Оптимизация. Градиентный спуск. Ускорение. SGD

Машинное обучение

Александр Безносиков

Московский физико-технический институт

04 октября 2025

Задача оптимизации

$$\min_{w \in \mathbb{R}^d} f(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(g(w, x_i), y_i) \tag{1}$$

Задача оптимизации

$$\min_{w \in \mathbb{R}^d} f(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(g(w, x_i), y_i) \tag{1}$$

Вопрос: что можно сказать про эту задач? сложная ли эта задача?

Задачи оптимизации. Первые наблюдения.

1. В общем случае задачи оптимизации могут не иметь решения. Например, задача $\min\limits_{w \in \mathbb{R}} w$ не имеет решения.
2. Задачи оптимизации часто нельзя решить аналитически.
3. Их сложность зависит от вида целевой функции f, а также от размерности w.

Задачи оптимизации. Первые наблюдения.

1. В общем случае задачи оптимизации могут не иметь решения. Например, задача $\min\limits_{w \in \mathbb{R}} w$ не имеет решения.
2. Задачи оптимизации часто нельзя решить аналитически.
3. Их сложность зависит от вида целевой функции f, а также от размерности w.

Если же задача оптимизации имеет решение, то на практике её обычно решают, вообще говоря, приближённо. Для этого применяются специальные алгоритмы, которые и называют методами оптимизации.

Методы оптимизации

- Нет смысла искать лучший метод для решения конкретной задачи. Например, лучший метод для решений задачи $\min_{w \in \mathbb{R}^d} \|w\|^2$ сходится за 1 итерацию: этот метод просто всегда выдаёт ответ $w^* = 0$. Очевидно, что для других задач такой метод не пригоден.
- Эффективность метода определяется для класса задач, т.к. обычно численные методы разрабатываются для *приближённого* решения множества однотипных задач.
- Метод разрабатывается для класса задач $\Longrightarrow$ метод не может иметь с самого начала полной информации о задаче. Вместо этого метод использует модель задачи, например, формулировку задачи, описание функциональных компонент, множества, на котором происходит оптимизация и т.д.

- Предполагается, что численный метод может накапливать специфическую информацию о задаче при помощи некоторого устройства (программы, процедуры), которое отвечает на последовательные вопросы численного метода.

- Предполагается, что численный метод может накапливать специфическую информацию о задаче при помощи некоторого устройства (программы, процедуры), которое отвечает на последовательные вопросы численного метода.

Вопрос: Какого рода информацию о функции можно запросить у устройства?

- Предполагается, что численный метод может накапливать специфическую информацию о задаче при помощи некоторого устройства (программы, процедуры), которое отвечает на последовательные вопросы численного метода.

Вопрос: Какого рода информацию о функции можно запросить у устройства?

Примеры

- В запрашиваемой точке w возвращает значение целевой функции $f(w)$.
- В запрашиваемой точке возвращает значение функции $f(w)$ и её градиент в данной точке $\nabla f(w) = \left(\frac{\partial f(w)}{\partial w_1}, \ldots, \frac{\partial f(w)}{\partial w_w} \right)^\top$.

Общая итеративная схема метода оптимизации $\mathcal{M}$

Входные данные: начальная точка w^0 (0 – верхний индекс), требуемая точность решения задачи $\varepsilon > 0$.

Общая итеративная схема метода оптимизации $\mathcal{M}$

Входные данные: начальная точка w^0 (0 – верхний индекс), требуемая точность решения задачи $\varepsilon > 0$.
Настройка. Задать $k = 0$ (счётчик итераций) и $I_{-1} = \varnothing$ (накапливаемая информационная модель решаемой задачи).

Общая итеративная схема метода оптимизации $\mathcal{M}$

Входные данные: начальная точка w^0 (0 – верхний индекс), требуемая точность решения задачи $\varepsilon > 0$.
Настройка. Задать $k = 0$ (счётчик итераций) и $I_{-1} = \varnothing$ (накапливаемая информационная модель решаемой задачи).
Основной цикл

1. Задать вопрос в точке w^k к схеме $\mathcal{O}$, который знает некоторую информацию о нашей функции (производная, градиент, etc).

Общая итеративная схема метода оптимизации $\mathcal{M}$

Входные данные: начальная точка w^0 (0 – верхний индекс), требуемая точность решения задачи $\varepsilon > 0$.
Настройка. Задать $k = 0$ (счётчик итераций) и $I_{-1} = \varnothing$ (накапливаемая информационная модель решаемой задачи).
Основной цикл

1. Задать вопрос в точке w^k к схеме $\mathcal{O}$, который знает некоторую информацию о нашей функции (производная, градиент, etc).
2. Пересчитать информационную модель: $I_k = I_{k-1} \cup (w^k, \mathcal{O}(w^k))$.

Общая итеративная схема метода оптимизации $\mathcal{M}$

Входные данные: начальная точка w^0 (0 – верхний индекс), требуемая точность решения задачи $\varepsilon > 0$.
Настройка. Задать $k = 0$ (счётчик итераций) и $I_{-1} = \varnothing$ (накапливаемая информационная модель решаемой задачи).
Основной цикл

1. Задать вопрос в точке w^k к схеме $\mathcal{O}$, который знает некоторую информацию о нашей функции (производная, градиент, etc).
2. Пересчитать информационную модель: $I_k = I_{k-1} \cup (w^k, \mathcal{O}(w^k))$.
3. Применить правило метода $\mathcal{M}$ для получения новой точки w^{k+1} по модели I_k.

Общая итеративная схема метода оптимизации $\mathcal{M}$

Входные данные: начальная точка w^0 (0 – верхний индекс), требуемая точность решения задачи $\varepsilon > 0$.
Настройка. Задать $k = 0$ (счётчик итераций) и $I_{-1} = \varnothing$ (накапливаемая информационная модель решаемой задачи).
Основной цикл

1. Задать вопрос в точке w^k к схеме $\mathcal{O}$, который знает некоторую информацию о нашей функции (производная, градиент, etc).
2. Пересчитать информационную модель: $I_k = I_{k-1} \cup (w^k, \mathcal{O}(w^k))$.
3. Применить правило метода $\mathcal{M}$ для получения новой точки w^{k+1} по модели I_k.
4. Проверить критерий остановки $\mathcal{T}_\varepsilon$. Если критерий выполнен, то выдать ответ $\bar{w}$, иначе положить $k := k + 1$ и вернуться на шаг 1.

Критерии останова

- По аргументу:

$$\|w^k - w^*\| \leq \varepsilon.$$

Вопрос: какие проблемы тут видим?

Критерии останова

- По аргументу:

$$\|w^k - w^*\| \leq \varepsilon.$$

Вопрос: какие проблемы тут видим?

- w^* – неизвестно, но можно так

$$\|w^{k+1} - w^k\| \leq \|w^{k+1} - w^*\| + \|w^k - w^*\| \leq 2\varepsilon.$$

Из $\|w^{k+1} - w^k\| \leq \|w^k - w^*\| \leq \varepsilon/2$, следует $\|w^{k+1} - w^k\| \leq \varepsilon$ (в обратную сторону, очевидно, неверно). $\|w^{k+1} - w^k\| \leq \varepsilon$ – это скорее практический вариант критерия, который работает, если есть понимание (интуиция), что $\|w^k - w^*\| \to 0$.

Критерии останова

- По аргументу:

$$\|w^k - w^*\| \leq \varepsilon.$$

Вопрос: какие проблемы тут видим?

- w^* – неизвестно, но можно так

$$\|w^{k+1} - w^k\| \leq \|w^{k+1} - w^*\| + \|w^k - w^*\| \leq 2\varepsilon.$$

Из $\|w^{k+1} - w^k\| \leq \|w^k - w^*\| \leq \varepsilon/2$, следует $\|w^{k+1} - w^k\| \leq \varepsilon$ (в обратную сторону, очевидно, неверно). $\|w^{k+1} - w^k\| \leq \varepsilon$ – это скорее практический вариант критерия, который работает, если есть понимание (интуиция), что $\|w^k - w^*\| \to 0$.

- w^* – не уникально. Тогда можно поменять критерий

Критерии останова

- По функции:

$$f(w^k) - f^* \leq \varepsilon.$$

Часто f^* известно, например, для $f(w) = \|Aw - b\|^2$. На практике можно использовать $|f(w^k) - f(w^{k+1})|$.

Критерии останова

- По функции:

$$f(w^k) - f^* \leq \varepsilon.$$

  Часто f^* известно, например, для $f(w) = \|Aw - b\|^2$. На практике можно использовать $|f(w^k) - f(w^{k+1})|$.

- По норме градиента:

$$\|\nabla f(w^k)\| \leq \varepsilon.$$

Класс задач минимизации липшицевых функций

$$\min_{w \in B_d} f(w) \tag{2}$$

- $B_d = \{w \in \mathbb{R}^d \mid 0 \leq w_i \leq 1, \quad i = 1, \ldots, d\}$
- Функция $f(w)$ является M-липшицевой на B_d относительно ℓ_∞-нормы:

$$\forall x, y \; |f(x) - f(y)| \leq M\|x - y\|_\infty = M \max_{i=1,\ldots,d} |x_i - y_i|.$$

Класс задач минимизации липшицевых функций

Наблюдение

Множество B_d является ограниченным и замкнутым, т.е. компактом, а из липшицевости функции f следует и её непрерывность, поэтому задача (2) имеет решение, ибо непрерывная на компакте функция достигает своих минимального и максимального значений. Введем обозначение $f^* = \min\limits_{w \in B_d} f(w)$.

- **Класс методов.** Для данной задачи рассмотрим методы нулевого порядка.
- **Цель:** найти $\bar{w} \in B_d$: $f(\bar{w}) - f^* \leq \varepsilon$.

Метод перебора

Рассмотрим один из самых простых способов решения этой задачи — метод равномерного перебора.

Алгоритм Метод равномерного перебора

Вход: целочисленный параметр перебора $p \geq 1$

1: Сформировать $(p+1)^d$ точек вида $w_{(i_1,\ldots,i_d)} = \left(\frac{i_1}{p}, \frac{i_2}{p}, \ldots, \frac{i_d}{p}\right)^\top$, где $(i_1, \ldots, i_d) \in \{0, 1, \ldots, p\}^d$

2: Среди точек $w_{(i_1,\ldots,i_d)}$ найти точку $\bar{w}$ с наименьшим значением целевой функции f.

Выход: $\bar{w}, f(\bar{w})$

Гарантии

Теорема

Алгоритм с параметром p возвращает такую точку $\bar{w}$, что

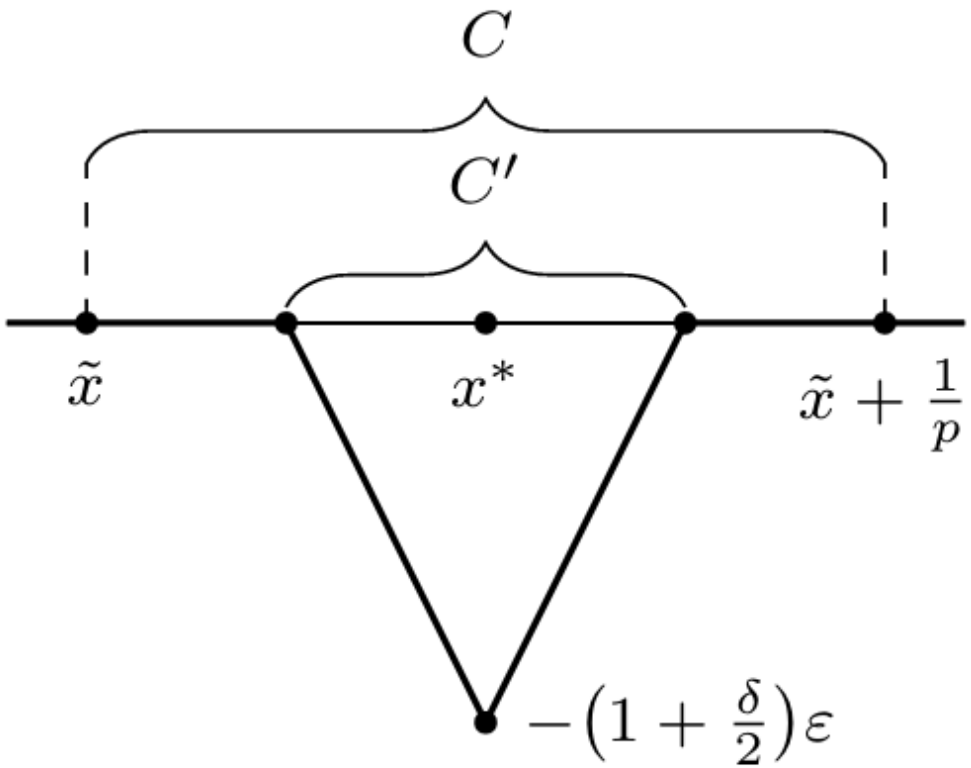
$$f(\bar{w}) - f^* \leq \frac{M}{2p}, \tag{3}$$

откуда следует, что методу равномерного перебора нужно в худшем случае

$$\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2 \right)^d \tag{4}$$

обращений к оракулу, чтобы гарантировать $f(\bar{w}) - f^* \leq \varepsilon$.

Пример функции



Метод перебора: анализ

Вопрос: хороший результат получили или нет?

Метод перебора: анализ

Вопрос: хороший результат получили или нет?

- Предположим $M = 2$, $d = 13$ и $\varepsilon = 0.01$, то есть размерность задачи сравнительно небольшая и точность решения задачи не слишком высокая.

Метод перебора: анализ

Вопрос: хороший результат получили или нет?

- Предположим $M = 2$, $d = 13$ и $\varepsilon = 0.01$, то есть размерность задачи сравнительно небольшая и точность решения задачи не слишком высокая.
- Необходимое число обращений к оракулу: $\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2\right)^d = 102^{13} > 10^{26}$.

Метод перебора: анализ

Вопрос: хороший результат получили или нет?

- Предположим $M = 2$, $d = 13$ и $\varepsilon = 0.01$, то есть размерность задачи сравнительно небольшая и точность решения задачи не слишком высокая.
- Необходимое число обращений к оракулу: $\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2\right)^d = 102^{13} > 10^{26}$.
- Сложность одного вызова оракула не менее 1, но если потребовать, чтобы он обязательно считал все переданные ему точки, то сложность не менее d операции.

Метод перебора: анализ

Вопрос: хороший результат получили или нет?

- Предположим $M = 2$, $d = 13$ и $\varepsilon = 0.01$, то есть размерность задачи сравнительно небольшая и точность решения задачи не слишком высокая.
- Необходимое число обращений к оракулу: $\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2\right)^d = 102^{13} > 10^{26}$.
- Сложность одного вызова оракула не менее 1, но если потребовать, чтобы он обязательно считал все переданные ему точки, то сложность не менее d операции.
- Производительность компьютера: 10^{11} арифметических операций в секунду.

Метод перебора: анализ

Вопрос: хороший результат получили или нет?

- Предположим $M = 2$, $d = 13$ и $\varepsilon = 0.01$, то есть размерность задачи сравнительно небольшая и точность решения задачи не слишком высокая.
- Необходимое число обращений к оракулу: $\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor + 2\right)^d = 102^{13} > 10^{26}$.
- Сложность одного вызова оракула не менее 1, но если потребовать, чтобы он обязательно считал все переданные ему точки, то сложность не менее d операции.
- Производительность компьютера: 10^{11} арифметических операций в секунду.
- Общее время: хотя бы 10^{15} секунд, что больше 30 миллионов лет.

Верхние и нижние оценки

- **Вопрос:** что мы сейчас получили? верхнюю или нижнюю оценку? что такое верхняя оценка?

Верхние и нижние оценки

- **Вопрос:** что мы сейчас получили? верхнюю или нижнюю оценку? что такое верхняя оценка?
- **Верхняя оценка** – гарантии нахождения решения определенным методом из рассматриваемого класса методов (например, методы с оракулом нулевого порядка) для любой задачи из класса (Липшецева целевая функция на кубе).
- **Нижняя оценка** – гарантия, что для любого метода из класса существует «плохая» задача из класса такая, что метод будет сходиться не лучше, чем утверждает нижняя оценка.
- Возникает вопрос: может мы плохо вывели верхнюю оценку (неидеальный анализ), может ли предложить другой метод из рассматриваемого класса, который будет находить приближённое решение существенно быстрее? На этот вопрос и даст ответ нижняя оценка.

Нижняя оценка

Теорема

Пусть $\varepsilon < \frac{M}{2}$. Тогда аналитическая сложность описанного класса задач, т.е. аналитическая сложность метода на «худшей» для него задаче из данного класса, составляет по крайней мере

$$\left(\left\lfloor \frac{M}{2\varepsilon} \right\rfloor \right)^d \quad \text{вызовов оракула.} \tag{5}$$

Итак, в указанном классе у любого метода оценки на скорость сходимости весьма пессимистичные. Возникает вопрос: какие свойства нужно потребовать от класса оптимизируемых функций, чтобы оценки стали более оптимистичными?

Сильная выпуклость и гладкость: определения

Определение μ-сильно выпуклой функции

Пусть дана непрерывно дифференцируемая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что она является μ-сильно выпуклой ($\mu > 0$), если для любых $x, y \in \mathbb{R}^d$ выполнено

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2.$$

Сильная выпуклость и гладкость: определения

Определение μ-сильно выпуклой функции

Пусть дана непрерывно дифференцируемая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что она является μ-сильно выпуклой ($\mu > 0$), если для любых $x, y \in \mathbb{R}^d$ выполнено

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|_2^2.$$

Определение L-гладкой функции

Пусть дана непрерывно дифференцируемая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что данная функция имеет L-Липшицев градиент (т.е является L-гладкой), если для любых $x, y \in \mathbb{R}^d$ выполнено

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2.$$

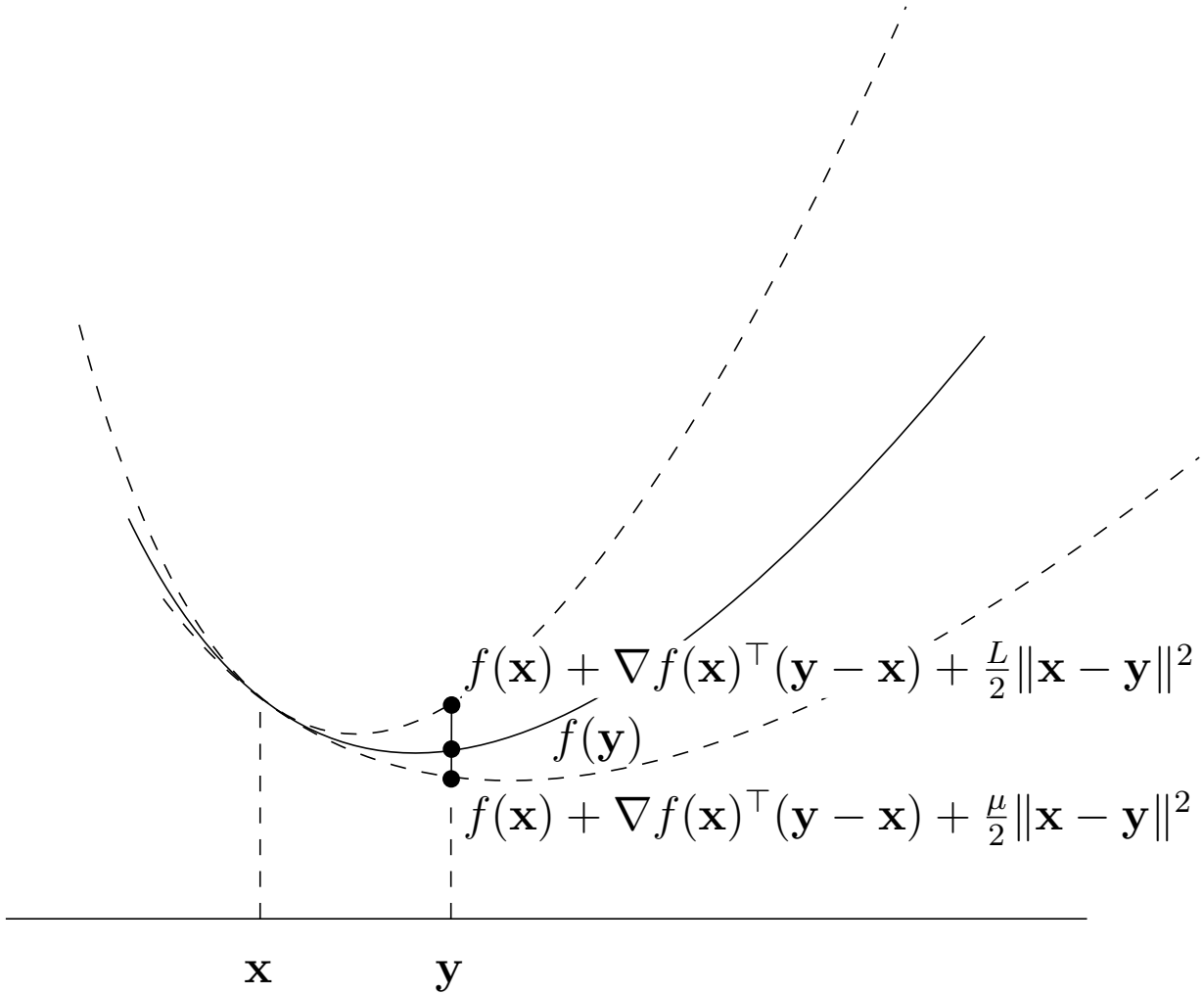
Гладкость: свойство

Теорема (свойство L - гладкой функции)

Пусть дана L - гладкая функция $f : \mathbb{R}^d \to \mathbb{R}$. Тогда для любых $x, y \in \mathbb{R}^d$ выполнено

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|x - y\|_2^2.$$

Гладкость: физический смысл



Градиентный спуск

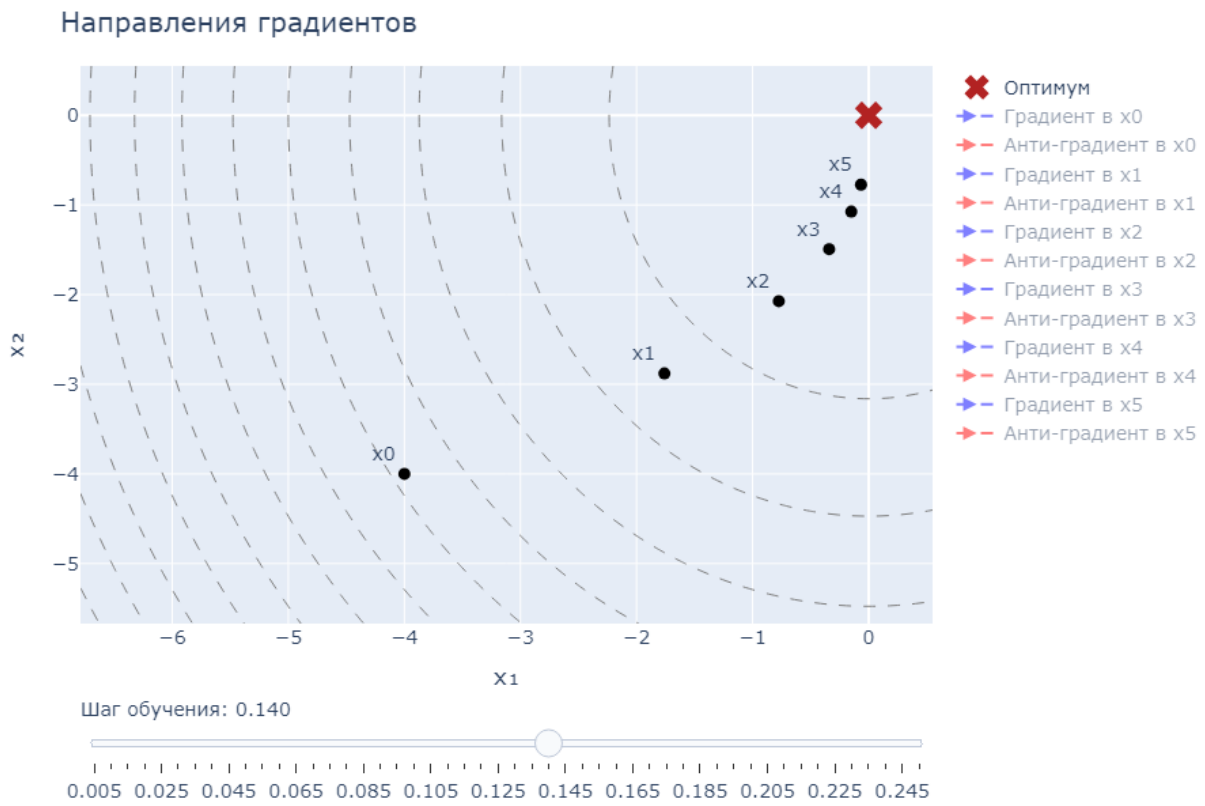
Алгоритм Градиентный спуск

Вход: размеры шагов $\{\gamma_k\}_{k=0} > 0$, стартовая точка $w^0 \in \mathbb{R}^d$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $\nabla f(w^k)$
3: $w^{k+1} = w^k - \gamma_k \nabla f(w^k)$
4: **end for**

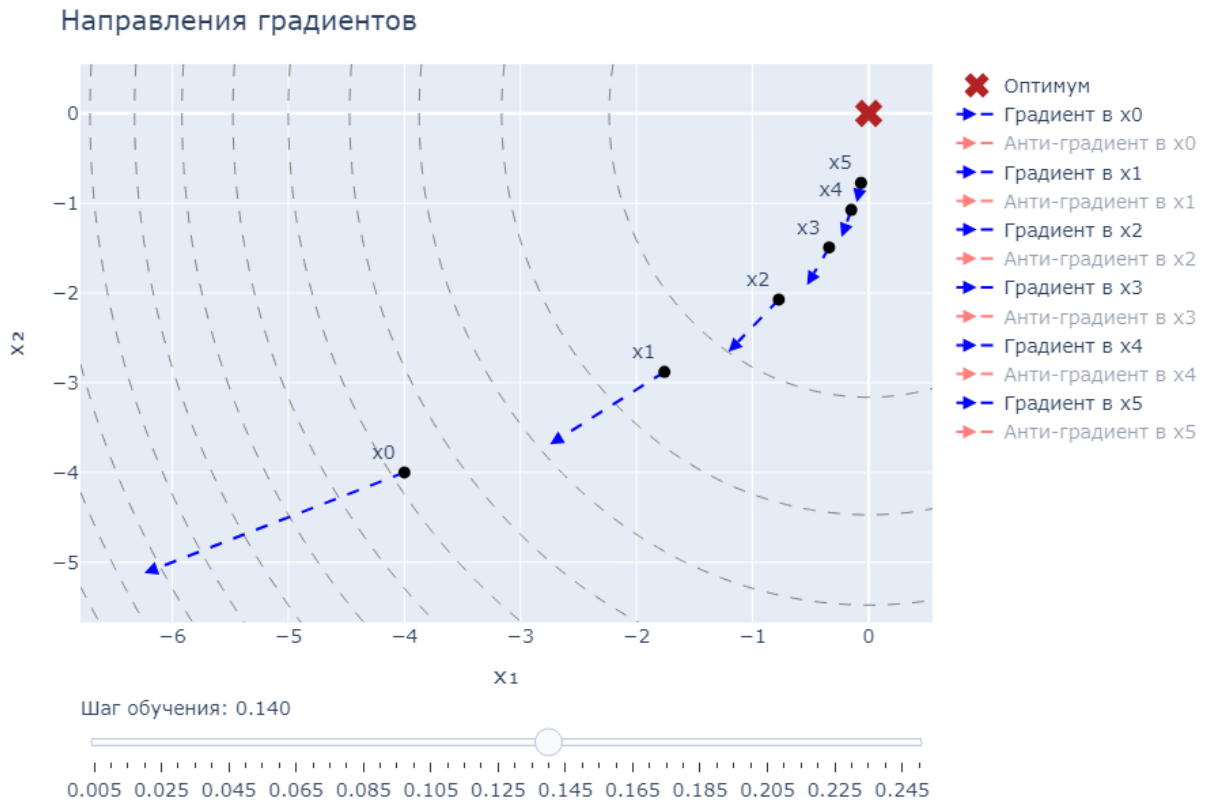
Выход: w^K

Пример

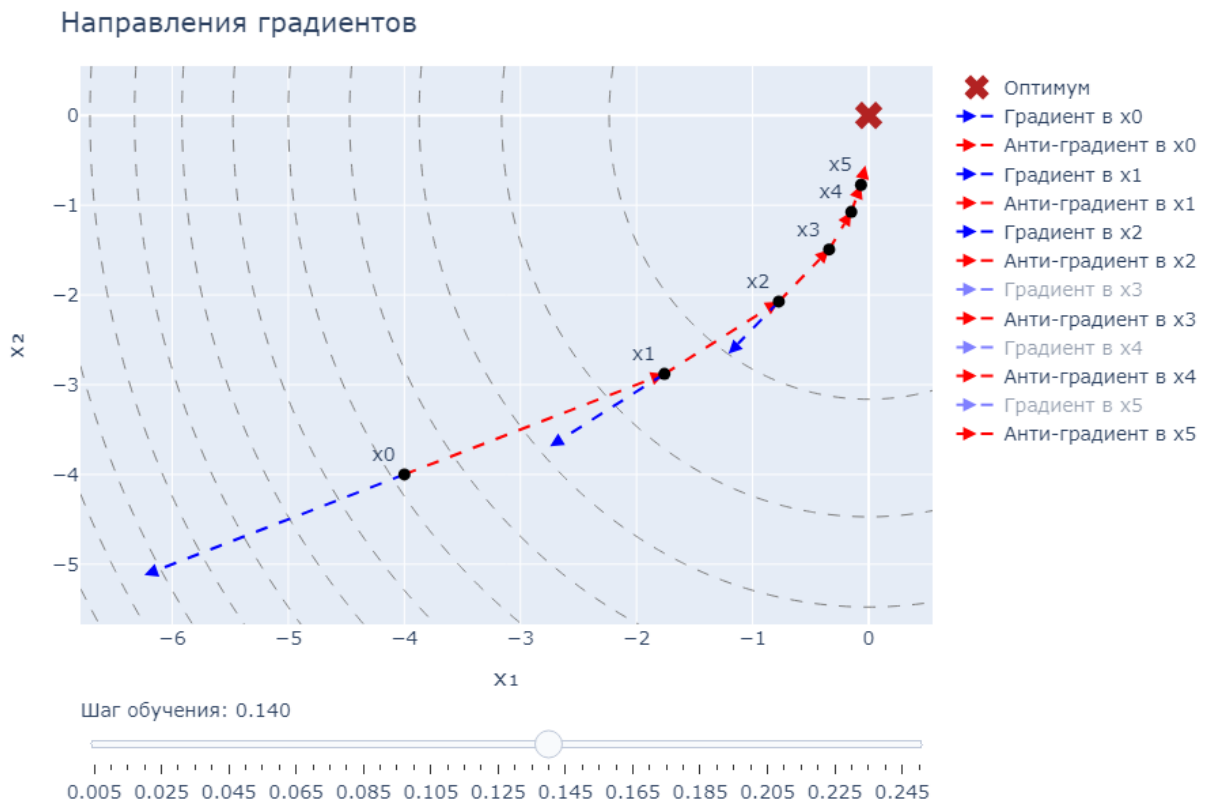


Вопрос: куда направлен градиент в точке x_1? x_2?

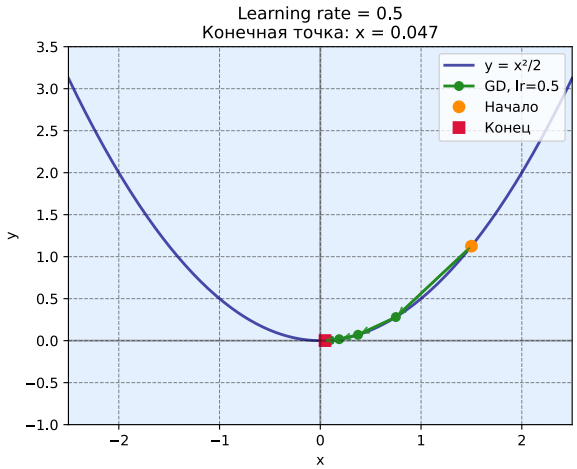
Пример

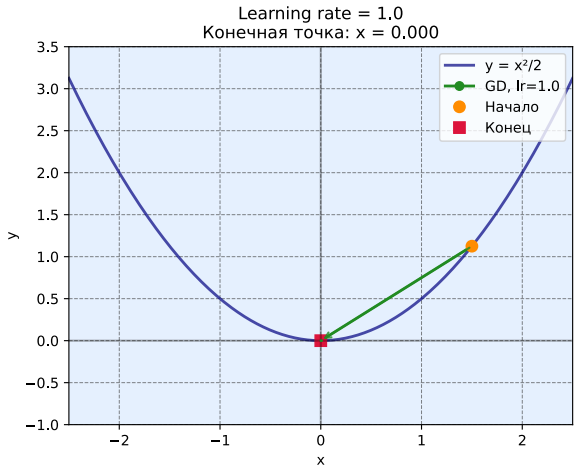

Направления градиентов
Оптимум
Градиент в x0
Анти-градиент в x0
Градиент в x1
Анти-градиент в x1
Градиент в x2
Анти-градиент в x2
Градиент в x3
Анти-градиент в x3
Градиент в x4
Анти-градиент в x4
Градиент в x5
Анти-градиент в x5
x0
x1
x2
x3
x4
x5
x1
x2
Шаг обучения: 0.140
0.005 0.025 0.045 0.065 0.085 0.105 0.125 0.145 0.165 0.185 0.205 0.225 0.245

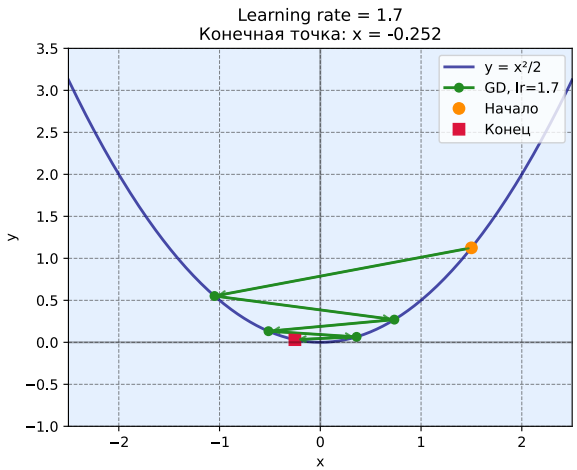
Пример

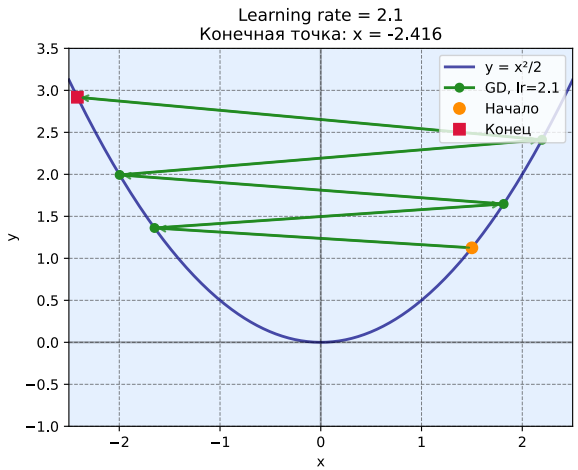

Направления градиентов
Оптимум
Градиент в x0
Анти-градиент в x0
Градиент в x1
Анти-градиент в x1
Градиент в x2
Анти-градиент в x2
Градиент в x3
Анти-градиент в x3
Градиент в x4
Анти-градиент в x4
Градиент в x5
Анти-градиент в x5
x0
x1
x2
x3
x4
x5
x1
x2
Шаг обучения: 0.140
0.005 0.025 0.045 0.065 0.085 0.105 0.125 0.145 0.165 0.185 0.205 0.225 0.245

Зависимость от шага

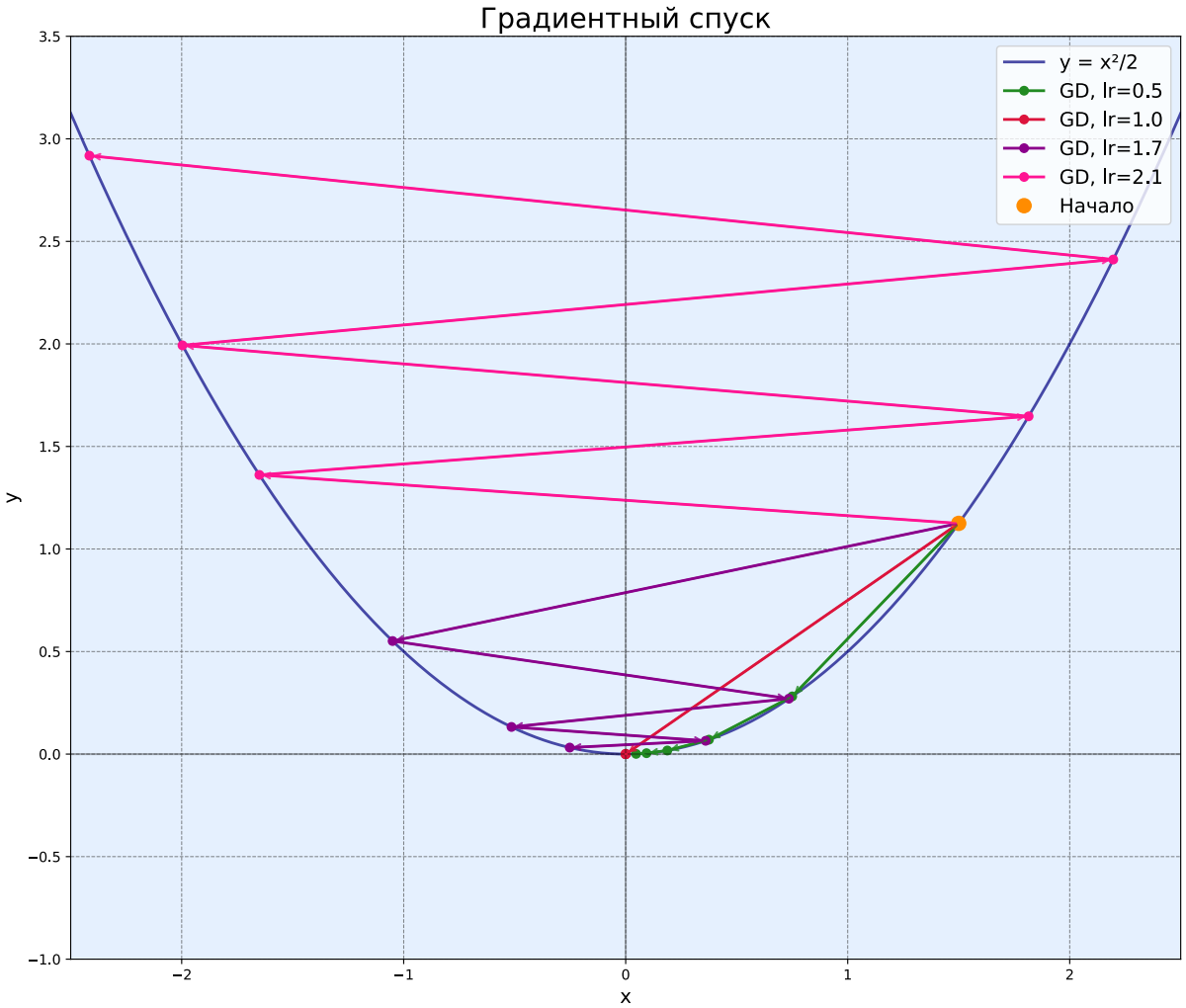








Зависимость от шага



Сходимость

Теорема: сходимость градиентного спуска для L-гладких и μ-сильно выпуклых функций

Пусть задача безусловной оптимизации с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью градиентного спуска с $\gamma_k \leq \frac{1}{L}$. Тогда справедлива следующая оценка сходимости

$$\|w^{k+1} - w^*\|^2 \leq (1 - \gamma_k \mu)\|w^k - w^*\|^2.$$

Сходимость GD: анализ

- Возьмем постоянный шаг $\gamma_k \equiv \gamma = \frac{1}{L}$, тогда

$$\begin{aligned}\|w^k - w^*\|^2 \leq & (1 - \gamma\mu)\|w^{k-1} - w^*\|^2 \\ \leq & (1 - \gamma\mu)^2\|w^{k-2} - w^*\|^2 \\ \leq & \dots \\ \leq & (1 - \gamma\mu)^k\|w^0 - w^*\|^2 \\ = & \left(1 - \frac{\mu}{L}\right)^k \|w^0 - w^*\|^2\end{aligned}$$

Сходимость GD: анализ

- Возьмем постоянный шаг $\gamma_k \equiv \gamma = \frac{1}{L}$, тогда

$$\begin{aligned}\|w^k - w^*\|^2 \leq & (1 - \gamma\mu)\|w^{k-1} - w^*\|^2 \\ \leq & (1 - \gamma\mu)^2\|w^{k-2} - w^*\|^2 \\ \leq & \dots \\ \leq & (1 - \gamma\mu)^k\|w^0 - w^*\|^2 \\ = & \left(1 - \frac{\mu}{L}\right)^k \|w^0 - w^*\|^2\end{aligned}$$

- Получилась линейная сходимость (скорость геометрической прогрессии) к решению.

Подбор шага: Поляк-Шор

Согласно теореме об оценке сходимости L-гладких и μ-сильно выпуклых функций, верно $\|w^{k+1} - w^*\|^2 \leq (1 - \gamma_k \mu)\|w^k - w^*\|^2$. $1 - \gamma_k \mu < 1$, откуда верно $\|w^{k+1} - w^*\|^2 \leq \|w^k - w^*\|^2$. В доказательстве градиентного спуска получаем следующее:

$$\|w^{k+1} - w^*\|_2^2 \leq \|w^k - w^*\|_2^2 - 2\gamma_k \left(f(w^k) - f(w^*)\right) + \gamma_k^2 \|\nabla f(w^k)\|_2^2$$

Подбор шага: Поляк-Шор

Согласно теореме об оценке сходимости L-гладких и μ-сильно выпуклых функций, верно $\|w^{k+1} - w^*\|^2 \leq (1 - \gamma_k \mu)\|w^k - w^*\|^2$. $1 - \gamma_k \mu < 1$, откуда верно $\|w^{k+1} - w^*\|^2 \leq \|w^k - w^*\|^2$. В доказательстве градиентного спуска получаем следующее:

$$\|w^{k+1} - w^*\|_2^2 \leq \|w^k - w^*\|_2^2 - 2\gamma_k \left(f(w^k) - f(w^*)\right) + \gamma_k^2 \|\nabla f(w^k)\|_2^2$$

Вопрос: как можно подобрать γ_k оптимально в этой ситуации?

Подбор шага: Поляк-Шор

Согласно теореме об оценке сходимости L-гладких и μ-сильно выпуклых функций, верно $\|w^{k+1} - w^*\|^2 \leq (1 - \gamma_k \mu)\|w^k - w^*\|^2$. $1 - \gamma_k \mu < 1$, откуда верно $\|w^{k+1} - w^*\|^2 \leq \|w^k - w^*\|^2$. В доказательстве градиентного спуска получаем следующее:

$$\|w^{k+1} - w^*\|_2^2 \leq \|w^k - w^*\|_2^2 - 2\gamma_k \left(f(w^k) - f(w^*)\right) + \gamma_k^2 \|\nabla f(w^k)\|_2^2$$

Вопрос: как можно подобрать γ_k оптимально в этой ситуации? $\arg\min_{\gamma_k} \left(-2\gamma_k \left(f(w^k) - f(w^*)\right) + \gamma_k^2 \|\nabla f(w^k)\|_2^2\right)$?

Подбор шага: Поляк-Шор

Согласно теореме об оценке сходимости L-гладких и μ-сильно выпуклых функций, верно $\|w^{k+1} - w^*\|^2 \leq (1 - \gamma_k \mu)\|w^k - w^*\|^2$. $1 - \gamma_k \mu < 1$, откуда верно $\|w^{k+1} - w^*\|^2 \leq \|w^k - w^*\|^2$. В доказательстве градиентного спуска получаем следующее:

$$\|w^{k+1} - w^*\|_2^2 \leq \|w^k - w^*\|_2^2 - 2\gamma_k \left(f(w^k) - f(w^*)\right) + \gamma_k^2 \|\nabla f(w^k)\|_2^2$$

Вопрос: как можно подобрать γ_k оптимально в этой ситуации? $\arg\min_{\gamma_k} \left(-2\gamma_k \left(f(w^k) - f(w^*)\right) + \gamma_k^2 \|\nabla f(w^k)\|_2^2\right)$?

$$\gamma_k = \frac{f(w^k) - f(w^*)}{\|\nabla f(w^k)\|_2^2}$$

Вопрос: какие видите проблемы?

Подбор шага: Поляк-Шор

Согласно теореме об оценке сходимости L-гладких и μ-сильно выпуклых функций, верно $\|w^{k+1} - w^*\|^2 \leq (1 - \gamma_k \mu)\|w^k - w^*\|^2$. $1 - \gamma_k \mu < 1$, откуда верно $\|w^{k+1} - w^*\|^2 \leq \|w^k - w^*\|^2$. В доказательстве градиентного спуска получаем следующее:

$$\|w^{k+1} - w^*\|_2^2 \leq \|w^k - w^*\|_2^2 - 2\gamma_k \left(f(w^k) - f(w^*)\right) + \gamma_k^2 \|\nabla f(w^k)\|_2^2$$

Вопрос: как можно подобрать γ_k оптимально в этой ситуации? $\arg\min_{\gamma_k} \left(-2\gamma_k \left(f(w^k) - f(w^*)\right) + \gamma_k^2 \|\nabla f(w^k)\|_2^2\right)$?

$$\gamma_k = \frac{f(w^k) - f(w^*)}{\|\nabla f(w^k)\|_2^2}$$

Вопрос: какие видите проблемы? $f(w^*)$ – иногда известно, а иногда можно оценить.

Подбор шага

Шаг Поляка-Шора:

$$\gamma_k = \frac{f(w^k) - f(w^*)}{\alpha \|\nabla f(w^k)\|_2^2}, \quad \alpha \geq 1 \quad \text{(надо подбирать)}$$

Верхняя оценка

Теорема: сходимость градиентного спуска для L-гладких и μ-сильно выпуклых функций

Пусть задача безусловной оптимизации с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью градиентного спуска с $\gamma_k \leq \frac{1}{L}$. Тогда, чтобы добиться точности ε по аргументу ($\|x^k - x^*\|_2 \leq \varepsilon$), необходимо

$$K = O\left(\frac{L}{\mu} \log \frac{\|x^0 - x^*\|_2}{\varepsilon}\right) = \tilde{O}\left(\frac{L}{\mu}\right) \text{ итераций.}$$

Верхняя оценка

Теорема: сходимость градиентного спуска для L-гладких и μ-сильно выпуклых функций

Пусть задача безусловной оптимизации с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью градиентного спуска с $\gamma_k \leq \frac{1}{L}$. Тогда, чтобы добиться точности ε по аргументу ($\|x^k - x^*\|_2 \leq \varepsilon$), необходимо

$$K = O\left(\frac{L}{\mu} \log \frac{\|x^0 - x^*\|_2}{\varepsilon}\right) = \tilde{O}\left(\frac{L}{\mu}\right) \text{ итераций.}$$

Вопрос: а можно ли лучше?

Метод тяжелого шарика

- Б.Т. Поляк в 1964 году предложил метод тяжелого шарика.

Алгоритм Метод тяжелого шарика

Вход: размер шагов $\{\gamma_k\}_{k=0} > 0$, моментумы $\{\tau_k\}_{k=0} \in [0; 1]$, стартовая точка $w^0 = w^{-1} \in \mathbb{R}^d$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $\nabla f(w^k)$
3: $w^{k+1} = w^k - \gamma_k \nabla f(w^k) + \tau_k (w^k - w^{k-1})$
4: **end for**

Выход: w^K

Метод тяжелого шарика

- Б.Т. Поляк в 1964 году предложил метод тяжелого шарика.

Алгоритм Метод тяжелого шарика

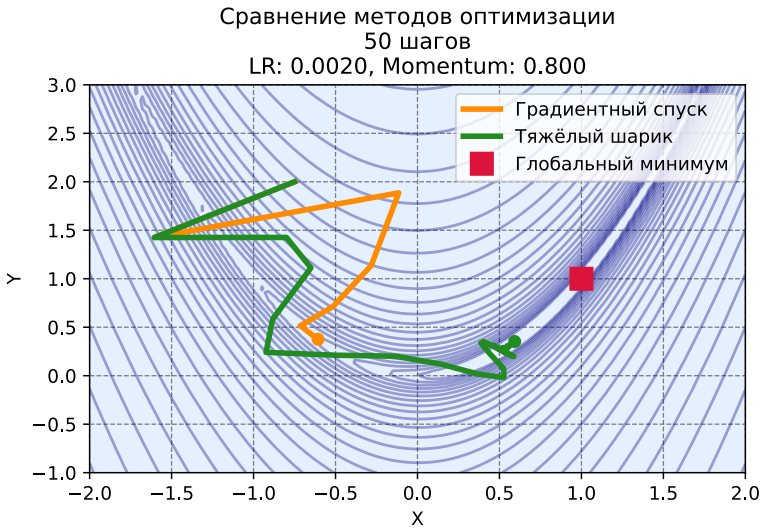
Вход: размер шагов $\{\gamma_k\}_{k=0} > 0$, моментумы $\{\tau_k\}_{k=0} \in [0; 1]$, стартовая точка $w^0 = w^{-1} \in \mathbb{R}^d$, количество итераций K

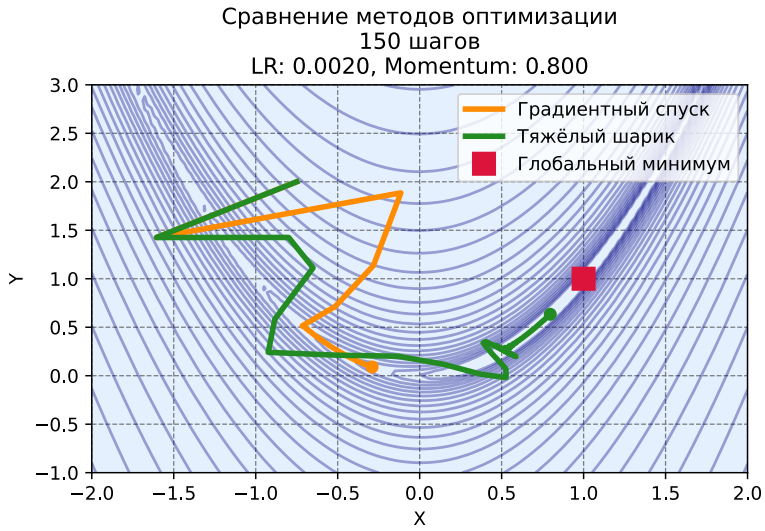
1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $\nabla f(w^k)$
3: $w^{k+1} = w^k - \gamma_k \nabla f(w^k) + \tau_k (w^k - w^{k-1})$
4: **end for**

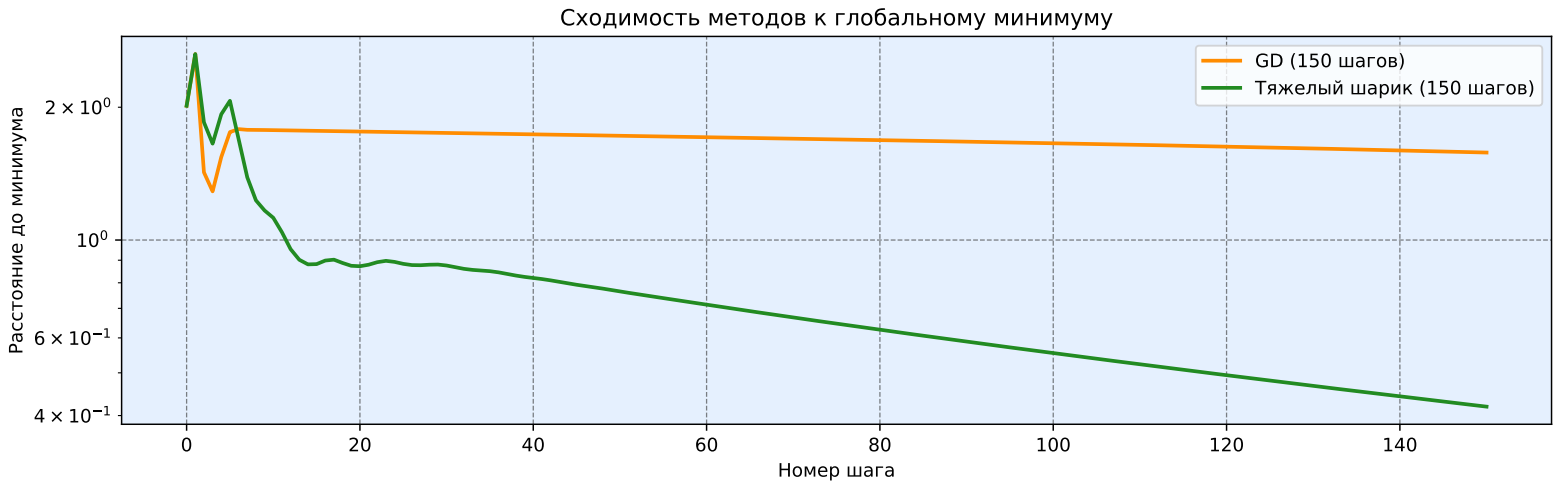
Выход: w^K

- Добавим к градиентному спуску моментумный член — предположим, что у точки, отвечающей за текущее положение значение w^k есть инерция.

Сравнение тяжелого шарика и градиентного спуска







На чем держится ML

- В библиотеке pytorch (основная библиотека Deep Learning) реализован следующий метод:

$$v^{k+1} = \beta v^k + \nabla f(w^k) \quad \beta \in [0; 1)$$

$$w^{k+1} = w^k - \gamma v^{k+1}$$

На чем держится ML

- В библиотеке pytorch (основная библиотека Deep Learning) реализован следующий метод:

$$v^{k+1} = \beta v^k + \nabla f(w^k) \quad \beta \in [0; 1)$$

$$w^{k+1} = w^k - \gamma v^{k+1}$$

Вопрос: как это метод связан с методом тяжелого шарика?

На чем держится ML

- В библиотеке pytorch (основная библиотека Deep Learning) реализован следующий метод:

$$v^{k+1} = \beta v^k + \nabla f(w^k) \quad \beta \in [0; 1)$$

$$w^{k+1} = w^k - \gamma v^{k+1}$$

Вопрос: как это метод связан с методом тяжелого шарика? Это практически он и есть. Поставим первую строку во вторую:

$$w^{k+1} = w^k - \gamma \nabla f(w^k) - \gamma \beta v^k$$

На чем держится ML

- В библиотеке pytorch (основная библиотека Deep Learning) реализован следующий метод:

$$v^{k+1} = \beta v^k + \nabla f(w^k) \quad \beta \in [0; 1)$$

$$w^{k+1} = w^k - \gamma v^{k+1}$$

Вопрос: как это метод связан с методом тяжелого шарика? Это практически он и есть. Поставим первую строку во вторую:

$$w^{k+1} = w^k - \gamma \nabla f(w^k) - \gamma \beta v^k$$

Из второй строки для k шага:

$$-\gamma v^k = w^k - w^{k-1}$$

На чем держится ML

- В библиотеке pytorch (основная библиотека Deep Learning) реализован следующий метод:

$$v^{k+1} = \beta v^k + \nabla f(w^k) \quad \beta \in [0; 1)$$

$$w^{k+1} = w^k - \gamma v^{k+1}$$

Вопрос: как это метод связан с методом тяжелого шарика? Это практически он и есть. Поставим первую строку во вторую:

$$w^{k+1} = w^k - \gamma \nabla f(w^k) - \gamma \beta v^k$$

Из второй строки для k шага:

$$-\gamma v^k = w^k - w^{k-1}$$

Тогда подставим в предыдущие и получим

$$w^{k+1} = w^k - \gamma \nabla f(w^k) + \beta(w^k - w^{k-1})$$

Это показывает еще одну физику метода тяжелого шарика – мы идем по аккумулированному градиенту (старые забываются).

Плюсы и минусы

Вопрос: какие плюсы и минусы видите у метода тяжелого шарика?

Плюсы и минусы

Вопрос: какие плюсы и минусы видите у метода тяжелого шарика?

Плюсы

- Понятная физика и интуиция.
- Легкость в имплантации.
- Дешевизна вычислений.

Минусы

- Нужно подбирать теперь 2 параметра. Мы сейчас умеем только в теории оценивать γ_k. Теперь что-то нужно делать с τ_k... Типично τ_k берут близким к единице или устремляют к единице.
- Мы шли за ускорением градиентного спуска. А оно вообще есть в общем случае?

Плюсы и минусы

Вопрос: какие плюсы и минусы видите у метода тяжелого шарика?

Плюсы

- Понятная физика и интуиция.
- Легкость в имплантации.
- Дешевизна вычислений.

Минусы

- Нужно подбирать теперь 2 параметра. Мы сейчас умеем только в теории оценивать γ_k. Теперь что-то нужно делать с τ_k... Типично τ_k берут близким к единице или устремляют к единице.
- Мы шли за ускорением градиентного спуска. А оно вообще есть в общем случае? Нет...

Ускоренный градиентный метод

- Ю.Е. Нестеров в 1983 году предложил ускоренный градиентный метод.

Алгоритм Ускоренный градиентный метод

Вход: размер шагов $\{\gamma_k\}_{k=0} > 0$, моментумы $\{\tau_k\}_{k=0} \in [0; 1]$, стартовая точка $w^0 = y^0 \in \mathbb{R}^d$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $\nabla f(y^k)$
3: $w^{k+1} = y^k - \gamma_k \nabla f(y^k)$
4: $y^{k+1} = w^{k+1} + \tau_k(w^{k+1} - w^k)$
5: **end for**

Выход: w^K

Ускоренный градиентный метод и тяжелый шарик

- **Вопрос:**В чем ключевое отличие метода Нестерова от тяжелого шарика?
  Тяжелый шарик:

$$w^{k+1} = w^k - \gamma_k \nabla f(w^k) + \tau_k (w^k - w^{k-1})$$

  Ускоренный градиентный метод:

$$w^{k+1} = y^k - \gamma_k \nabla f(y^k)$$

$$y^{k+1} = w^{k+1} + \tau_k (w^{k+1} - w^k)$$

Ускоренный градиентный метод и тяжелый шарик

- **Вопрос:**В чем ключевое отличие метода Нестерова от тяжелого шарика?
  Тяжелый шарик:

$$w^{k+1} = w^k - \gamma_k \nabla f(w^k) + \tau_k (w^k - w^{k-1})$$

  Ускоренный градиентный метод:

$$w^{k+1} = y^k - \gamma_k \nabla f(y^k)$$
$$y^{k+1} = w^{k+1} + \tau_k (w^{k+1} - w^k)$$

- Перепишем ускоренный градиентный метод:

$$w^{k+1} = w^k + \tau_k (w^k - w^{k-1}) - \gamma_k \nabla f(w^k + \tau_k (w^k - w^{k-1})).$$

  Моментум в точке подсчета градиента/«взгляд вперед»/экстраполяция

Сходимость

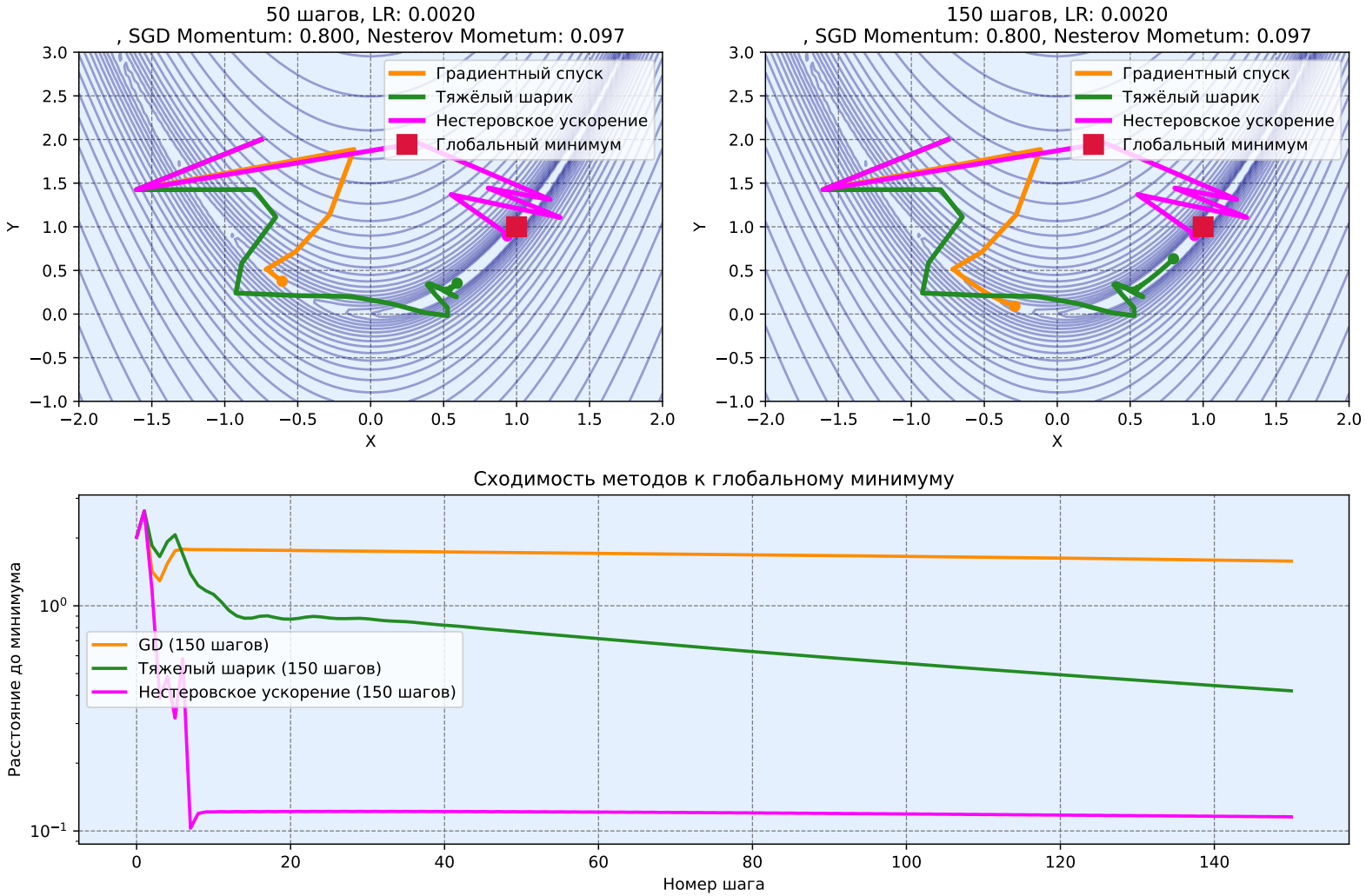
Теорема: сходимость ускоренного градиентного метода

Пусть задача безусловной оптимизации с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью ускоренного градиентного метода. Тогда при $\gamma_k = \frac{1}{L}$ и $\tau_k = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$, справедлива следующая оценка сходимости:

$$f(x^K) - f^* \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^K \cdot L\|x^0 - x^*\|_2^2.$$

Вопрос: А какой была оценка сходимости у GD?

Сравнение ускоренных методов



Плюсы и минусы

Вопрос: какие плюсы и минусы видите у метода Нестерова?

Плюсы и минусы

Вопрос: какие плюсы и минусы видите у метода Нестерова?

Плюсы

- В большом наборе постановок (гладкая функция, сильно выпуклая функция) и практических случаев сходится быстрее, чем методы градиентного спуска и тяжёлого шарика.

Минусы

- Во многих задачах машинного обучения и проще, и эффективнее использовать стохастические методы (SGD, Adam).

Стохастическая оптимизация: постановка

Вспомним, что f имеет следующий вид:

$$\min_{w \in \mathbb{R}^d} f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(g(w, x_i), y_i).$$

Этот вид оптимизационной задачи также называется *оффлайн* постановкой машинного обучения.

Вопрос. Зачем разбивать функцию f на сумму функций f_i?

- Если n велико (т.е. данных очень много), то искать полный градиент вычислительно затратно, поэтому в большинстве случаев используется градиент по какому-то набору сэмплов.
- Борьба с переобучением. Использование полных градиентов ведет к быстрой минимизации f на обучающей выборке, что в свою очередь приводит к переобучению (обсуждалось на предыдущей лекции).

Стохастический градиентный спуск

- Простая идея – модифицировать градиентный спуск и посмотреть, что будет.

Алгоритм Стохастический градиентный спуск (SGD)

Вход: размеры шагов $\{\gamma_k\}_{k=0} > 0$, стартовая точка $w^0 \in \mathbb{R}^d$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Сгенерировать независимо $\xi^k = (x_i, y_i)$, где i генерируется независимо и равномерно из $[n]$
3: Вычислить стохастический градиент $\nabla f(w^k, \xi^k)$
4: $w^{k+1} = w^k - \gamma_k \nabla f(w^k, \xi^k)$
5: **end for**

Выход: w^K

Сходимость SGD

Теорема сходимость SGD в случае ограниченной дисперсии

Пусть задача безусловной стохастической оптимизации с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью SGD с $\gamma_k \leq \frac{1}{L}$ в условиях несмещенности и ограниченности дисперсии стохастического градиента. Тогда справедлива следующая оценка сходимости

$$\mathbb{E}\left[\|w^{k+1} - w^*\|^2\right] \leq (1 - \gamma_k \mu)\mathbb{E}\left[\|w^k - w^*\|^2\right] + \gamma_k^2 \sigma^2.$$

Сходимость SGD: анализ

- Возьмем постоянный шаг $\gamma_k \equiv \gamma$, тогда

$$\begin{aligned}
\mathbb{E}\left[\|w^k - w^*\|^2\right] \leq & (1-\gamma\mu)\mathbb{E}\left[\|w^{k-1} - w^*\|^2\right] + \gamma^2\sigma^2 \\
\leq & (1-\gamma\mu)^2\mathbb{E}\left[\|w^{k-2} - w^*\|^2\right] \\
& + (1-\gamma\mu)\gamma^2\sigma^2 + \gamma^2\sigma^2 \\
\leq & \dots \\
\leq & (1-\gamma\mu)^k\mathbb{E}\left[\|w^0 - w^*\|^2\right] + \gamma^2\sigma^2\sum_{i=0}^{k-1}(1-\gamma\mu)^i.
\end{aligned}$$

Сходимость SGD: анализ

- Возьмем постоянный шаг $\gamma_k \equiv \gamma$, тогда

$$\begin{aligned}
\mathbb{E}\left[\|w^k - w^*\|^2\right] \leq & (1-\gamma\mu)\mathbb{E}\left[\|w^{k-1} - w^*\|^2\right] + \gamma^2\sigma^2 \\
\leq & (1-\gamma\mu)^2\mathbb{E}\left[\|w^{k-2} - w^*\|^2\right] \\
& + (1-\gamma\mu)\gamma^2\sigma^2 + \gamma^2\sigma^2 \\
\leq & \dots \\
\leq & (1-\gamma\mu)^k\mathbb{E}\left[\|w^0 - w^*\|^2\right] + \gamma^2\sigma^2\sum_{i=0}^{k-1}(1-\gamma\mu)^i.
\end{aligned}$$

- **Вопрос:** как оценить второе слагаемое?

Сходимость SGD: анализ

- Возьмем постоянный шаг $\gamma_k \equiv \gamma$, тогда

$$\begin{aligned}\mathbb{E}\left[\|w^k - w^*\|^2\right] \leq & (1-\gamma\mu)\mathbb{E}\left[\|w^{k-1} - w^*\|^2\right] + \gamma^2\sigma^2 \\ \leq & (1-\gamma\mu)^2\mathbb{E}\left[\|w^{k-2} - w^*\|^2\right] \\ & + (1-\gamma\mu)\gamma^2\sigma^2 + \gamma^2\sigma^2 \\ \leq & \dots \\ \leq & (1-\gamma\mu)^k\mathbb{E}\left[\|w^0 - w^*\|^2\right] + \gamma^2\sigma^2\sum_{i=0}^{k-1}(1-\gamma\mu)^i.\end{aligned}$$

- **Вопрос:** как оценить второе слагаемое? Геометрическая прогрессия: $\sum_{i=0}^{k-1}(1-\gamma\mu)^i \leq \sum_{i=0}^{+\infty}(1-\gamma\mu)^i = \frac{1}{\gamma\mu}$:

$$\mathbb{E}\left[\|w^k - w^*\|^2\right] \leq (1-\gamma\mu)^k\mathbb{E}\left[\|w^0 - w^*\|^2\right] + \frac{\gamma\sigma^2}{\mu}.$$

Сходимость SGD: анализ

- Результат вида:

$$\mathbb{E}\left[\|w^k - w^*\|^2\right] \leq (1-\gamma\mu)^k \mathbb{E}\left[\|w^0 - w^*\|^2\right] + \frac{\gamma\sigma^2}{\mu},$$

  похож на то, что мы уже видели для градиентного спуска.
- Первый член – линейная сходимость к решению

Сходимость SGD: анализ

- Результат вида:

$$\mathbb{E}\left[\|w^k - w^*\|^2\right] \leq (1-\gamma\mu)^k \mathbb{E}\left[\|w^0 - w^*\|^2\right] + \frac{\gamma\sigma^2}{\mu},$$

  похож на то, что мы уже видели для градиентного спуска.
- Первый член – линейная сходимость к решению
- Второй член – говорит о том, что некоторую точность (зависящую от γ, σ и μ) метод преодолеть не может и начинает осциллировать, больше не приближаясь к решению.

Сходимость SGD: проблема сходимости

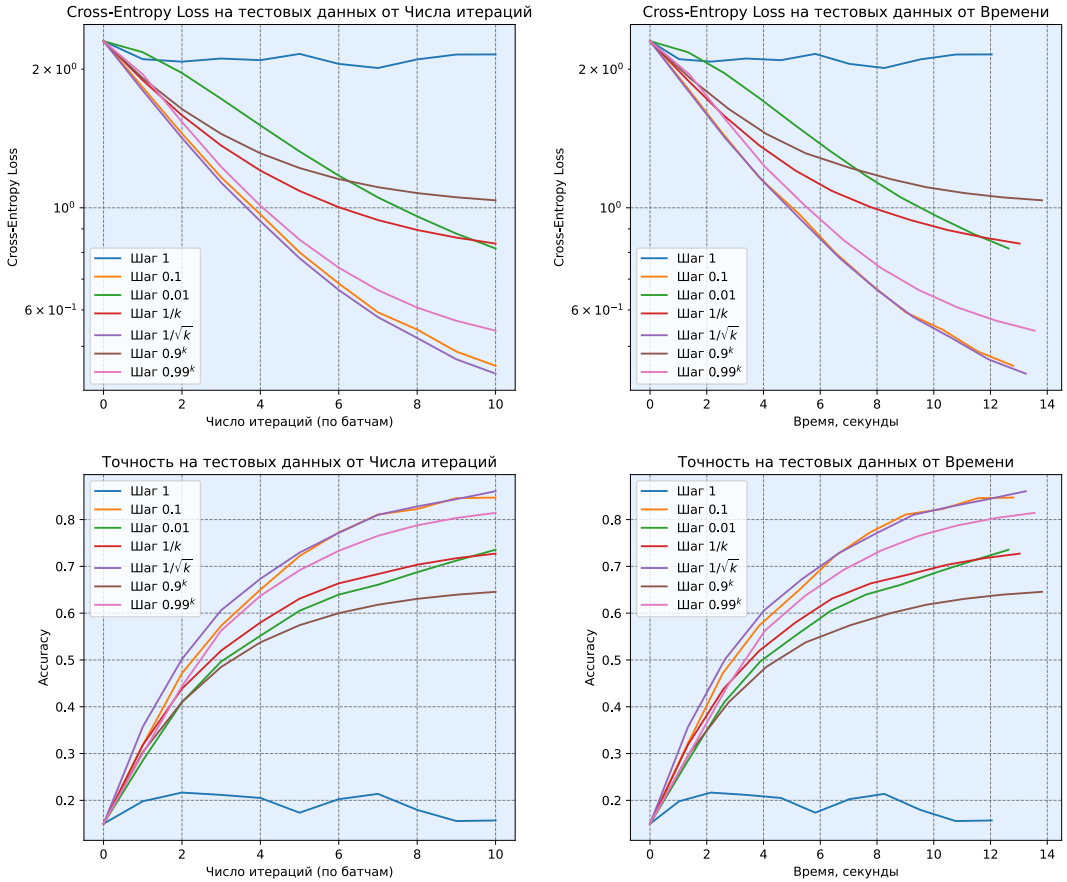
Как можно попробовать решить проблемы неточной сходимости?

Сходимость SGD: проблема сходимости

Как можно попробовать решить проблемы неточной сходимости?

- Уменьшить шаг. Например, брать $\gamma_k = \frac{1}{k+1}$ или $\gamma_k = \frac{1}{\sqrt{k+1}}$.
  Вопрос: какой видно плюс и минус?

Сходимость SGD: выбор γ_k



Точнее сходимость, но теряется линейная сходимость в начале.

Сходимость SGD: проблема сходимости

Как можно попробовать решить проблемы неточной сходимости?

Сходимость SGD: проблема сходимости

Как можно попробовать решить проблемы неточной сходимости?

- Уменьшить σ. **Вопрос:** а как?

Сходимость SGD: проблема сходимости

Как можно попробовать решить проблемы неточной сходимости?

- Уменьшить σ. **Вопрос:** а как? С помощью техники батчинга/батчирования:

$$\nabla f(w^k, \xi^k) \quad \to \quad \frac{1}{b} \sum_{j \in S^k} \nabla f(x, \xi_j),$$

где S^k – набор индексов из $[n]$, $|S^k| = b$, и все индексы генерируются независимо друг от друга.

Сходимость SGD: батчинг

- **Вопрос:** что можем сказать про

$$\mathbb{E}\left[\frac{1}{b}\sum_{j\in S^k}\nabla f(x,\xi_j)\mid w^k\right], \quad \mathbb{E}\left[\left\|\frac{1}{b}\sum_{j\in S^k}(\nabla f(x,\xi_j)-\nabla f(x))\right\|_2^2\mid w^k\right]?$$

Сходимость SGD: батчинг

- **Вопрос:** что можем сказать про

$$\mathbb{E}\left[\frac{1}{b}\sum_{j\in S^k}\nabla f(x,\xi_j)\mid w^k\right], \quad \mathbb{E}\left[\left\|\frac{1}{b}\sum_{j\in S^k}(\nabla f(x,\xi_j)-\nabla f(x))\right\|_2^2\mid w^k\right]?$$

- Независимость дает

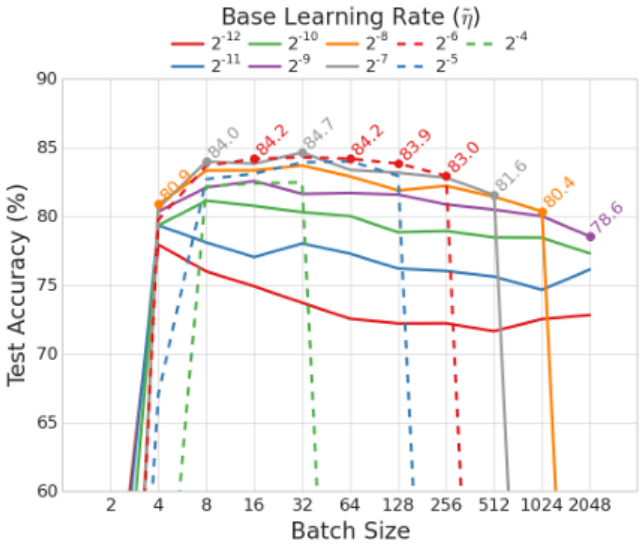
$$\mathbb{E}\left[\frac{1}{b}\sum_{j\in S^k}\nabla f(x,\xi_j)\mid w^k\right]=\nabla f(x),$$

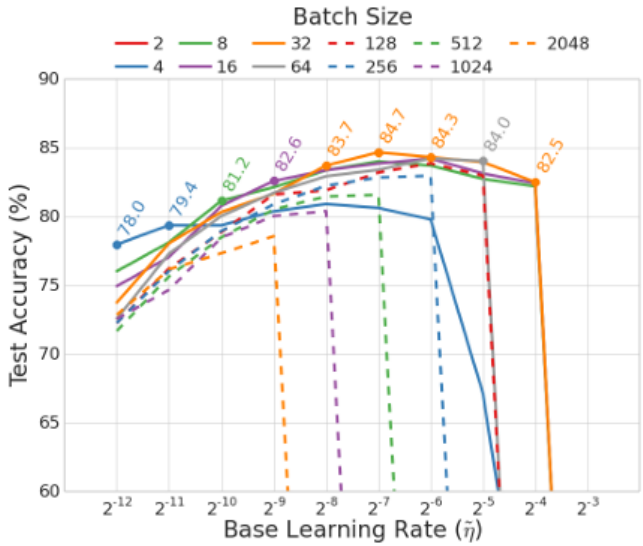
$$\mathbb{E}\left[\left\|\frac{1}{b}\sum_{j\in S^k}(\nabla f(x,\xi_j)-\nabla f(x))\right\|_2^2\mid w^k\right]\leq\frac{\sigma^2}{b}$$

- Получается дисперсию можно уменьшить в b раз, но тогда и вычисление стохастического градиента подорожает.

Сходимость SGD: батчинг

В статье[1] обучали AlexNet на датасете CIFAR-10[2] с разными размерами батчей и learning rate. Результатом стало то, что оптимальный размер батча в данной задаче лежит от 4 до 32.





[1] Revisiting Small Batch Training for Deep Neural Networks, Dominic Masters and Carlo Luschi, 2018

[2] Датасеты CIFAR-10 и CIFAR-100

Сходимость SGD

- В итоге можно подобрать стратегию выбора шагов и добиться следующей оценки сходимости:

$$\mathbb{E}\left[\|w^k - w^*\|^2\right] \leq \left(1 - \frac{\mu}{L}\right)^k \mathbb{E}\left[\|w^0 - w^*\|^2\right] + \frac{\sigma^2}{\mu^2 bk}.$$

Линейная по «детерминистической» части и сублинейная по «стохастической».

Сходимость SGD

- В итоге можно подобрать стратегию выбора шагов и добиться следующей оценки сходимости:

$$\mathbb{E}\left[\|w^k - w^*\|^2\right] \leq \left(1 - \frac{\mu}{L}\right)^k \mathbb{E}\left[\|w^0 - w^*\|^2\right] + \frac{\sigma^2}{\mu^2 bk}.$$

  Линейная по «детерминистической» части и сублинейная по «стохастической».

- Ускорение Нестерова возможно:

$$\mathbb{E}\left[\|w^k - w^*\|^2\right] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \mathbb{E}\left[\|w^0 - w^*\|^2\right] + \frac{\sigma^2}{\mu^2 bk}.$$

Сходимость SGD

- В итоге можно подобрать стратегию выбора шагов и добиться следующей оценки сходимости:

$$\mathbb{E}\left[\|w^k - w^*\|^2\right] \leq \left(1 - \frac{\mu}{L}\right)^k \mathbb{E}\left[\|w^0 - w^*\|^2\right] + \frac{\sigma^2}{\mu^2 bk}.$$

  Линейная по «детерминистической» части и сублинейная по «стохастической».

- Ускорение Нестерова возможно:

$$\mathbb{E}\left[\|w^k - w^*\|^2\right] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \mathbb{E}\left[\|w^0 - w^*\|^2\right] + \frac{\sigma^2}{\mu^2 bk}.$$

  Важно! Мы улучшаем/ускоряем только первый член, второй член (который и возникает из-за стохастики) остался прежним.

Сходимость SGD

- В итоге можно подобрать стратегию выбора шагов и добиться следующей оценки сходимости:

$$\mathbb{E}\left[\|w^k - w^*\|^2\right] \leq \left(1 - \frac{\mu}{L}\right)^k \mathbb{E}\left[\|w^0 - w^*\|^2\right] + \frac{\sigma^2}{\mu^2 bk}.$$

  Линейная по «детерминистической» части и сублинейная по «стохастической».
- Ускорение Нестерова возможно:

$$\mathbb{E}\left[\|w^k - w^*\|^2\right] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \mathbb{E}\left[\|w^0 - w^*\|^2\right] + \frac{\sigma^2}{\mu^2 bk}.$$

  Важно! Мы улучшаем/ускоряем только первый член, второй член (который и возникает из-за стохастики) остался прежним.
- Ускорение Нестерова может быть более эффективно вместе с батчингом.