# Постановка задачи

В данном проекте будет рассматриваться следущая задача оптимизации:
$$\min_x f(x) := \mathbb{E}[F(x,\xi)]$$
при условии $x \in X$, $X$ – замкнутое выпуклое подмножество $\mathbb{R}^n$, $F(x,\xi) \ : \ X \to \mathbb{R}$ – функция из класса $C^1(X)$, $\xi$ – случайная величина.

# Метод SGD

Метод стохастического градиента получается из классического метода градиентого спуска заменой градиента $\nabla f(x)$ на стохастический градиент $G(x,\xi) = \partial_x F(x,\xi)$. Если функция $f$ является $L$ – гладкой и $\exists \sigma > 0$ $\mathbb{E}\|G(x,\xi) - \mathbb{E}[G(x,\xi)]\|_*^2 \leq \sigma^2$, то оптимальная скорость сходимости принадлежит $\mathcal{O}(\frac{L}{k^2} + \frac{\sigma}{\sqrt{k}})$. Исследуемый в данном проекте алгоритм достигает указанной оптимальной скорости сходимости.

# Цель проекта

В данном проекте будет рассмотрен метод стохастического градиентного спуска A2Grad. А также будет достигнута оптимальная скорость сходимости $\mathcal{O}(\frac{L}{k^2} + \frac{\sigma}{\sqrt{k}})$, а также исследованы разные варианты размеров шагов.

# Стохастический градиентный спуск

Нашей целью будет найти новый градиентный метод, который за основу берёт продвинутый алгоритм Нестерова, использовующий метод моментов и адаптивный алгоритм стохастического градиентного спуска.
Приведём алгоритм Adaptive SGD.

**Algorithm 1:** Adaptive algorithm

**Input:** $x_0, v_0$
**for** `k=0,1,2,...,K` **do**
  Получаем $\xi_k$ и вычисляем $G_k \in \nabla F(x_k,\xi_k)$; $v_k = \psi(G_0^2, G_1^2, G_2^2, ..., G_k^2)$;
  $G_k = \phi(G_0, G_1, G_2, ..., G_k)$;
  $x_{k+1} = x_k - \beta_k G_k / \sqrt{v_k}$;
**end**

# Adaptive ASGD

Определение: $D_\phi(x,y) := \phi(x) - \phi(y) - \langle \phi(y), x-y \rangle$, где $\phi$ — выпуклая гладкая функция. Для удобства положим $D(x,y) = D_\psi(x,y)$, где $\psi(x) = \frac{1}{2}\|x\|^2$.

Общий вид рассматриваемого в статье метода выглядит так:

**Algorithm 2:** A2Grad algorithm

**Input:** $x_0, \overline{x_0}, \gamma_k, \beta_k > 0$
**for** `k=0,1,2,...,K` **do**
  $\underline{x_k} = (1-\alpha_k)\overline{x_k} + \alpha_k x_k$;
  Получаем $\xi_k$, вычисляем $\underline{G_k} \in \nabla F(\underline{x_k}, \xi_k)$ и $\phi_k(\cdot)$;
  $x_{k+1} = \arg\min\{\langle \underline{G_k}, x\rangle + \gamma_k D(x_k, x) + \beta_k D_{\phi_k}(x_k, x)\}$;
  $\overline{x_{k+1}} = (1-\alpha_k)\overline{x_k} + \alpha_k x_{k+1}$;
**end**
**Output:** $\overline{x_{K+1}}$

Осталось разобраться с выбором констант и функций.
В качестве $\phi_k$ возьмём $\frac{1}{2}\|x\|^2_{h_k} = \frac{1}{2}\langle x^T, diag(h_k)x\rangle$, где $h_k \in \mathbb{R}^d$, $h_{k,i} > 0$. Тогда $D_{\phi_k} = \frac{1}{2}\langle (x-y)^T, diag(h_k)(x-y)\rangle$. То, что получилось, будем называть diagonal scaling. В такой версии алгоритма $x_{k+1} = proj\ (x_k - \frac{1}{\gamma_k + \beta_k h_k}\underline{G_k})$.
Положим $\alpha_k = \frac{2}{k+2}, \gamma_k = \frac{2L}{k+1}, \beta_k = \beta$, $\beta$ — параметр алгоритма.
Для выбора $h_k$ мы рассмотрим 3 способа:

- Uniform moving average
  Положим $v_{-1} = 0,\ v_k = v_{k-1} + \delta_k^2,\ h_k = \sqrt{v_k}$
- Incremental moving average
  Положим $v_{-1} = 0,\ v_k = \frac{k^2}{(k+1)^2}v_{k-1} + \delta_k^2,\ h_k = \sqrt{v_k}$
- Exponential moving average
  $$\widetilde{v_k} = \begin{cases} 0, & \text{если } k=-1 \\ \delta_k^2, & \text{если } k=0 \\ \rho\widetilde{v_{k-1}} + (1-\rho)\delta_k^2, & \text{иначе} \end{cases}$$
  $v_k = max(v_{k-1}, \widetilde{v_k})$
  $h_k = \sqrt{(k+1)h_k}$

Здесь $\delta_k = \underline{G_k} - \frac{1}{k+1}\sum\limits_{i=0}^{k}\underline{G_i}$.

# Сходимость

**Теорема.** Если в алгоритме 2 функция $f$ выпуклая и $L$-гладкая с константой $L$, и $\{\alpha_k\}, \{\gamma_k\}$ удовлетворяют следующим условиям:

$$L\alpha_k \leq \gamma_k$$

$$\lambda_{k+1}\alpha_{k+1}\gamma_{k+1} \leq \lambda_k\alpha_k\gamma_k$$

где последовательность $\{\lambda_k\}$ - это:

$$\lambda_0 = 1$$

$$\lambda_k = \frac{1}{\prod_{i=1}^{k}(1-\alpha_i)}$$

Тогда верно следующее неравенство:

$$\lambda_K\left[f(\overline{x}_{K+1}) - f(x)\right] \leq (1-\alpha_0)\left[f(\overline{x}_0) - f(x)\right] + \alpha_0\gamma_0 D(x_0, x) +$$

$$+\sum_{k=0} K\lambda_k\left[\lambda_k\frac{\alpha_k||\delta_k||^2_{\phi_k*}}{2\beta_k} + \alpha_k\langle\delta_k, x - x_k\rangle + \alpha_k R_k\right]$$

, где $\delta_k = \underline{G}_k - \nabla f(x_k)$ и $R_k = \beta_k D_{\phi_k}(x_k, x) - \beta_k D_{\phi_k}(x_{k+1}, x)$

**Теорема.** Пусть $x^*$ — глобальный минимум, причем $\exists B > 0 \ \|x_k - x^*\|^2_\infty < B$. Пусть в схеме exponential moving average $\rho \in (0,1)$, и распределение ошибки $\delta_k$ таково, что $\exists \sigma \geq 0 \ \forall t > 0 \ \mathbb{E}e^{t\delta_k} \leq e^{t^2\sigma^2/2}$. Тогда

$$\mathbb{E}[f(\overline{x_{K+1}}) - f(x^*)] \leq \frac{2L\|x^* - x_0\|2}{(K+1)(K+2)} + 2\beta B\frac{\sqrt{2\log(2(K+1))}\sigma}{\sqrt{K+2}} + \frac{\sqrt{2\varpi}d\sigma}{2\beta(1-\rho)\sqrt{K+2}}$$

.

# Анализ литературы

В ходе анализа литературы было получено, что описанный в статье метод является state-of-the-art среди стохастических градиентных методов.

# Эмпирические результаты

Описанные в статье три версии A2Grad (uniform, incremental, exponential moving average) были протестированы на линейной регрессии и логистической регрессии против Adam, SGD и LBFGS (квазиньютоновский метод). Итак, на логистической регрессии описанные в статье методы показали лучший результат. На линейной регрессии incremental moving average сходится медленно, что согласуется с результатами в статье.

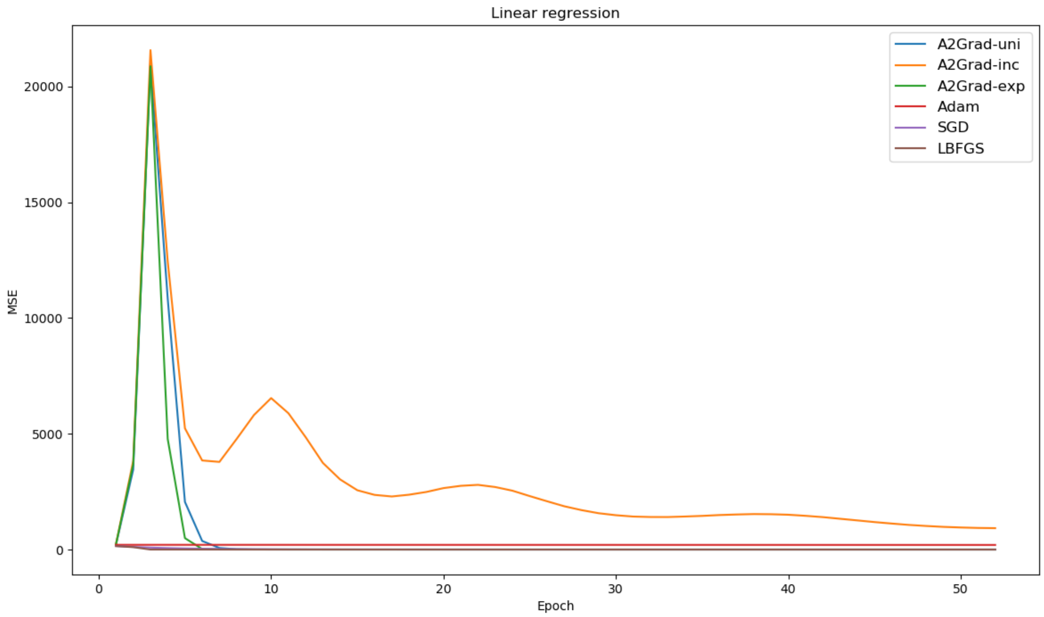


Рис. 1: Linear regression

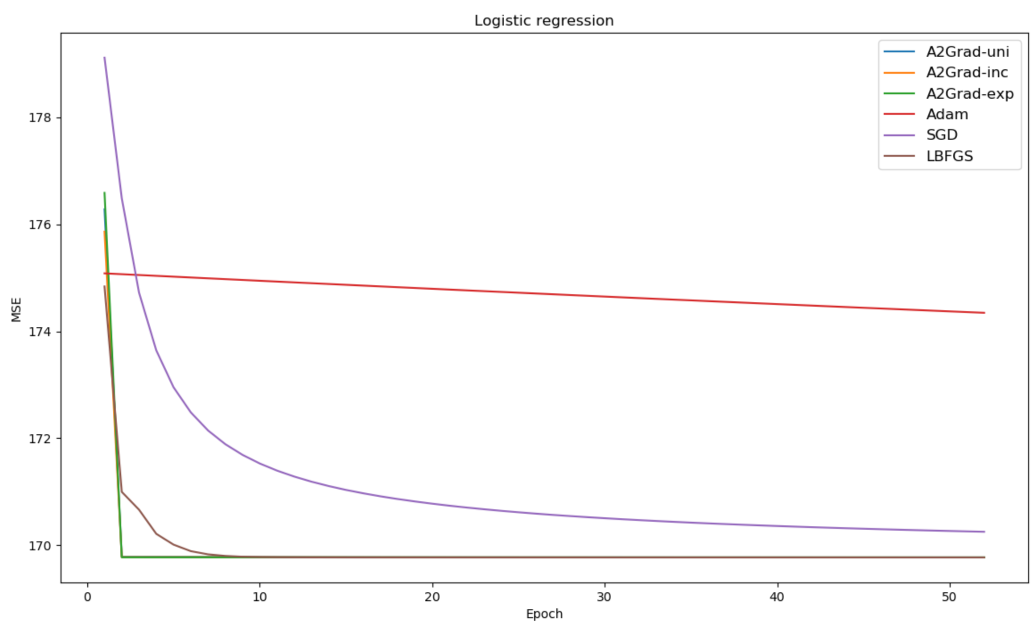


Рис. 2: Logistic regression