Градиентный метод.
Семинар 2. 18 февраля 2020 г.

Семинарист: Данилова М.

Градиентный метод

Безусловная минимизация.

Метод градиентного спуска применяется для минимизации дифференцируемых функций $f(x)$ на $\mathbb{R}^n$.

В качестве направления h_k берется антиградиент функции, то есть $h_k = -\nabla f(x_k)$.

Шаг α_k выбирается по одному из указанных выше способов.

Algorithm 1 Градиентный метод

1: Пусть $x_0 \in \mathbb{R}^n$, $h_k = -\nabla f(x_k)$
2: Вычислим $x_{k+1} = x_k - \alpha_k \nabla f(x_k) \quad k = 0, 1, ..$

Выбор направления $h_k = -\nabla f(x_k)$

Направление антиградиента - лучшее направление с точки зрения линейной аппроксимации.

Пусть h задает некое направление в пространстве $\mathbb{R}^n$, $\|h\| = 1$. Рассмотрим производную по направлению h:

$$f'(x, h) = \lim_{\alpha \to 0} \frac{f(x + \alpha h) - f(x)}{\alpha}.$$

Рассмотрим линейную аппроксимацию:

$$f(x + \alpha h) = f(x) + \alpha \langle \nabla f(x), h \rangle + o(\alpha).$$

Используя неравенство Коши-Буняковского

$$-\|x\|\|y\| \leq \langle x, y \rangle \leq \|x\|\|y\|$$

получаем, что

$$f'(x, h) = \langle \nabla f(x), h \rangle \geq -\|\nabla f(x)\|.$$

$$h = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$$

Таким образом, направление $-\nabla f(x)$ (антиградиент) является направлением наискорейшего локального убывания функции $f(x)$.

Сходимость градиентного метода (Поляк)

Теорема 1. Пусть $f(x)$ дифференцируема на $\mathbb{R}^n$, градиент $f(x)$ удовлетворяет условию Липшица:

$$\|\nabla f(x) - \nabla f(y)\| \leq \mathrm{L}\|x - y\|,$$

$f(x)$ ограничена снизу:

$$f(x) \geq f^* > -\infty$$

$\alpha_k = \alpha$ и удовлетворяет условию

$$0 < \alpha < \frac{2}{\mathrm{L}}.$$

Тогда в методе градиентного спуска:

$$\lim_{k\to\infty} \nabla f(x_k) = 0.$$

а функция $f(x)$ монотонно убывает: $f(x_{k+1}) \leq f(x_k)$.

Теорема 2. Пусть $f(x)$ дифференцируема на $\mathbb{R}^n$, градиент $f(x)$ удовлетворяет условию Липшица с константой L и $f(x)$ является сильно выпуклой функцией с константой μ. Тогда при $0 < \alpha < \frac{2}{\mathrm{L}}$ метод градиентного спуска сходится к единственной точке глобального минимума x^* со скоростью геометрической прогрессии:

$$\|x_k - x^*\| \leq Cq^k,\ 0 \leq q \leq 1.$$

Теорема 3. Пусть $f(x)$ дважды дифференцируема на $\mathbb{R}^n$ и

$$\mu I \leq \nabla^2 f(x) \leq \mathrm{L}I,\ \mu > 0,$$

для всех x. Тогда при $0 < \alpha < \frac{2}{\mathrm{L}}$

$$\|x_k - x^*\| \leq \|x_0 - x^*\|q^k,$$

$$q = \max\{|1 - \alpha\mu|, |1 - \alpha\mathrm{L}|\} < 1.$$

Величина q минимальна и равна

$$q^* = \frac{\mathrm{L} - \mu}{\mathrm{L} + \mu} \quad \text{при} \quad \alpha = \alpha^* = \frac{2}{\mathrm{L} + \mu}.$$

Замечания

1. Оценка скорости сходимости, даваемая теоремой 3 точная, она достигается для любой квадратичной функции;
2. $$q^* = \frac{L - \mu}{L + \mu} = \frac{\frac{L}{\mu} - 1}{\frac{L}{\mu} + 1} = \frac{\kappa - 1}{\kappa + 1} = 1 - \frac{2}{\kappa + 1},$$

   где $\kappa = \frac{L}{\mu}$ - число обусловленности матрицы $\nabla^2 f(x)$.

   При $\kappa \to \infty \quad q^* \to 1$ - медленная скорость сходимости.

Метод наискорейшего спуска

Если в общем градиентном методе на каждом шаге выбирать α_k по следующему правилу:

$$\alpha_k = \arg\min_{\alpha \geq 0} \phi_k(\alpha) \qquad \phi_k(\alpha) = f(x_k - \alpha \nabla f(x_k))$$

При этом мы получаем **метод наискорейшего спуска.**

Теорема 4. Пусть $f(x)$ - непрерывно дифференцируемая функция и $\{x : \ f(x) \leq f(x_0)\}$ ограничено. Тогда в методе наискорешего спуска $\nabla f(x_k) \to 0$ и у последовательности x_k существуют предельные точки, каждая из которых стационарна, т.е. найдется подпоследовательность $x_{k_i} \to x^*$ и $\nabla f(x^*) = 0$.

По сравнению с теоремой 1 здесь условие Липшица на градиент удается заменить более слабым требованием непрерывности градиента. Это естественно, поскольку способ выбора длины шага в наискорейшем спуске является более гибким, чем $\alpha_k = \alpha$.

Упражнения

1. Сделать шаг методом наискорейшего спуска

$$\min_{x \in R^n} \left(-e^{-x^T x} \right)$$

$$x_0 \in \mathbb{R}^n$$

2. Найти $\alpha_k \, (h_k, x_k) = ?$

$$f(x) = \frac{1}{2} x^T A x + b^T x$$

3. Подробно разберите поведение градиентного метода с постоянным шагом для функций на $\mathbb{R}^1$:

   - $|x|^{1+\alpha}, \ 0 < \alpha < 1;$
   - $|x|^{2+\alpha}, \ \alpha > 0;$
   - $x^2;$
   - $(1 + x^2)^{-1}.$

   При каких x_0, γ метод сходится, при каких расходится?

Скорость сходимости градиентного спуска (Семинар)

Липшицевы функции

Введём в рассмотрение следующий класс функций, который достаточно интересен с точки зрения численных методов оптимизации, а именно, *функции с Липшицевым градиентом.*

Определение 1. Будем говорить, что у дифференцируемой функции $f : Q \to \mathbb{R}$, $Q \subseteq \mathbb{R}^n$ *градиент Липшицев относительно нормы* $\|\cdot\|_2$ *(обычной евклидовой нормы) с константой* L, если

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y \in Q.$$

Оказывается, что у класса функций, имеющих Липшицев градиент, есть простая геометрическая интерпретация, о чём нам говорит следующая теорема.

Теорема 5. Пусть функция $f : Q \to \mathbb{R}$ имеет Липшицев градиент с константой L относительно нормы $\|\cdot\|_2$, а множество Q является выпуклым. Тогда

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2 \quad \forall x, y \in Q. \tag{1}$$

Доказательство. Рассмотрим две произвольные точки $x, y \in Q$. Так как множество Q является выпкулым, то для любого $\tau \in [0, 1]$ точка $x + \tau(y - x)$ принадлежит множеству Q (иными словами, множество Q вместе с любыми двумя точками содержит и отрезок, их соединяющий). Рассмотрим функцию $\varphi(\tau) = f(x + \tau(y - x))$. Функция $\varphi(\tau)$ дифференцируема по τ на $[0, 1]$ и $\varphi'(\tau) = \langle \nabla f(x + \tau(y - x)), y - x\rangle$. Следовательно,

$$\begin{aligned}
f(y) - f(x) &= \varphi(1) - \varphi(0) = \int_0^1 \langle \nabla f(x + \tau(y - x)), y - x\rangle d\tau \\
&= \int_0^1 \Big(\langle \nabla f(x + \tau(y - x)), y - x\rangle - \underbrace{\langle \nabla f(x), y - x\rangle + \langle \nabla f(x), y - x\rangle}_{\text{не зависит от } \tau}\Big) d\tau \\
&= \langle \nabla f(x), y - x\rangle + \int_0^1 \langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x\rangle d\tau \\
&\overset{\text{К.-Б.}}{\leq} \langle \nabla f(x), y - x\rangle + \int_0^1 \|\nabla f(x + \tau(y - x)) - \nabla f(x)\|_2 \|y - x\|_2 d\tau \\
&\overset{\text{Липш.}}{\leq} \langle \nabla f(x), y - x\rangle + \int_0^1 L\|x + \tau(y - x) - x\|_2 \|y - x\|_2 d\tau \\
&\leq \langle \nabla f(x), y - x\rangle + \int_0^1 L\|y - x\|_2^2 \tau d\tau = \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2,
\end{aligned}$$

где К.-Б. означает, что переход справедлив в силу неравенства Коши-Буняковского: $\forall a, b \in \mathbb{R}^n \to \langle a, b\rangle \leq \|a\|_2\|b\|_2$. □

Доказанная теорема доказывает следующую геометрическую интерпретацию функций с Липшицевым градиентом: это такие функции, которые в каждой точке можно оценить сверху некоторым парабалоидом (если рассматривать график функции f как множество в $\mathbb{R}^{n+1}$), причём оценить на всём множестве Q. Другие интересные свойства функций с Липшицевым градиентом (и не только) можно прочитать в книге Ю. Е. Нестерова "Введение в выпуклую оптимизацию" (глава 2, §2.1.1).

Геометрия градиентного спуска

Теперь рассмотрим задачу оптимизации выпуклой на $\mathbb{R}^n$ функции f с Липшицевым градиентом с константой L:

$$f(x) \to \min_{x\in\mathbb{R}^n}$$

Рассмотрим градиентный спуск с постоянным шагом равным $\frac{1}{L}$ (здесь x^0 — стартовая точка, N — число итераций).

1: **procedure** GRADIENT_DESCENT(f, x^0, N)
2: **for** $k = 0, 1, 2, \ldots, N-1$ **do**
3: $x^{k+1} := x^k - \frac{1}{L}\nabla f(x^k)$
4: **end for**
5: **return** x^N
6: **end procedure**

Рассмотрим следующую геометрическую интерпретацию градиентного спуска для функций с Липшицевым градиентом. Допустим мы построили k-е приближение точки минимума x^* функции $f(x)$ и хотим построить следующее приближение. Как это сделать? Один из способов следующий. Давайте запишем верхнюю квадратичную аппроксимацию (1) в точке x^k (возьмём в неравенстве (1) $x = x^k$)

$$f(y) \leq f(x^k) + \langle \nabla f(x^k), y - x^k\rangle + \frac{L}{2}\|y - x^k\|_2^2$$

и попробуем построить x^{k+1} путём минимизации по y правой части предыдущего неравенства. Иными словами, пусть

$$x^{k+1} = \operatorname*{argmin}_{y\in\mathbb{R}^n}\left\{f(x^k) + \langle \nabla f(x^k), y - x^k\rangle + \frac{L}{2}\|y - x^k\|_2^2\right\}.$$

Преобразуем выражение, стоящее в правой части предыдущего равенства:

$$\begin{aligned}
&\operatorname*{argmin}_{y\in\mathbb{R}^n}\left\{\underbrace{f(x^k)}_{\text{не зависит от } y}+\langle\nabla f(x^k),y-x^k\rangle+\frac{L}{2}\|y-x^k\|_2^2\right\}\\
&=\operatorname*{argmin}_{y\in\mathbb{R}^n}\underbrace{\left\{\underbrace{\frac{1}{2L}\|\nabla f(x^k)\|_2^2}_{\text{не зависит от } y}+\langle\nabla f(x^k),y-x^k\rangle+\frac{L}{2}\|y-x^k\|_2^2\right\}}_{\text{полный квадрат}}\\
&=\operatorname*{argmin}_{y\in\mathbb{R}^n}\left\{\left\|\frac{1}{\sqrt{2L}}\nabla f(x^k)+\sqrt{\frac{L}{2}}(y-x^k)\right\|_2^2\right\}\\
&=x^k-\frac{1}{L}\nabla f(x^k).
\end{aligned}$$

Итак, мы получили, что $x^{k+1}=x^k-\frac{1}{L}\nabla f(x^k)$, что сопадает с формулой для шага градиентного спуска с постоянной длиной шага $\frac{1}{L}$, то есть градиентный спуск на каждом шаге минимизирует квадратичную аппроксимацию (1), записанную относительно точки x^k, то есть он переходит в точку, соответствующую вершине параболоида $f(x^k)+\langle\nabla f(x^k),y-x^k\rangle+\frac{L}{2}\|y-x^k\|_2^2$ (см. рисунок 1).

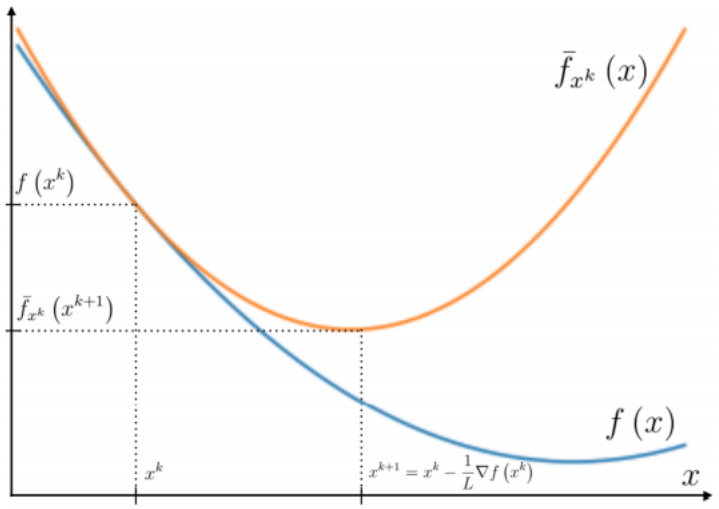

Рис. 1: Геометрия градиентного спуска (источник). Здесь $\overline{f}_{x^k}(x)=f(x^k)+\langle\nabla f(x^k),x-x^k\rangle+\frac{L}{2}\|x-x^k\|_2^2$.

Сходимость градиентного спуска

Теорема 6. Пусть функция f — выпуклая на $\mathbb{R}^n$ с Липшицевым на $\mathbb{R}^n$ градиентом с константой L в евклидовой норме. Тогда процедура `Gradient_Descent`(f, x^0, N) вернёт точку x^N, для которой

$$f(x^N) - f(x^*) \leq \frac{2LR^2}{N}, \tag{2}$$

где x^* — ближайшая точка минимума f к стартовой точке x^0, а $R = \max\limits_{x: f(x) \leq f(x^0)} \|x - x^*\|_2$ (на самом деле, можно показать, что $\|x^k - x^*\|_2 \leq \|x^0 - x^*\|_2$ и в качестве R можно просто брать расстояние до решения из стартовой точки: $R = \|x^0 - x^*\|_2^2$; подробности см. в книге А. В. Гасникова, "Современные численные методы оптимизации. Метод универсального градиентного спуска").

Доказательство. Подставим в неравенство (1) точки $y = x^{k+1} = x^k - \frac{1}{L}\nabla f(x), x = x^k$:

$$\begin{aligned} f(x^{k+1}) - f(x^k) &\leq \langle -\frac{1}{L}\nabla f(x^k), \nabla f(x^k)\rangle + \frac{L}{2}\left\|\frac{1}{L}\nabla f(x^k)\right\|_2^2 \\ &= -\frac{1}{L}\|\nabla f(x^k)\|_2^2 + \frac{1}{2L}\|\nabla f(x^k)\|_2^2 = -\frac{1}{2L}\|\nabla f(x^k)\|_2^2, \end{aligned}$$

откуда получаем, что

$$\|\nabla f(x^k)\|_2^2 \leq 2L(f(x^k) - f(x^{k+1})). \tag{3}$$

В частности из этого неравенства следует, что $f(x^k) \geq f(x^{k+1})$ для всех k, ибо $\|\nabla f(x^k)\|_2^2 \geq 0$. Следовательно, для всех $k \geq 0$ выполнено неравенство $f(x^k) \leq f(x^0)$, а значит, по определению числа R получаем, что $\|x^k - x^*\|_2 \leq R$ для всех k. Поэтому из выпуклости функции f и неравенства Коши-Буняковского получаем неравенства:

$$f(x^k) - f(x^*) \leq \langle \nabla f(x^k), x^k - x^*\rangle \leq \|\nabla f(x^k)\|_2^2 \underbrace{\|x^k - x^*\|_2}_{\leq R} \leq R\|\nabla f(x^k)\|_2.$$

Возведём предыдущее неравенство в квадрат и введём обозначение $D_k = f(x^k) - f(x^*)$. Отсюда и из доказанного неравенства $\|\nabla f(x^k)\|_2^2 \leq 2L(f(x^k) - f(x^{k+1})) \leq 2L(D_k - D_{k+1})$ получаем

$$D_k^2 \leq 2LR^2(D_k - D_{k+1}) \overset{\text{делим на } D_k \cdot D_{k+1}}{\Longrightarrow} \frac{D_k}{D_{k+1}} \leq 2LR^2\left(\frac{1}{D_{k+1}} - \frac{1}{D_k}\right).$$

Отметим, что в силу $f(x^k) \geq f(x^{k+1})$ выполняется неравенство $D_k \geq D_{k+1}$, а значит,

$$2LR^2\left(\frac{1}{D_{k+1}} - \frac{1}{D_k}\right) \geq \frac{D_k}{D_{k+1}} \geq 1,$$

откуда

$$\frac{1}{D_{k+1}} - \frac{1}{D_k} \geq \frac{1}{2LR^2}.$$

Складывая полученные выше неравенства для $k = 0, 1, 2, \ldots, N-1$, получим:

$$\frac{N}{2LR^2} \leq \frac{1}{D_1} - \frac{1}{D_0} + \frac{1}{D_2} - \frac{1}{D_1} + \ldots + \frac{1}{D_N} - \frac{1}{D_{N+1}} = \frac{1}{D_N} - \frac{1}{D_0} \leq \frac{1}{D_N}.$$

Вспоминаем, что $D_N = f(x^N) - f(x^*)$, и из последнего неравенства получаем, что $f(x^N) - f(x^*) \leq \frac{2LR^2}{N}$. □

Следствие 1. Для $N \geq \frac{2LR^2}{\varepsilon}, \varepsilon > 0$ процедура `Gradient_Descent`(f, x^0, N), где f — выпуклая на $\mathbb{R}^n$ функция с Липшицевым на $\mathbb{R}^n$ градиентом с константой Липшица L в евклидовой норме, $R = \max\limits_{x: f(x) \leq f(x^0)} \|x - x^*\|_2$, x^* — ближайшая к x^0 точка минимума функции f, вернёт такую точку x^N, что

$$f(x^N) - f(x^*) \leq \varepsilon.$$

Иными словами, чтобы получить приближение решения с точностью ε по значению функции для функции f, удовлетворяющей описанным выше свойствам, достаточно сделать $O\left(\frac{LR^2}{\varepsilon}\right)$ шагов градиентного спуска.

Сильно выпуклые функции

На прошлых семинарах мы познакомились с определением *сильно выпуклой на* $\mathbb{R}^n$ *функции с параметром* μ:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{\mu}{2}\alpha(1-\alpha)\|x-y\|_2^2 \quad \forall x, y \in \mathbb{R}^n, \alpha \in [0,1]. \tag{4}$$

Для дифференцируемых функций существует эквивалентное определение, через нижнюю квадратичную границу:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y-x\|_2^2 \quad \forall x, y \in \mathbb{R}^n. \tag{5}$$

Следующая теорема говорит о том, что для дифференцируемых функций из (5) следует (4) (на самом деле верно и обратное, но на доказательстве обратного утверждения мы останавливаться не будем).

Теорема 7. Пусть функция f дифференцируема и для неё выполнено условие (5). Тогда для ней выполняется и условие (4).

Доказательство. Рассмотрим две произвольных точки $x, y \in \mathbb{R}^n$ и произвольное число $\alpha \in [0,1]$. Покажем, что выполнено неравенство

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{\mu}{2}\alpha(1-\alpha)\|x-y\|_2^2.$$

Пусть $x_\alpha = \alpha x + (1-\alpha)y$. Запишем условие (5) сначала для пары точек $y = x, x = x_\alpha$, а затем для пары точек $y = y, x = x_\alpha$:

$$\begin{aligned} f(x) &\geq f(x_\alpha) + \langle \nabla f(x_\alpha), (1-\alpha)(x-y) \rangle + \frac{\mu}{2}\|(1-\alpha)(x-y)\|_2^2 \\ f(y) &\geq f(x_\alpha) + \langle \nabla f(x_\alpha), \alpha(y-x) \rangle + \frac{\mu}{2}\|\alpha(x-y)\|_2^2. \end{aligned}$$

Домножим первой неравенство на α, второе — на $1-\alpha$ и сложим полученные неравенства. В итоге слагаемые со скалярным произведением сократятся, т.к. они отличаются только знаком, и получим неравенство

$$\alpha f(x) + (1-\alpha)f(y) \geq \underbrace{(\alpha + 1 - \alpha)}_{=1} f(\alpha x + (1-\alpha)y) + \frac{\mu}{2}\underbrace{\left(\alpha(1-\alpha)^2 + \alpha^2(1-\alpha)\right)}_{=\alpha(1-\alpha)}\|x-y\|_2^2,$$

что и требовалось доказать. □

Докажем ещё один весьма простой и полезный факт о сильно выпуклых функциях.

Теорема 8. Если функция $f : \mathbb{R}^n \to \mathbb{R}$ удовлетворяет условию (5) с константой μ (сильно выпукла с константой μ в норме $\|\cdot\|_2$), то

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2\mu}\|\nabla f(x) - \nabla f(y)\|_2^2 \quad \forall x, y \in \mathbb{R}^n. \tag{6}$$

Доказательство. Рассмотрим произвольные точки $x, y \in \mathbb{R}^n$ и рассмотрим функцию $\varphi(y) = f(y) - \langle \nabla f(x), y\rangle$ для заданного фиксированного x. Нетрудно проверить, что данная функция также удовлетворяет условию (5). Кроме того, $\nabla\varphi(x) = \nabla f(x) = \nabla f(x) = 0$, а значит, из неравенства (5) для функции φ получаем, что для всех точек $y \in \mathbb{R}^n$

$$\varphi(y) \ge \varphi(x) + \frac{\mu}{2}\|y - x\|_2^2,$$

откуда следует, что x — точка минимума функции $\varphi(y)$ (причём единственная). Тогда, пользуясь (5) для функции φ, получим

$$\varphi(x) = \min_{v\in\mathbb{R}^n} \varphi(v) \ge \min_{v\in\mathbb{R}^n} \left\{\varphi(y) + \langle \nabla\varphi(y), v - y\rangle + \frac{\mu}{2}\|v - y\|_2^2\right\}.$$

Теперь распишем правую часть чуть более подробно:

$$\begin{aligned}
&\min_{v\in\mathbb{R}^n} \left\{\varphi(y) + \langle \nabla\varphi(y), v - y\rangle + \frac{\mu}{2}\|v - y\|_2^2\right\} \\
&= \varphi(y) + \min_{v\in\mathbb{R}^n} \left\{\langle \nabla\varphi(y), v - y\rangle + \frac{\mu}{2}\|v - y\|_2^2\right\} + \frac{1}{2\mu}\|\nabla\varphi(y)\|_2^2 - \frac{1}{2\mu}\|\nabla\varphi(y)\|_2^2 \\
&= \varphi(y) + \min_{v\in\mathbb{R}^n} \Bigg\{ \underbrace{\frac{1}{2\mu}\|\nabla\varphi(y)\|_2^2 + \langle \nabla\varphi(y), v - y\rangle + \frac{\mu}{2}\|v - y\|_2^2}_{\text{полный квадрат}} \Bigg\} - \frac{1}{2\mu}\|\nabla\varphi(y)\|_2^2 \\
&= \varphi(y) + \underbrace{\min_{v\in\mathbb{R}^n} \left\{ \left\| \sqrt{\frac{1}{2\mu}}\nabla\varphi(y) + \sqrt{\frac{\mu}{2}}(v - y)\right\|_2^2 \right\}}_{=0,\ \text{при } v = y - \frac{1}{\mu}\nabla\varphi(y)} - \frac{1}{2\mu}\|\nabla\varphi(y)\|_2^2 \\
&= \varphi(y) - \frac{1}{2\mu}\|\nabla\varphi(y)\|_2^2,
\end{aligned}$$

откуда получаем, что

$$\varphi(x) \ge \varphi(y) - \frac{1}{2\mu}\|\nabla\varphi(y)\|_2^2.$$

Если теперь подставить $\varphi(y) = f(y) - \langle \nabla f(x), y\rangle$, то получится неравенство (6). □

Следствие 2. В условиях Теоремы 8 выполняется следующее утверждение: если x^* — точка минимума функции f, то

$$\|\nabla f(x)\|_2^2 \ge 2\mu(f(x) - f(x^*)) \quad \forall x \in \mathbb{R}^n. \tag{7}$$

Доказательство. Чтобы доказать это неравенство, достаточно подставить в неравенство (6) точки $x = x^*$ и $y = x$ и учесть, что градиент в решении равен нулю. □