Задача 1. *(1) По определению оценки максимального правдоподобия параметра* θ *:*

$$\hat{\theta} = \arg\max p_\theta(X_1, X_2, \ldots, X_n)$$

каждая величина из гамма распределения и они независимы между собой:

$$p_\theta(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} p_\theta(X_i) = \frac{\theta^{\beta n}}{\Gamma(\beta)^n}(\prod_{i=1}^{n} X_i)^{\beta-1} e^{-\theta\sum_{i=1}^{n} X_i}$$

Натуральный логарифм монотонная возрастающая функция так, что можно максимизировать логарифм правдоподобия:

$$\arg\max p_\theta(X_1, X_2, \ldots, X_n) = \arg\max \ln(p_\theta(X_1, X_2, \ldots, X_n))$$

$$\ln(p_\theta(X_1, X_2, \ldots, X_n)) = \ln(\frac{\theta^{\beta n}}{\Gamma(\beta)^n}) + \ln((\prod_{i=1}^{n} X_i)^{\beta-1}) + \ln(e^{-\theta\sum_{i=1}^{n} X_i}) =$$

$$= \beta n \ln(\theta) - \ln(\Gamma(\beta)^n) + (\beta - 1)\sum_{i=1}^{n} \ln(X_i) - \theta \sum_{i=1}^{n} X_i$$

Возьмём производную по θ *и приравняв к нулю найдем экстремум:*

$$\frac{d \ln(p_\theta(X_1, X_2, \ldots, X_n))}{d\theta} = \frac{\beta n}{\theta} - \sum_{i=1}^{n} X_i$$

$$\Rightarrow \hat{\theta} = \frac{\beta n}{\sum_{i=1}^{n} X_i}$$

Вычислим вторую производную и убедимся что она отрицательна $\Rightarrow$ *выпукла вниз и* $\hat{\theta}-$ *максимум:*

$$\frac{d^2 p_\theta(X_1, X_2, \ldots, X_n)}{d^2\theta} = -\frac{\beta n}{\theta^2} < 0$$

(2) Рассмотрим $X_1, X_2, \ldots, X_n - i.i.d$ *случайные величины. Логарифмическая функция правдоподобия будет выглядеть так:*

$$\ln(p_\theta(X_1, X_2, \ldots, X_n)) = \ln(\prod_{i=1}^{n} p_\theta(X = x_1)) = \ln(\prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!}) = \sum_{i=1}^{n}(x_i \ln(\theta) - n\theta - \ln(x_i!))$$

рассмотрим теперь производные первого и второго порядка:

$$\frac{dp_\theta(X_1, X_2, \ldots, X_n)}{d\theta} = \sum_{i=1}^{n}(\frac{x_i}{\theta} - n)$$

$$\Rightarrow \hat{\theta} = \frac{\sum_{i=1}^{n} X_i}{n}$$

покажем, что функция выпукла вниз и $\hat{\theta}-$ *экстремум максимум*

$$\frac{d^2 p_\theta(X_1, X_2, \ldots, X_n)}{d^2\theta} = \frac{-\sum_{i=1}^{n} X_i}{\theta^2} < 0$$

Задача 2. *(1)* $X_i, \theta \in \mathbb{R}^D$ *(строки, не столбцы)*

Функция правдоподобия:

$$\mathcal{L}_y(\theta) = \prod_{i=1}^{n} \sigma(X_i\theta^T)^{y_i}(1 - \sigma(X_i\theta^T)^{1-y_i}$$

Логарифмическая функция правдоподобия:

$$l_y(\theta) = \sum_{i=1}^{n} y_i \ln(\sigma(X_i\theta^T)) + (1 - y_i)\ln(1 - \sigma(X_i\theta^T))$$

Градиент логарифма правдоподобия:

$$\frac{\partial l_y(\theta)}{\partial\theta} = \sum_{i=1}^{n}[y_i - \sigma(X_i\theta^T)]X_i$$

Матричный вид:

$$\frac{\partial l_y(\theta)}{\partial\theta} = X^T(y - S(\theta))$$

где:

$$S(\theta) = [\sigma(X_1\theta^T), \ldots, \sigma(X_n\theta^T)]$$

С регуляризацией:

$$F(\theta) = -l_y(\theta) + \lambda||\theta||^2 \to min$$

Градиентный спуск:

θ_t- *состояние весов на итерации* t

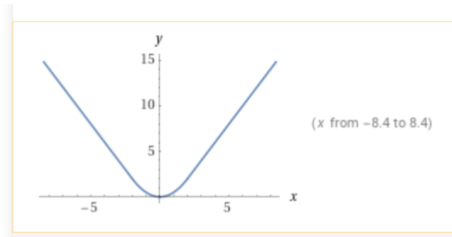
$$\theta_{t+1} = \theta_t - \alpha\nabla_\theta F(\theta)$$

$$\theta_{t+1} = \theta_t - \alpha \frac{1}{n} X^T (y - S(\theta_t)) - 2\alpha\lambda\theta$$

Стохастический градиентный спуск по батчам B:

$$\theta_{t+1} = \theta_t - \alpha \frac{1}{B} X_B^T (y_B - S_B(\theta_t)) - 2\alpha\lambda\theta$$

X_B, y_B, S_B- *подвыборки строк под батч*



Задача 3.

$$R(x) = \frac{x^2}{2} I\{|x| \leqslant C\} + C(|x| - \frac{C}{2}) I\{|x| > C\}$$

(1) *Особенность функции в том, что при достаточно больших значениях функция начинает штрафовать линейно от порядка значения и следовательно не так сильно реагирует на большие выбросы в данных. А при при ошибках ниже заданного порога-гиперпараметра минимизируется обычная квадратичная функция.*

(2) *Пусть строки(не столбцы)* $X_i, \theta \in \mathbb{R}^D$, *где* $D-$ *количество признаков* $i)|(Y_i - X_i\theta^T)| \leqslant C$

$$R(x) = \frac{x^2}{2}$$

$$\nabla_\theta R(Y_i - X_i\theta^T) = -X_i(Y_i - X_i\theta^T)$$

$ii)|(Y_i - X_i\theta^T)| > C$

$$R(x) = C(|x| - \frac{C}{2})$$

$$\nabla_\theta R(Y_i - X_i\theta^T) = C\nabla_\theta |Y_i - X_i\theta^T| = -CX_i sgn(Y_i - X_i\theta^T)$$

тогда

$$\nabla_\theta R(Y_i - x_i^T\theta) = \begin{cases} -X_i(Y_i - X_i^T\theta), \left|Y_i - X_i^T\theta\right| \leqslant c \\ -CX_i \cdot sgn(Y_i - X_i^T\theta), \left|Y_i - X_i^T\theta\right| > c \end{cases}$$

Распишем формулы град. спуска и стохастического град. спуска:

$$\theta_{t+1} = \theta_t - \alpha\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta R(Y_i - x_i^T\theta)$$

$$\theta_{t+1} = \theta_t - \alpha\frac{1}{B}\sum_{i=1}^{B} \nabla_\theta R(Y_i - x_i^T\theta)$$

где B *- размер батча*