Стандартная оптимизационная задача, которая возникает в машинном обучении имеет вид:

$$\min_{x\in\mathbb{R}^d}\left[f(x)=\frac{1}{n}\sum_{i=1}^{n}f_i(x)\right]. \tag{1}$$

Определение 1. *Дифференцируемая функция $\varphi:\mathbb{R}^d\to\mathbb{R}$ является сильно выпуклой функцией с константой μ, если для любых $x,y\in\mathbb{R}^d$ выполнено*

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

Для дважды непрерывно дифференцируемой фукнции сильная выпуклость эквивалентна условию, что минимальное собственное значение гессиана ограничено снизу положительной константой, т. е. $\nabla^2\varphi(x)\succeq\mu I$, где $I\in\mathbb{R}^{d\times d}$ — единичная матрица. Другими словами, $\lambda_{min}(\nabla^2\varphi(x))\geq\mu$.

Определение 2. *Дважды непрерывно дифференцируемая функция имеет липшецев гессиан с константой H, если для любых $x,y\in\mathbb{R}^d$ выполнено*

$$\left\|\nabla^2\varphi(x)-\nabla^2\varphi(y)\right\|\leq H\|x-y\|. \tag{3}$$

Для дважды непрерывно дифференцируемой функции с липшецевым гессианом справедлива лемма

Лемма 1. *Пусть функция $\varphi:\mathbb{R}^d\to\mathbb{R}$ дважды непрерывно дифференцируема, а ее гессиан липшецев с константой H. Тогда выполнены неравенства*

$$\left\|\nabla\varphi(y)-\nabla\varphi(x)-\nabla^2\varphi(x)(y-x)\right\|\leq\frac{H}{2}\|y-x\|^2, \tag{4}$$

$$\left|\varphi(y)-\varphi(x)-\langle\nabla\varphi(x),y-x\rangle-\frac{1}{2}\left\langle\nabla^2\varphi(x)(y-x),y-x\right\rangle\right|\leq\frac{H}{6}\|y-x\|^3. \tag{5}$$

Доказательство. Действительно,

$$\begin{aligned}\left\|\nabla\varphi(y)-\nabla\varphi(x)-\nabla^2\varphi(x)(y-x)\right\| &= \left\|\int_0^1\left[\nabla^2\varphi(x+\tau(y-x))-\nabla^2\varphi(x)\right](y-x)d\tau\right\|\\ &\leq H\|y-x\|^2\int_0^1\tau d\tau=\frac{H}{2}\|y-x\|^2.\end{aligned}$$

Следовательно,

$$\begin{aligned}&\left|\varphi(y)-\varphi(x)-\langle\nabla\varphi(x),y-x\rangle-\frac{1}{2}\left\langle\nabla^2\varphi(x)(y-x),y-x\right\rangle\right|\\ =&\left|\int_0^1\left\langle\nabla\varphi(x+t(y-x))-\nabla\varphi(x)-t\nabla^2\varphi(x)(y-x),y-x\right\rangle dt\right|\\ \leq&\frac{H}{2}\|y-x\|^3\int_0^1 t^2dt=\frac{H}{6}\|y-x\|^3.\end{aligned}$$

□

1 Стохастический метод Ньютона

Будем аппроксимировать гессиан и градиент функции, используя последнюю доступную информацию, т. е.

$$\nabla^2 f(x^k) \approx \frac{1}{n}\sum_{i=1}^{n} \nabla^2 f_i(w_i^k), \qquad \nabla f(x^k) \approx \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w_i^k),$$

где w_i^k — последний вектор, для которого были посчитаны ∇f_i и $\nabla^2 f_i$.

Algorithm 1 Стохастический метод Ньютона

Initialize: Choose starting iterates $w_1^0, w_2^0, \dots, w_0^n \in \mathbb{R}^d$ and minibatch size $\tau \in \{1, 2, \dots, n\}$

for $k = 0, 1, 2, \dots$ **do**

$x^{k+1} = \left[\frac{1}{n}\sum\limits_{i=1}^{n} \nabla^2 f_i(w_i^k)\right]^{-1} \left[\frac{1}{n}\sum\limits_{i=1}^{n} \nabla^2 f_i(w_i^k) w_i^k - \nabla f_i(w_i^k)\right]$

Choose a subset $S^k \subseteq \{1, \dots, n\}$ of size τ uniformly at random

$w_i^{k+1} = \begin{cases} x^{k+1}, & i \in S^k \\ w_i^k, & i \notin S^k \end{cases}$

end for

1.1 Локальная сходимость метода

Обозначим за $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid x^k, w_1^k, \dots, w_n^k]$ условное матожидание, связанное со всей информацией предшествующей итерации $k+1$. Обозначим x^* как решение исходной задачи. Введем функцию Ляпунова вида

$$\mathcal{W}^k := \frac{1}{n}\sum_{i=1}^{n} \|w_i^k - x^*\|^2.$$

Лемма 2. *Пусть функция f_i является сильно выпуклой с константой μ и имеет липшецев гессиан с константой H для всех $i \in \{1, 2, \dots, n\}$. Тогда на следующем шаге Алгоритма 1 выполнено*

$$\|x^{k+1} - x^*\| \leq \frac{H}{2\mu}\mathcal{W}^k. \tag{6}$$

Доказательство. Пусть $H^k := \frac{1}{n}\sum_{i=1}^{n} \nabla^2 f_i(w_i^k)$, тогда шаг алгоритма может быть записан в виде

$$x^{k+1} = \left(H^k\right)^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} \nabla^2 f_i(w_i^k) w_i^k - \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w_i^k)\right].$$

Кроме этого, имеем $x^* = \left(H^k\right)^{-1} H^k x^*$ и $\sum_{i=1}^{n} \nabla f_i(x^*) = 0$. Тогда это ведет к равенству

$$x^{k+1} - x^* = \left(H^k\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \left[\nabla^2 f_i(w_i^k)(w_i^k - x^*) - \left(\nabla f_i(w_i^k) - \nabla f_i(x^*)\right)\right]. \tag{7}$$

Раз f_i сильно выпукла, то $\nabla f_i^2(w_i^k) \succeq \mu I$ для всех i. Как следствие, $H^k \succeq \mu I$, что ведет к неравенству

$$\left\|\left(H^k\right)^{-1}\right\| \leq \frac{1}{\mu}. \tag{8}$$

$$
\begin{aligned}
\|x^{k+1}-x^*\| &\overset{(7)}{\leq} \left\|\left(H^k\right)^{-1}\right\| \left\| \frac{1}{n}\sum_{i=1}^{n}\left[\nabla^2 f_i(w_i^k)(w_i^k-x^*)-\left(\nabla f_i(w_i^k)-\nabla f_i(x^*)\right)\right]\right\| \\
&\overset{(8)}{\leq} \frac{1}{\mu}\left\|\frac{1}{n}\sum_{i=1}^{n}\left[\left(\nabla f_i(x^*)-\nabla f_i(w_i^k)\right)-\nabla^2 f_i(w_i^k)(x^*-w_i^k)\right]\right\| \\
&\leq \frac{1}{n\mu}\sum_{i=1}^{n}\left\|\nabla f_i(x^*)-\nabla f_i(w_i^k)-\nabla^2 f_i(w_i^k)(x^*-w_i^k)\right\| \\
&\overset{(4)}{\leq} \frac{H}{2n\mu}\sum_{i=1}^{n}\|w_i^k-x^*\|^2 = \frac{H}{2\mu}\mathcal{W}^k.
\end{aligned}
$$

□

Лемма 3. *Пусть каждая функция f_i является сильно выпуклой с константой μ и имеет липшецев гессиан с константой H. Если $\|w_i^0 - x^*\| \leq \frac{\mu}{H}$ для всех $i \in \{1, 2, \ldots, n\}$, тогда для всех k имеем*

$$
\mathcal{W}^k \leq \frac{\mu^2}{H^2}. \tag{9}
$$

Доказательство. Покажем, что

$$
\|w_i^k - x^*\|^2 \leq \frac{\mu^2}{H^2} \quad \forall i \in \{1, 2, \ldots, n\}. \tag{10}
$$

Тогда ограничение сверху на $\mathcal{W}^k$ будет следовать из этого. Теперь докажем утверждение выше по индукции. Пусть оно верно для k, покажем, что оно верно и для $k+1$. Если $i \notin S^k$, то $w_i^k = w_i^{k+1}$, и утверждение выполнено по предположению индукции. Если $i \in S^k$, то

$$
\|w_i^{k+1} - x^*\| = \|x^{k+1} - x^*\| \overset{(2)}{\leq} \frac{H}{2\mu}\frac{1}{n}\sum_{j=1}^{n}\|w_j^k - x^*\|^2 \leq \frac{H}{2\mu}\frac{1}{n}\sum_{j=1}^{n}\frac{\mu^2}{H^2} < \frac{\mu}{H}.
$$

Поэтому мы снова получаем (10) верно. □

Лемма 4. *Случайные переменные Алгоритма* (1) *удовлетворяют равенству*

$$
\mathbb{E}_k\left[\mathcal{W}^{k+1}\right] = \frac{\tau}{n}\mathbb{E}_k\left[\|x^{k+1} - x^*\|^2\right] + \left(1 - \frac{\tau}{n}\right)\mathcal{W}^k. \tag{11}
$$

Доказательство. Для каждого i переменная w_i^{k+1} равна x^{k+1} с вероятностью $\frac{\tau}{n}$ и равна w_i^k с вероятностью $1 - \frac{\tau}{n}$, что завершает доказательство. □

Теорема 1. *Для случайных переменных Алгоритма 1 верна рекурсия*

$$
\mathbb{E}_k\left[\mathcal{W}^{k+1}\right] \leq \left(1 - \frac{\tau}{n} + \frac{\tau}{n}\frac{H^2}{4\mu^2}cW^k\right)\mathcal{W}^k. \tag{12}
$$

Более того, если $\|w_i^0 - x^*\| < \frac{\mu}{H}$ для всех $i \in \{1, 2, \ldots, n\}$, тогда

$$
\mathbb{E}_k\left[\mathcal{W}^{k+1}\right] \leq \left(1 - \frac{3\tau}{4n}\right)\mathcal{W}^k. \tag{13}
$$

Как следствие, если $\tau = n$, то Алгоритм 1 имеет локальную квадратичную сходимость, как и стандартный метод Ньютона. Если $\tau = 1$, то мы получаем локальную линейную сходимость, не зависящую от обусловленности функции, т. е. метод адаптируется к кривизне функции.

Доказательство. Используя Лемму 2 и Лемму 4, имеем

$$\begin{aligned}
\mathbb{E}_k\left[\mathcal{W}^{k+1}\right] &\overset{(4)}{=} \frac{\tau}{n}\|x^{k+1}-x^*\|^2 + \left(1-\frac{\tau}{n}\right)\mathcal{W}^k \\
&\overset{(2)}{\leq} \frac{\tau}{n}\frac{H^2}{4\mu^2}\left(\mathcal{W}^k\right)^2 + \left(1-\frac{\tau}{n}\right)\mathcal{W}^k \\
&= \left(1-\frac{\tau}{n}+\frac{\tau}{n}\frac{H^2}{4\mu^2}\mathcal{W}^k\right)\mathcal{W}^k \\
&\overset{(3)}{\leq} \left(1-\frac{3\tau}{4n}\right)\mathcal{W}^k,
\end{aligned}$$

где на последнем шаге использовано предположение на ограниченность на $\|w_i^0 - x^*\|$ для всех $i \in \{1, 2, \ldots, n\}$. □