1. Рассмотрим задачу минимизации суммы

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x) \right\}$$

где функции $f_i(x)$ являются выпуклыми и L-гладкими, а функция $f(x)$ является μ-сильно выпуклой. Пусть m - очень большое (порядка 10^6). Выберите правильное утверждение:

(a) SGD с постоянным размером шага достигает не очень высокой точности решения быстрее, чем GD

(b) GD сходится быстрее, чем SGD, поскольку стох. градиент плохо аппроксимирует градиент.

(c) SGD достигает любую наперёд заданную точность решения задачи быстрее, чем GD

(d) Метод Нестерова сходится быстрее, чем SGD, поскольку является оптимальным методом для задач (сильно) выпуклой гладкой оптимизации.

Ответ: a

2. Рассмотрим SGD с постоянным шагом в случае, когда стохастический градиент имеет равномерно ограниченную дисперсию. Как было показано на лекции, SGD сходится линейно (экспоненциально быстро) к некоторой окрестности решения. Выберите неправильное утверждение:

(a) Размер окрестности можно уменьшить, если уменьшить размер шага

(b) Размер окрестности можно уменьшить, если увеличить размер батча

(c) Размер окрестности не зависит ни от размера шага, ни от размера батча

Сходимость происходит линейно к окресности с радиусом $\sqrt{\frac{\gamma\sigma^2}{\mu r}}$, где γ - размер шага, r - размер батча. Отсюда следует, что неверный ответ c

Ответ: c

3. Выберите правильное утверждение про методы редукции дисперсии:

(a) Методы редукции дисперсии бесполезны, поскольку требуют знания градиентов слагаемых в решении задачи

(b) Методы редукции дисперсии - это стохастические методы, которые сходятся линейно всегда, когда SGD не сходится линейно

(c) Полноградиентные методы всегда работают быстрее, чем SVRG/L-SVRG, поскольку в SVRG/L-SVRG требуется вычислять полный градиент в точке w/w^k

(d) Методы редукции дисперсии - это стохастические методы, итерации которых сравнимы по сложности с итерациями SGD, но которые сходятся линейно в случае, когда $f(x)$ является сильно выпуклой, а $f_i(x)$ - выпуклые и L-гладкие функции.

Ответ: d

4. Предположим, что мы хотим решить задачу из первого вопроса достаточно точно. Выберите правильное утверждение:

(a) Если m достаточно большое, то выгоднее использовать полноградиентные методы, например, GD или Ускоренный метод Нестерова.

(b) Если m достаточно большое, то выгоднее всего использовать методы редукции дисперсии.

(c) Если m достаточно большое, то выгоднее всего использовать методы редукции дисперсии только в случае, когда $f(x)$ является сильно выпуклой ($\mu > 0$).

(d) Если m достаточно большое, то выгоднее всего использовать SGD.

Вариант a будет сходиться очень долго, вариант d не гарантирует точной сходимости к решению, вариант c не верен, т.к. методы редукции дисперсии хорошо работают и на выпуклых (не сильно выпуклых) функциях

Ответ: b

5. Выберите неправильное утверждение про over-parameterized модели:

(a) Для over-parameterized моделей типично, что целевая функция $f(x)$ является выпуклой.

(b) Over-parameterized модели при определённых предположениях имеют хорошую обобщающую способность

(c) Для over-parameterized моделей типично, что $n >> m$

(d) Для over-parameterized моделей типично, что $f(x^*) = 0$ или $f(x^*) \approx 0$, т.е. модель может идеально или почти идеально "подогнаться"под данные.

Неверен вариант a, т.к. обычно функция не является выпуклой, но удовлетворяет условию Поляка-Лоясевича (что обычно свойственно сильно-выпуклым функциям) практически везде в пространстве.

Ответ: a

Ответы:

1. a
2. c
3. d
4. b
5. a