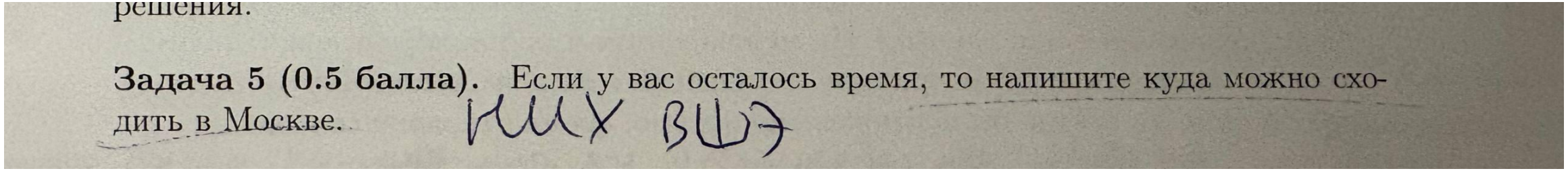
Модификации GD

Семинар 5


решения.

Задача 5 (0.5 балла). Если у вас осталось время, то напишите куда можно сходить в Москве. КИХ ВШЭ

План

- Разбор проверочной
- Модификации GD
- Метрики классификации

Разбор проверочной

Задача 1 (2 балла). Предложите для каждой из перечисленных ниже задач, как сформулировать их в терминах машинного обучения: укажите, что будет являться объектом и целевой переменной, а также напишите тип задачи (например, регрессия).

1. Доставка еды "ТЫндекс Еда"хочет научиться определять время, которое займет доставка продуктов курьером. Для предсказания им доступны координаты магазина выдачи и координаты клиента, время доставки, а так же информация о пробках в городе в этот момент.

Объект: Заказ/доставка
Целевая переменная: Время доставки
Тип задачи: Регрессия

Разбор проверочной

2. Новый мессенджер ШМАКС хочет создать систему автоматического определения фрода(умышленный обман, совершаемый для получения незаконной прибыли, кражи денег или данных). Им нужна программа для определения, является ли сообщение фродом или нет. Сообщение содержит текст, имя автора и получателя, ip, откуда было отправлено письмо, время отправки сообщения.

Объект: Сообщение
Целевая переменная: Фрод/не фрод
Тип задачи: Бинарная классификация

Разбор проверочной

Задача 2 (3 балла). Ответьте на вопросы:

1. Как выглядит модель линейной регрессии? Напишите формулу линейной регрессии и напишите явную формулу решения для матрицы весов w. Что означают слишком большие веса в модели линейной регрессии? К чему они могут привести?

$$a(x) = w_0 + \sum_{j=1}^{d} w_j x_j = \langle w, x \rangle.$$

$a(X) = Xw$

$w = (X^T X)^{-1} X^T y$

Большие веса - это признак переобучения.
Модель придает очень большое значение определенным признакам, небольшое изменение такого признака приведет к огромному изменению предсказания.

Разбор проверочной

2. Что такое функция потерь? Выпишите функцию потерь MSE, MAE и Huber Loss. Приведите пример, когда одна из функций может работать лучше другой?

$$MSE = (a(x) - y)^2$$

$$MAE = |a(x) - y|$$

$$HuberLoss = \begin{cases} \frac{1}{2}(y-a)^2, & |y-a| < \delta \\ \delta\left(|y-a| - \frac{1}{2}\delta\right), & |y-a| \geqslant \delta \end{cases}$$

MAE более устойчива к выбросам, а MSE с точки зрения оптимизации

Если в данных много выбросов, то MSE будет пытаться переобучиться на них, сильно искажая модель. В таком случае Huber Loss или MAE будут работать лучше, так как они не придают выбросам слишком большого значения.

Разбор проверочной

3. Что такое переобучение? Как помогает бороться с переобучением регуляризация? Запишите формулы для L1- и L2-регуляризации. В чем их отличие после обучения модели на каждой из этих регуляризаций? За что отвечает коэффициент регуляризации λ? Что будет происходить с моделью при его увеличении/уменьшении?

Переобучение - ситуация, когда модель слишком хорошо запоминает обучающие данные, улавливая не только основные закономерности, но и шум и выбросы. В результате модель отлично показывает себя на обучающей выборке, но плохо работает на новых, невиданных ранее данных.

Регуляризация - метод борьбы с переобучением путем штрафа за большие веса модели. Регуляризация добавляет в функцию потерь штрафное слагаемое, которое заставляет веса быть маленькими.

$$L_1 = Loss + \lambda \sum_{i=0}^{n} |w_i| \qquad L_2 = Loss + \lambda \sum_{i=0}^{n} w_i^2$$

Разбор проверочной

3. Что такое переобучение? Как помогает бороться с переобучением регуляризация? Запишите формулы для L1- и L2-регуляризации. В чем их отличие после обучения модели на каждой из этих регуляризаций? За что отвечает коэффициент регуляризации λ? Что будет происходить с моделью при его увеличении/уменьшении?

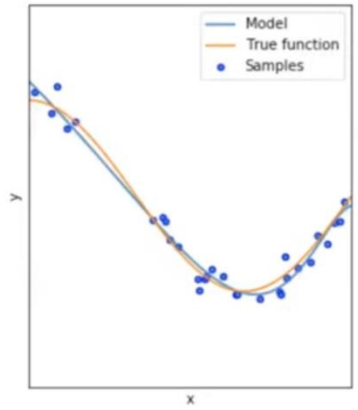
$$L_1 = Loss + \lambda \sum_{i=0}^{n} |w_i| \qquad L_2 = Loss + \lambda \sum_{i=0}^{n} w_i^2$$

L1 регуляризация может обнулять веса, L2 регуляризация стремится сделать веса равномерно маленькими, но редко обнуляет их полностью.

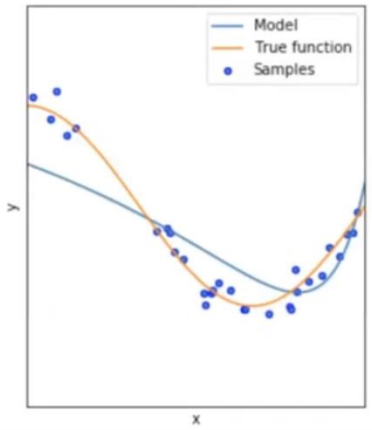
λ - гиперпараметр, который контролирует силу регуляризации.

Разбор проверочной

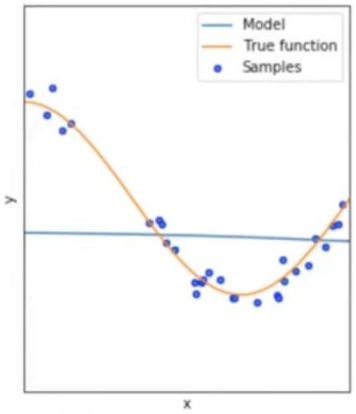
$$\frac{1}{\ell}\sum_{i=1}^{\ell}(a(x_i)-y_i)^2 + 0.01\ \|w\|^2 \to \min_w$$



$$\frac{1}{\ell}\sum_{i=1}^{\ell}(a(x_i)-y_i)^2 + 1\ \|w\|^2 \to \min_w$$



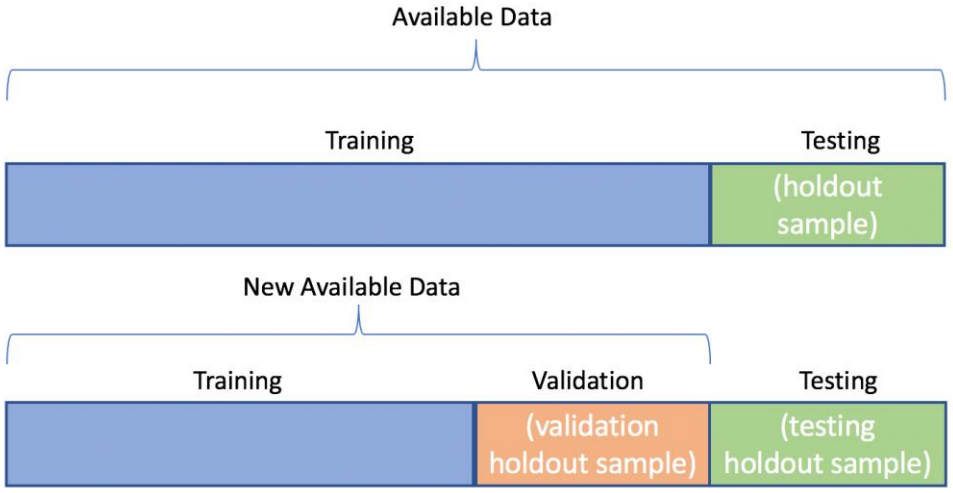
$$\frac{1}{\ell}\sum_{i=1}^{\ell}(a(x_i)-y_i)^2 + 100\ \|w\|^2 \to \min_w$$



Разбор проверочной

Задача 3 (2 балла). Для чего нужны обучающая и валидационная выборки? Почему нельзя обойтись только обучающей выборкой? Расскажите,какпроходиткросс-валидация. В чем плюсы и минусы кросс-валидации?

Train - для обучения модели, для подбора параметров модели(например, весов w в линейной регрессии).
Val - для оценки качества модели после обучения и для настройки ее гиперпараметров (например, λ)
Если оценивать модель на тех же данных, на которых она обучалась, мы получим слишком оптимистичную оценку. Мы не узнаем, как модель поведет себя в реальном мире. Тестовая выборка нужна для честной оценки способности модели к обобщению.



Разбор проверочной

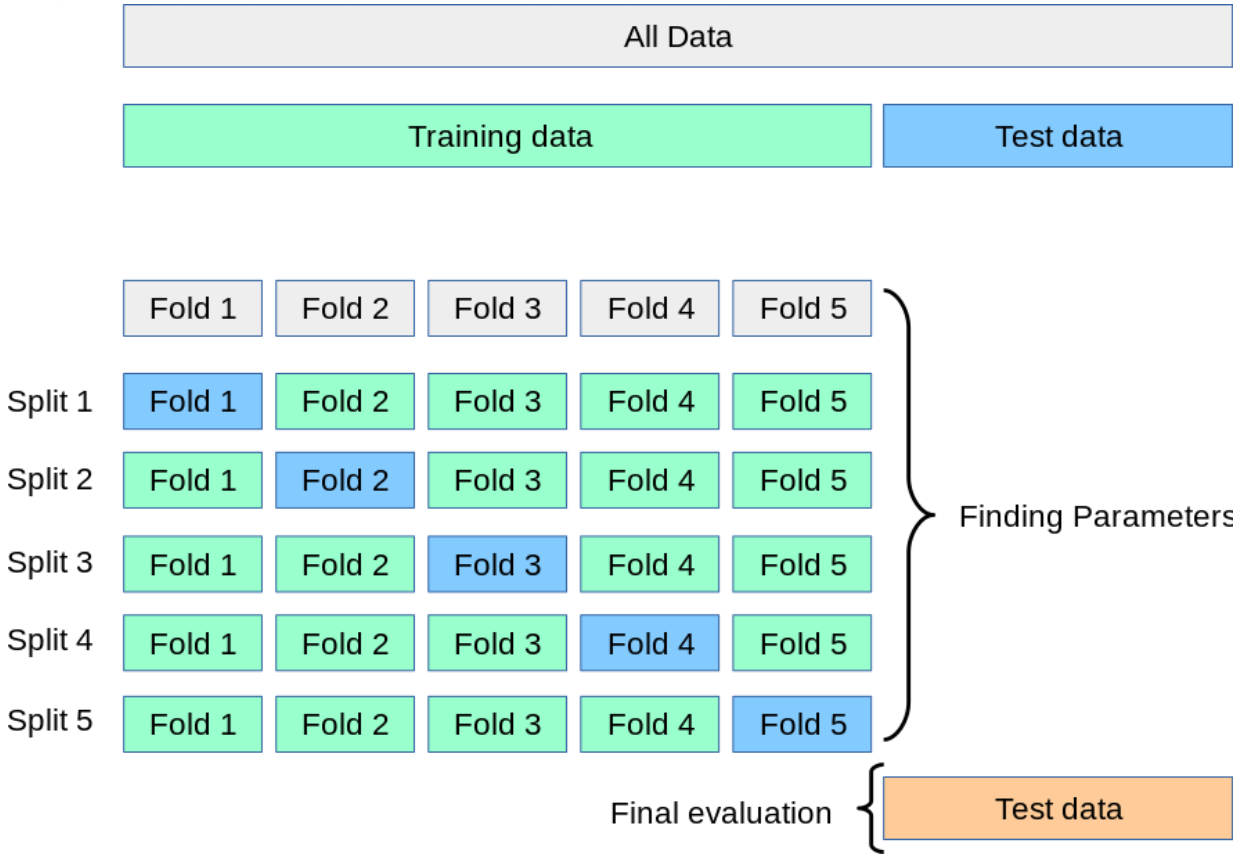
Задача 3 (2 балла). Для чего нужны обучающая и валидационная выборки? Почему нельзя обойтись только обучающей выборкой? Расскажите,какпроходиткросс-валидация. В чем плюсы и минусы кросс-валидации?

Плюсы:

- Более надежная оценка качества модели

Минусы:

- Высокая вычислительная стоимость, так как модель нужно обучать k раз



Разбор проверочной

Задача 4 (3 балла). Допустим, что есть онлайн-курс по машинному обучению. Преподаватели в какой-то момент решили научиться по результатам решения домашних заданий предсказывать, как хорошо студенты будут писать тесты после них. Пусть у них есть данные со следующими признаками:

- student – имя и фамилия студента
- background – какой бэкграунд у студента (математический, гуманитарный или экономический)
- date – дата сдачи домашнего задания
- task_1_points – количество баллов за 1 задание ДЗ
- task_2_points – количество баллов за 2 задание ДЗ
- hw_comments – комментарий преподавателя к решению ДЗ
- hw_points – общее количество баллов за ДЗ
- test_points – общее количество баллов за тест

Известно, что за каждое задание в ДЗ можно получить от 0 до 5 баллов. За тест может ставиться любая оценка из отрезка $[0, 100]$, поэтому было решено использовать модель линейной регрессии. Какую предобработку данных нужно сделать? Есть ли какие-то проблемы с данными? Если да, то поясните какие и предложите пути их решения.

Модификации GD

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

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

$$w^{(k)} = w^{(k-1)} - \eta_k \nabla_w Q(w^{(k-1)})$$ Mini-batch SGD

$$\nabla_w Q(w) \approx \frac{1}{n} \sum_{j=1}^{n} \nabla_w q_{i_{kj}}(w),$$

Модификации GD

Проблемы:

- метод может застревать в локальных минимумах, и у нас **нет никакой гарантии попадания в глобальный минимум**

Модификации GD

Проблемы:

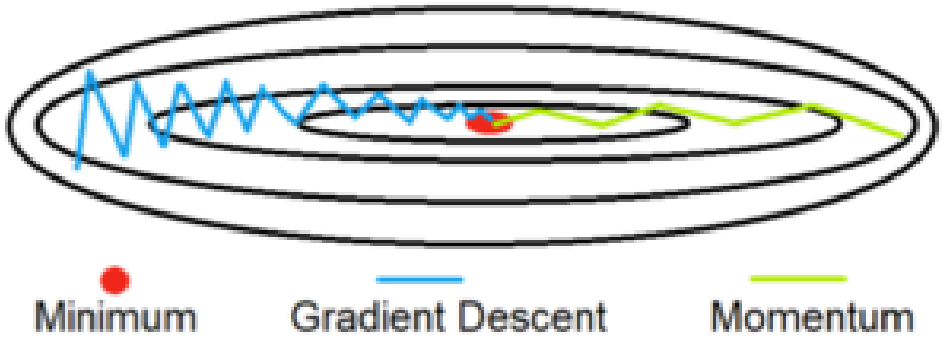
- **метод может медленно сходиться** из-за разной скорости оптимизации по каждому весу в отдельности

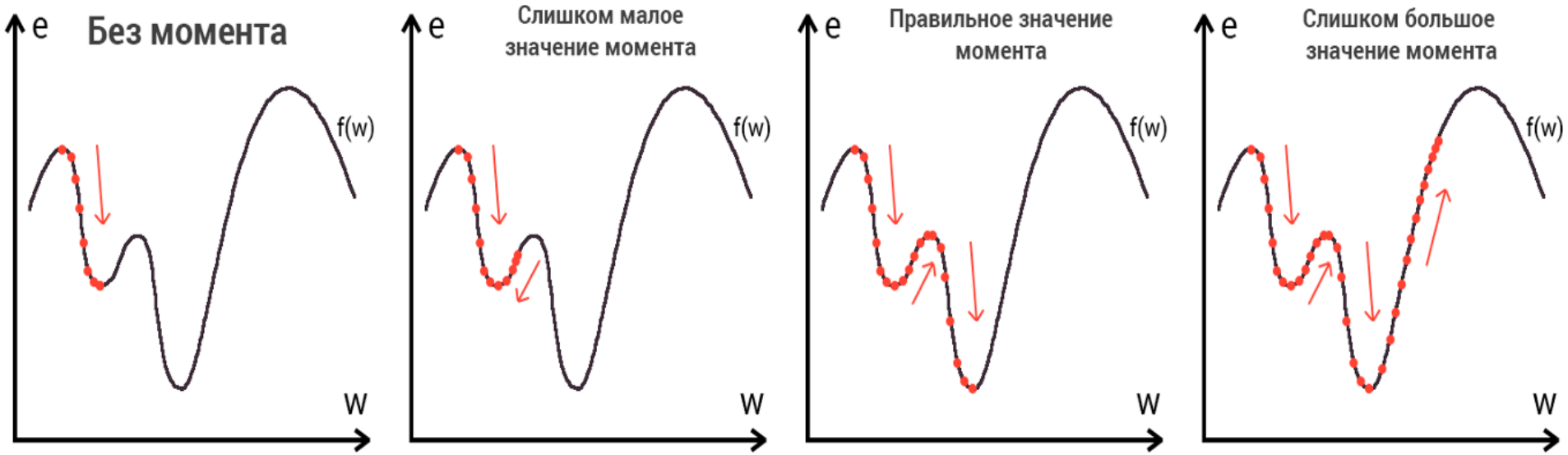
momentum

$$h_0 = 0;$$

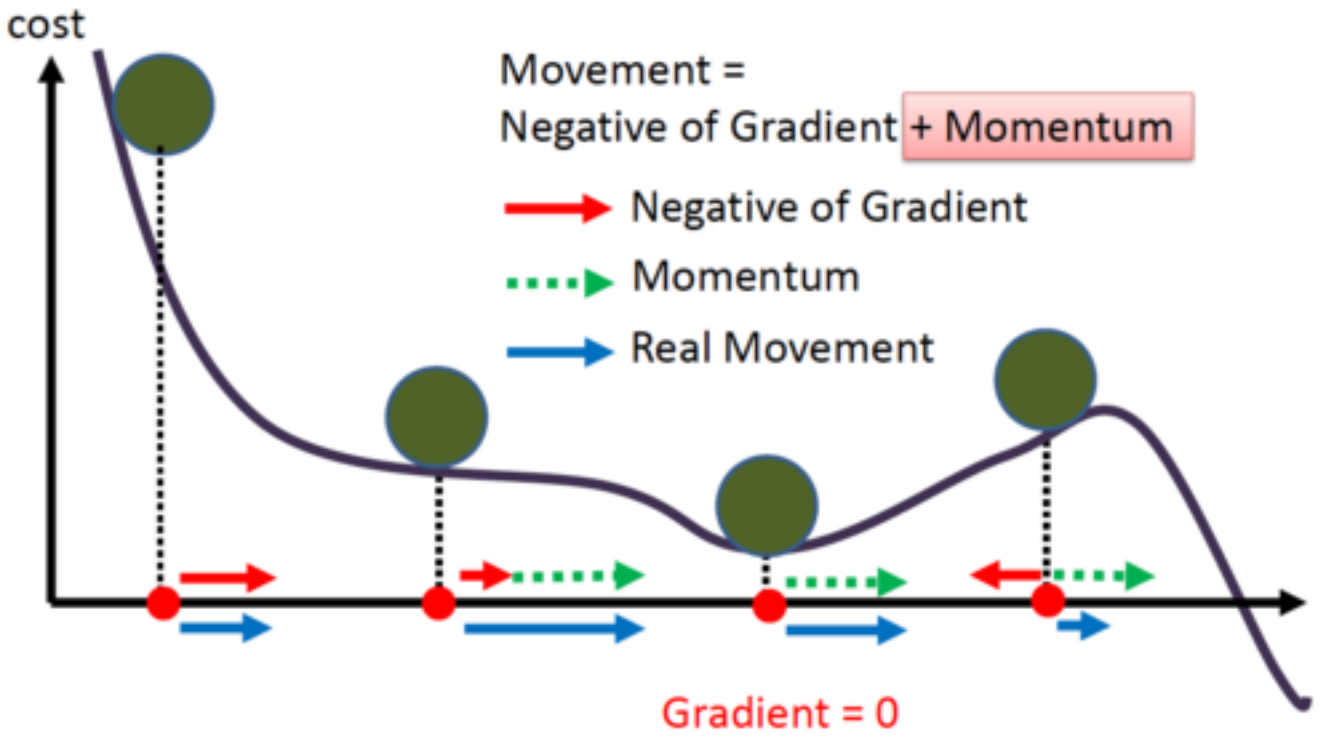
$$h_k = \alpha h_{k-1} + \eta_k \nabla_w Q(w^{(k-1)})$$

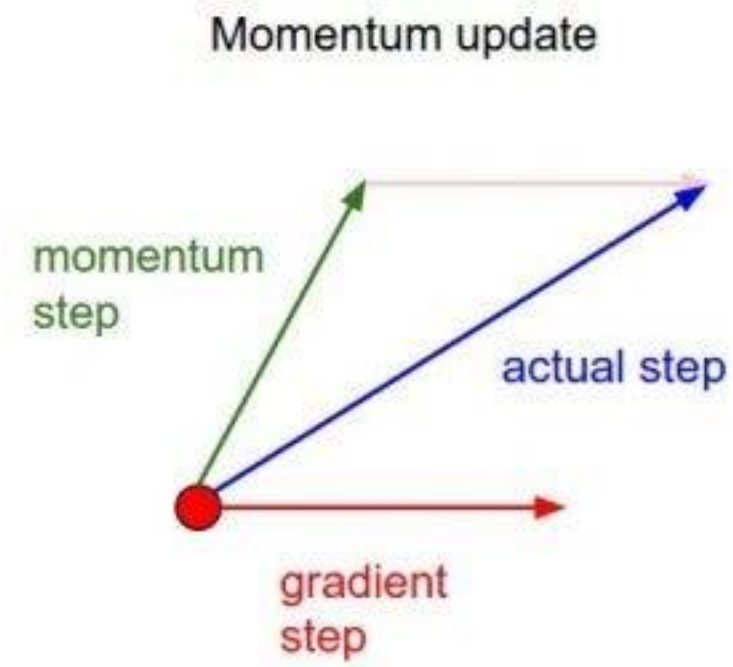
$$w^{(k)} = w^{(k-1)} - h_k$$





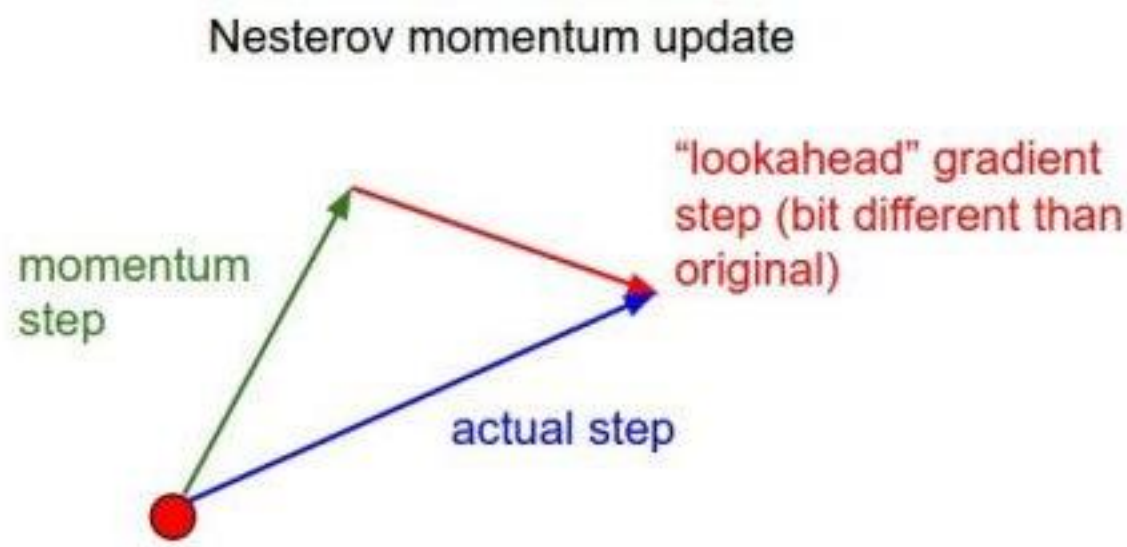
momentum


cost
Movement =
Negative of Gradient + Momentum
Negative of Gradient
Momentum
Real Movement
Gradient = 0


Momentum update
momentum step
actual step
gradient step

Nesterov momentum

$$h_t = \alpha \cdot h_{t-1} + \eta \cdot \nabla Q(w_{t-1} - \alpha \cdot h_{t-1})$$



На практике обычно используют следующие значения гиперпараметров:

- $\alpha = 0.9$
- $\eta = 0.01$

AdaGrad

$$G_{kj} = G_{k-1,j} + (\nabla_w Q(w^{(k-1)}))_j^2;$$
$$w_j^{(k)} = w_j^{(k-1)} - \frac{\eta_t}{\sqrt{G_{kj} + \varepsilon}}(\nabla_w Q(w^{(k-1)}))_j$$

RMSProp

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

На практике обычно используют следующие значения гиперпараметров:

- $\alpha = 0.9$
- $\eta = 0.01$

Adam

Nesterov momentum + RMSProp = Adam