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. Выберите неправильное утверждение о данной задаче.

(a) Из условия слабого роста следует, что у функций $f_i(x)$ существует общая точка минимума.

(b) Если выполнены условия интерполяции, то это не означает, что выполнено условие сильного роста.

(c) Из условия сильного роста следует условие слабого роста.

(d) Из условия слабого роста следует условие сильного роста.

Т.к. из условия интерполяции следует условие слабого роста, а из условия слабого роста (учитывая что $f(x)$ является μ-сильно выпуклой) следует условие сильного роста, то неверным утверждением будет b
Ответ: b

2. Рассмотрим теперь задачу

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

где функции $f_i(x)$ являются выпуклыми, $f_i(x)=\ell(h_i(x)),\ \ell:\mathbb{R}\to\mathbb{R}_+$ - выпуклая неотрицательная 1-гладкая функция, $\inf_{x\in\mathbb{R}}\ell(x)=0,\ h_i$ - Липшицева с константой M. Выберите правильное утверждение о сходимости стохастичекого субградиентного спуска (SSGD).

(a) В данной задаче выполнены условия интерполяции, поэтому SSGD сходится за $O\left(\frac{M^2R_0^2}{\varepsilon}\right)$ итераций

(b) Поскольку рассматриваемая функция является негладкой, SSGD сходится за $O\left(\frac{M^2R_0^2}{\varepsilon}\right)$ итераций, что соответствует нижней оценке

(c) Если существует x^*, такая что $f(x^*)=0$, то SSGD сходится за $O\left(\frac{M^2R_0^2}{\varepsilon}\right)$ итераций
Ответ: c

3. Рассмотрим задачу минимизации математического ожидания:

$$\min_{x\in\mathbb{R}^n}\left\{f(x)=\mathbb{E}_\xi[f(x,\xi)]\right\},$$

где функция f является L-гладкой и ограниченной снизу значением f^*, $\mathbb{E}_\xi[\|\nabla f(x,\xi)-\nabla f(x)\|_2^2]\le\sigma^2$ для всех $x\in\mathbb{R}^n$. Выберите неправильное утверждение.

(a) SGD находит ε-стационарную точку, используя $O\left(\max\left\{\frac{L(f(x_0)-f^*)}{\varepsilon^2}, \frac{L(f(x_0)-f^*)\sigma^2}{\varepsilon^4}\right\}\right)$ подсчётов стох. градиента

(b) SGD является оптимальным методом для поиска ε-стационарной точки в данной задаче

(c) Momentum-SGD часто работает на практике лучше, но в теории не превосходит SGD

(d) В указанной задаче любой локальный минимум является глобальным.

Ответ: d

4. Выберите правильное утверждение о методах редукции дисперсии и методе перестановок

(a) Методы редукции дисперсии как правило оказываются неэффективными при обучении глубоких нейросетей; метод перестановок на практике работает лучше, чем SGD, но в теории оказывается доказуемо лучше только при большом числе эпох (проходов по датасету)

(b) Методы редукции дисперсии хорошо работают на простых моделях (линейная регрессия, логистическая регрессия); метод перестановок - эвристика, которая работает лучше стандартного SGD на практике, но в теории никакого преимущества нет

(c) Методы редукции дисперсии как правило оказываются неэффективными при обучении глубоких нейросетей; метод перестановок - эвристика, которая работает лучше стандартного SGD на практике, но в теории никакого преимущества нет

(d) Методы редукции дисперсии хорошо работают на простых моделях (линейная регрессия, логистическая регрессия); метод перестановок и SGD работают примерно одинаково

Ответ: a

5. Выберите правильное утверждение о Momentum-SGD и адаптивных методах (Adagrad, Adam)

(a) Momentum-SGD обычно работает лучше чем Adam, когда шум в стох. градиентах имеет лёгкие хвосты распределения ("картиночные"задачи); с точки зрения скорости поиска стационарных точек Adagrad и Adam имеют теоретические гарантии не лучше, чем у SGD

(b) Adam как правило работает лучше, чем Momentum-SGD, когда шум в стох. градиентах имеет тяжёлые хвосты распределения ("текстовые"задачи); этому есть строгое теоретическое обоснование в статье, рассмотренной в конце лекции

(c) Momentum-SGD обычно работает лучше, чем Adam, когда шум в стох. градиентах имеет лёгкие хвосты распределения ("картиночные"задачи); с точки зрения скорости поиска стационарных точек Adagrad и Adam имеют теоретические гарантии лучше, чем у SGD

(d) Adam как правило работает лучше, чем Momentum-SGD, когда шум в стох.

градиентах имеет тяжёлые хвосты распределения ("текстовые"задачи); согласно современным теоретическим результатом Adam работает лучше, чем Adagrad, в контексте поиска ε-стационарных точек
Ответ: a

Ответы:
1. b
2. c
3. d
4. a
5. a