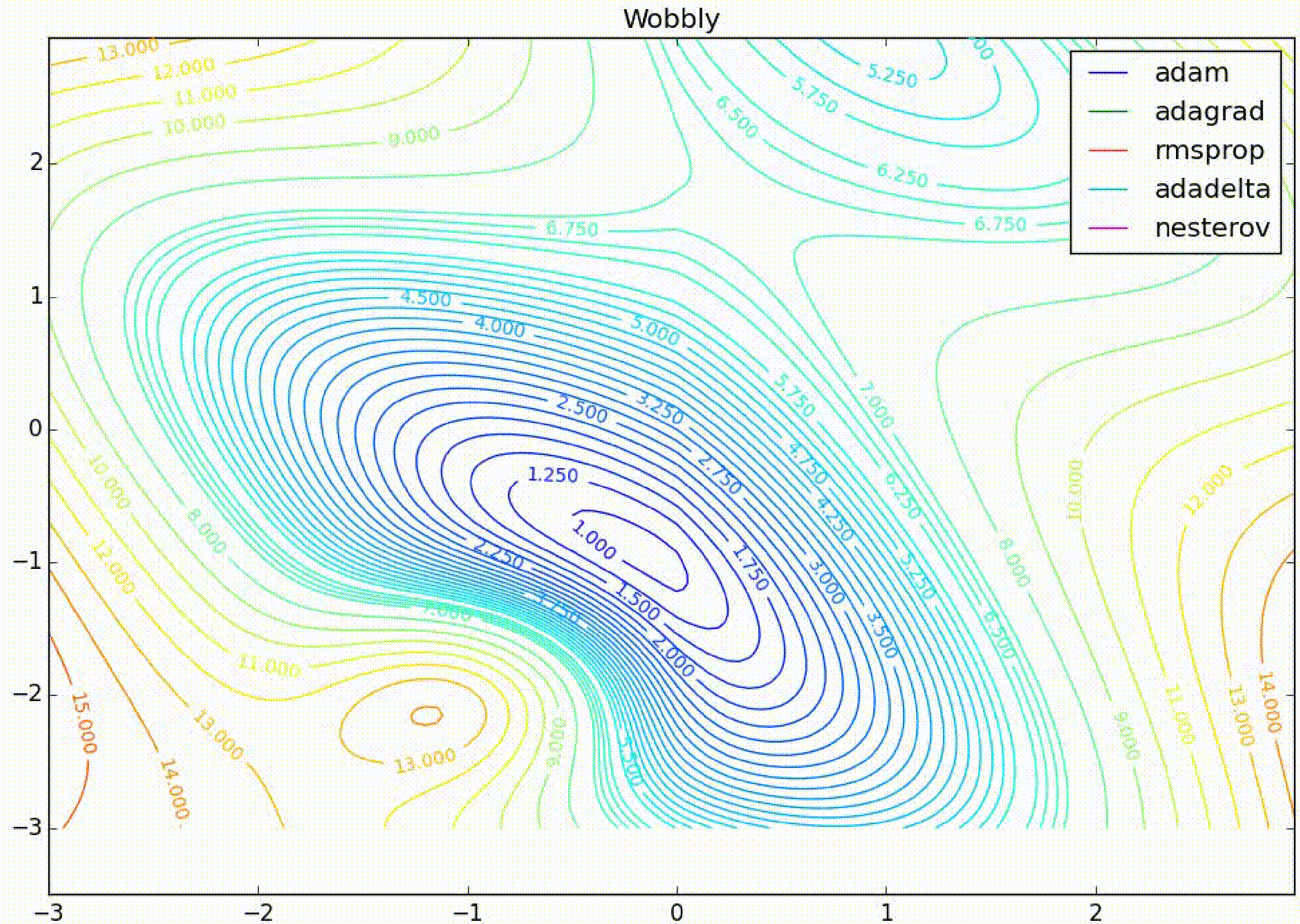
RMSPROP (ROOT MEAN SQUARE PROPAGATION)

Метод реализует экспоненциальное затухание градиентов

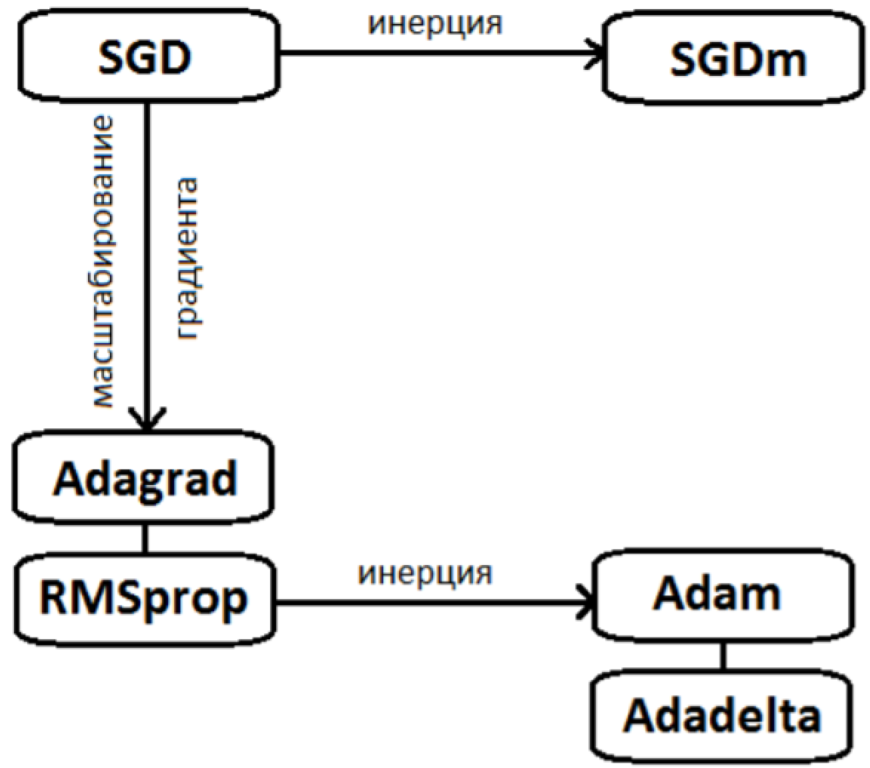
Формулы метода RMSprop (*усредненный по истории квадрат градиента*):

- $G_{k,j} = \boldsymbol{\alpha} \cdot G_{k-1,j} + \mathbf{(1-\boldsymbol{\alpha})} \cdot g_{k-1,j}$
- $\omega_j^{(k)} = \omega_j^{k-1} - \frac{\eta}{\sqrt{G_{k,j}+\varepsilon}} \cdot \left(\nabla Q\left(w^{(k-1)}\right)\right)_j$

МОДИФИКАЦИИ ГРАДИЕНТНОГО СПУСКА



МОДИФИКАЦИИ SGD



http://www.machinelearning.ru/wiki/images/a/a0/2016_417_ChabanenkoVD.pdf