Гладкость. Градиентный спуск

Методы оптимизации

Александр Безносиков

Московский физико-технический институт

19 сентября 2024

ФПМИ

Гладкость: определение

Определение L-гладкой функции

Пусть дана непрерывно дифференцируемая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что данная функция имеет L-Липшицев градиент (говорить, что она является L-гладкой), если для любых $x, y \in \mathbb{R}^d$ выполнено

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2.$$

Гладкость: определение

Определение L-гладкой функции

Пусть дана непрерывно дифференцируемая на $\mathbb{R}^d$ функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что данная функция имеет L-Липшицев градиент (говорить, что она является L-гладкой), если для любых $x, y \in \mathbb{R}^d$ выполнено

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2.$$

Определение L-гладкости можно писать и в не евклидовой норме. Поэтому формально в предыдущем определении можно указывать, что имеется в виду L-гладкость в терминах $\|\cdot\|_2$.

Гладкость: свойство

Теорема (свойство L - гладкой функции)

Пусть дана L - гладкая функция $f : \mathbb{R}^d \to \mathbb{R}$. Тогда для любых $x, y \in \mathbb{R}^d$ выполнено

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

Гладкость: свойство

Доказательство

Начнем с формулы Ньютона-Лейбница

$$f(y) - f(x) = \int\limits_0^1 \langle \nabla f(x + \tau(y - x)), y - x \rangle d\tau$$

$$= \langle \nabla f(x), y - x \rangle + \int\limits_0^1 \langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x \rangle d\tau$$

Гладкость: свойство

Доказательство

Начнем с формулы Ньютона-Лейбница

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + \tau(y-x)), y - x \rangle d\tau$$

$$= \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x + \tau(y-x)) - \nabla f(x), y - x \rangle d\tau$$

Тогда

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| = \left| \int_0^1 \langle \nabla f(x + \tau(y-x)) - \nabla f(x), y - x \rangle d\tau \right|$$

$$\leq \int_0^1 |\langle \nabla f(x + \tau(y-x)) - \nabla f(x), y - x \rangle| d\tau$$

Гладкость: свойство

Доказательство

Применим КБШ:

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq \int\limits_0^1 |\langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x\rangle| d\tau$$

$$\leq \int\limits_0^1 \|\nabla f(x + \tau(y - x)) - \nabla f(x)\|_2 \|y - x\|_2$$

Гладкость: свойство

Доказательство

Применим КБШ:

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq \int\limits_0^1 |\langle \nabla f(x + \tau(y - x)) - \nabla f(x), y - x\rangle| d\tau$$

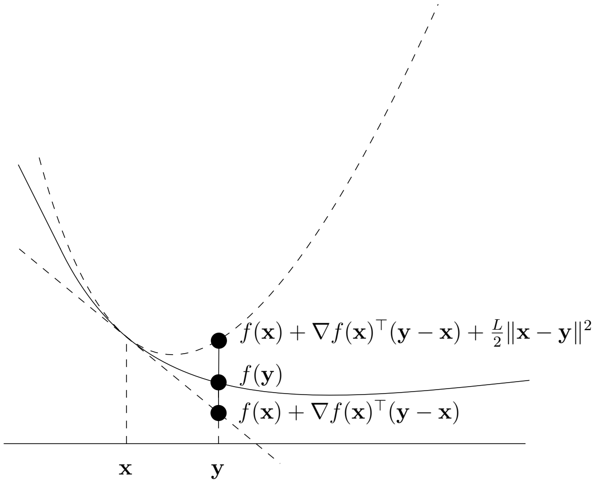
$$\leq \int\limits_0^1 \|\nabla f(x + \tau(y - x)) - \nabla f(x)\|_2 \|y - x\|_2$$

Далее определение L-гладкости:

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq L\|y - x\|_2^2 \int\limits_0^1 \tau d\tau$$

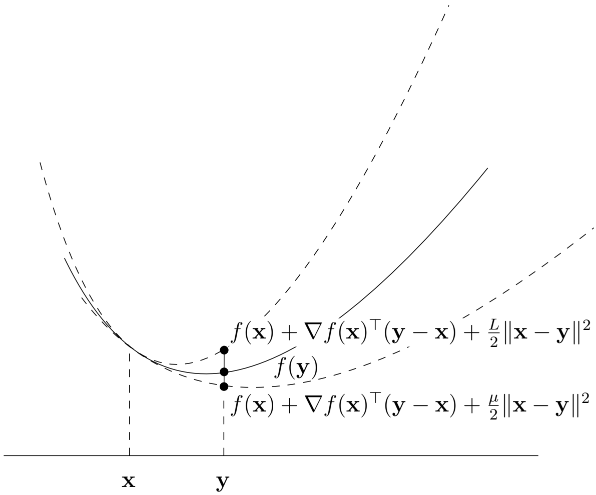
$$= \frac{L}{2}\|x - y\|_2^2$$

Гладкость: физический смысл



Ограничение сверху на поведение (рост) – растет не слишком быстро.

Гладкость: физический смысл



Градиентный спуск

- **Задача:** найти решение безусловной оптимизации:

$$\min_{x \in \mathbb{R}^d} f(x). \tag{1}$$

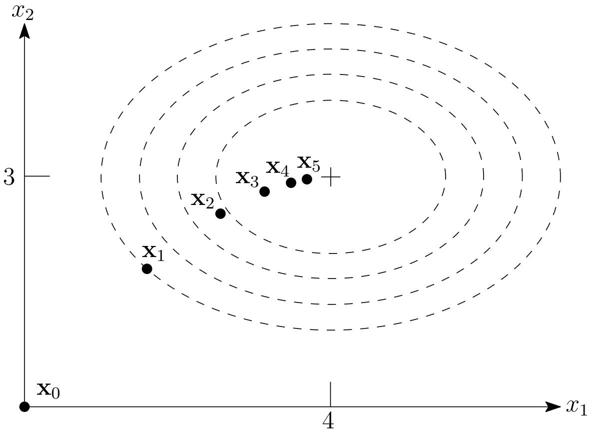
Алгоритм 1 Градиентный спуск

Вход: размеры шагов $\{\gamma_k\}_{k=0} > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $\nabla f(x^k)$
3: $x^{k+1} = x^k - \gamma_k \nabla f(x^k)$
4: **end for**

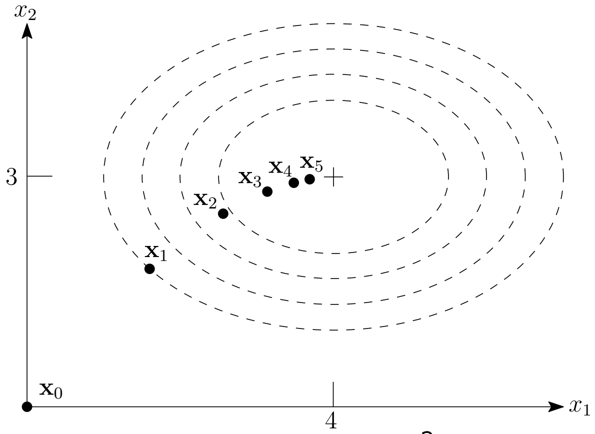
Выход: x^K

Пример



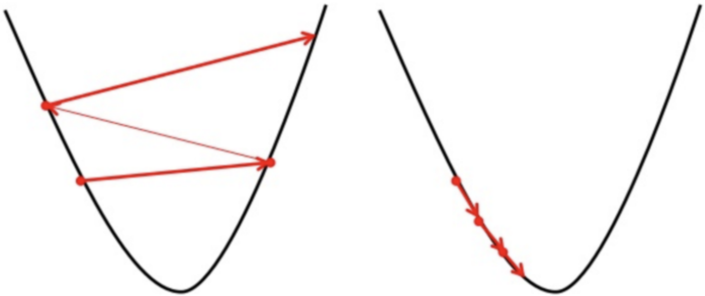
Вопрос: куда направлен градиент в точке x_1?

Пример



Вопрос: куда направлен градиент в точке x_1? направление роста

Зачем нужен шаг?



Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Знаем, что для сильно выпуклых функций решение уникально, попытаемся оценить, как меняется расстояние до него. Подставим итерацию:

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 &= \|x^k - \gamma_k \nabla f(x^k) - x^*\|_2^2 \\
&= \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k)\|_2^2
\end{aligned}
$$

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Знаем, что для сильно выпуклых функций решение уникально, попытаемся оценить, как меняется расстояние до него. Подставим итерацию:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|x^k - \gamma_k \nabla f(x^k) - x^*\|_2^2 \\ &= \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k)\|_2^2\end{aligned}$$

Вопрос: что дальше?

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Знаем, что для сильно выпуклых функций решение уникально, попытаемся оценить, как меняется расстояние до него. Подставим итерацию:

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 &= \|x^k - \gamma_k \nabla f(x^k) - x^*\|_2^2 \\
&= \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k)\|_2^2
\end{aligned}
$$

Вопрос: что дальше? Вспоминаем, что у нас есть гладкость $\|\nabla f(x) + \nabla f(y)\|_2^2 \leq L^2 \|x - y\|_2^2$ и сильная выпуклость в виде $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2$.

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Знаем, что для сильно выпуклых функций решение уникально, попытаемся оценить, как меняется расстояние до него. Подставим итерацию:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|x^k - \gamma_k \nabla f(x^k) - x^*\|_2^2 \\ &= \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k)\|_2^2\end{aligned}$$

Вопрос: что дальше? Вспоминаем, что у нас есть гладкость $\|\nabla f(x) + \nabla f(y)\|_2^2 \le L^2 \|x - y\|_2^2$ и сильная выпуклость в виде $\langle \nabla f(x) - \nabla f(y), x - y \rangle \ge \mu \|x - y\|_2^2$. Достаточно только вспомнить условие оптимальности $\nabla f(x^*) = 0$.

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Знаем, что для сильно выпуклых функций решение уникально, попытаемся оценить, как меняется расстояние до него. Подставим итерацию:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|x^k - \gamma_k \nabla f(x^k) - x^*\|_2^2 \\ &= \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k)\|_2^2\end{aligned}$$

Вопрос: что дальше? Вспоминаем, что у нас есть гладкость $\|\nabla f(x) + \nabla f(y)\|_2^2 \leq L^2 \|x - y\|_2^2$ и сильная выпуклость в виде $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2$. Достаточно только вспомнить условие оптимальности $\nabla f(x^*) = 0$.

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 = &\|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^* \rangle \\ &+ \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2\end{aligned}$$

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Гладкость $\|\nabla f(x) + \nabla f(y)\|_2^2 \leq L^2\|x-y\|_2^2$ и сильная выпуклость в виде $\langle \nabla f(x) - \nabla f(y), x-y\rangle \geq \mu\|x-y\|_2^2$:

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

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Гладкость $\|\nabla f(x) + \nabla f(y)\|_2^2 \leq L^2\|x-y\|_2^2$ и сильная выпуклость в виде $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x-y\|_2^2$:

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

Вопрос: а что мы хотим теперь?

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Гладкость $\|\nabla f(x) + \nabla f(y)\|_2^2 \leq L^2\|x - y\|_2^2$ и сильная выпуклость в виде $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x - y\|_2^2$:

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

Вопрос: а что мы хотим теперь? $(1 - 2\gamma_k\mu + \gamma_k^2 L^2) < 1$. Как подобрать?

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Гладкость $\|\nabla f(x) + \nabla f(y)\|_2^2 \leq L^2\|x-y\|_2^2$ и сильная выпуклость в виде $\langle \nabla f(x) - \nabla f(y), x-y\rangle \geq \mu\|x-y\|_2^2$:

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

Вопрос: а что мы хотим теперь? $(1 - 2\gamma_k\mu + \gamma_k^2L^2) < 1$. Как подобрать? $\arg\min_{\gamma_k}(1 - 2\gamma_k\mu + \gamma_k^2L^2)$?

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Гладкость $\|\nabla f(x) + \nabla f(y)\|_2^2 \leq L^2 \|x - y\|_2^2$ и сильная выпуклость в виде $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2$:

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

Вопрос: а что мы хотим теперь? $(1 - 2\gamma_k \mu + \gamma_k^2 L^2) < 1$. Как подобрать? $\arg\min_{\gamma_k} (1 - 2\gamma_k \mu + \gamma_k^2 L^2)$? $\gamma_k = \frac{\mu}{L^2}$ и $(1 - 2\gamma_k \mu + \gamma_k^2 L^2) = 1 - \frac{\mu^2}{L^2}$.

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Итого:

$$\|x^{k+1} - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right) \|x^k - x^*\|_2^2$$

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Итого:

$$\|x^{k+1} - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right) \|x^k - x^*\|_2^2$$

Запустим рекурсию:

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^K \|x^0 - x^*\|_2^2$$

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Итого:

$$\|x^{k+1} - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right) \|x^k - x^*\|_2^2$$

Запустим рекурсию:

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^K \|x^0 - x^*\|_2^2$$

С геометрической (геом. прогрессии) скоростью приближаемся к решению. Формализуем понятия скоростей сходимости.

Скорости сходимости/приближения к решению

- Сублинейная:

$$\|x^k - x^*\| \leq \frac{C}{k^\alpha},$$

где $\alpha > 0$, $C > 0$ – константа.

- Линейная:

$$\|x^k - x^*\| \leq Cq^k,$$

где $q \in (0; 1)$, $C > 0$ – константа.

- Сверхлинейная:

$$\|x^k - x^*\| \leq Cq^{k^p},$$

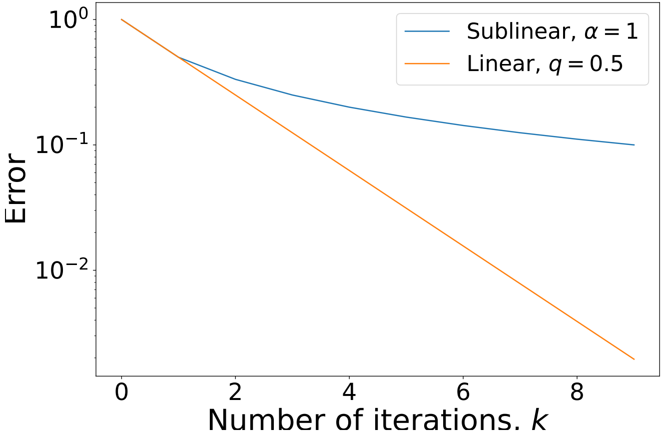
где $q \in (0; 1)$, $p > 1$, $C > 0$ – константа.

- Квадратичная:

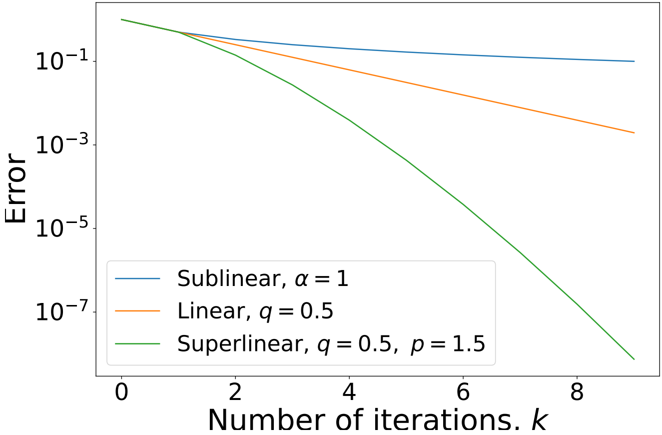
$$\|x^k - x^*\| \leq Cq^{2^k} \quad \text{или} \quad \|x^{k+1} - x^*\| \leq C\|x^k - x^*\|^2,$$

где $q \in (0; 1)$, $C > 0$ – константа.

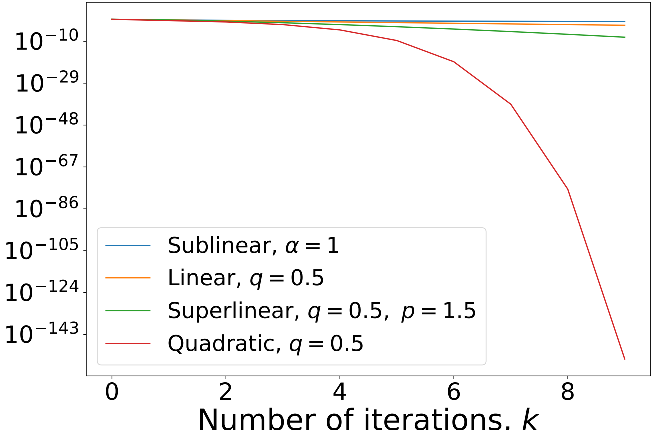
Скорости сходимости/приближения к решению



Скорости сходимости/приближения к решению



Скорости сходимости/приближения к решению



Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Возвращаемся с градиентному спуску:

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^K \|x^0 - x^*\|_2^2$$

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Возвращаемся с градиентному спуску:

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^K \|x^0 - x^*\|_2^2$$

Вопрос: какая это скорость сходимости?

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Возвращаемся с градиентному спуску:

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^K \|x^0 - x^*\|_2^2$$

Вопрос: какая это скорость сходимости? Линейная.

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Возвращаемся с градиентному спуску:

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^K \|x^0 - x^*\|_2^2$$

Вопрос: какая это скорость сходимости? Линейная. А как получить оценку на число итераций?

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Возвращаемся с градиентному спуску:

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^K \|x^0 - x^*\|_2^2$$

Вопрос: какая это скорость сходимости? Линейная. А как получить оценку на число итераций? (Здесь просто нужно вспомнить разложение экспоненты в ряд)

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^K \|x^0 - x^*\|_2^2 \leq \exp\left(-\frac{\mu^2}{L^2} \cdot K\right) \|x^0 - x^*\|_2^2$$

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Возвращаемся с градиентному спуску:

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^K \|x^0 - x^*\|_2^2$$

Вопрос: какая это скорость сходимости? Линейная. А как получить оценку на число итераций? (Здесь просто нужно вспомнить разложение экспоненты в ряд)

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu^2}{L^2}\right)^K \|x^0 - x^*\|_2^2 \leq \exp\left(-\frac{\mu^2}{L^2} \cdot K\right) \|x^0 - x^*\|_2^2$$

Мы хотим, чтобы гарантированно

$$\|x^K - x^*\|_2^2 \leq \exp\left(-\frac{\mu^2}{L^2} \cdot K\right) \|x^0 - x^*\|_2^2 \leq \varepsilon^2$$

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Тогда логарифируем и получаем

$$K \geq \frac{L^2}{\mu^2} \log \left(\frac{\|x^0 - x^*\|_2^2}{\varepsilon^2} \right)$$

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Тогда логарифмируем и получаем

$$K \geq \frac{L^2}{\mu^2} \log \left(\frac{\|x^0 - x^*\|_2^2}{\varepsilon^2} \right)$$

Итого: Not great, not terrible – можно лучше. Пример того, как в получении верхних оценок можно «загрубить». Будем исправлять.

Гладкость: свойства

Теорема (свойства L - гладкой выпуклой функции)

Пусть дана L - гладкая *выпуклая* функция $f : \mathbb{R}^d \to \mathbb{R}$. Тогда для любых $x, y \in \mathbb{R}^d$ выполнено

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2} \|x - y\|_2^2$$

и

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

Гладкость: свойства

Теорема (свойства L - гладкой выпуклой функции)

Пусть дана L - гладкая *выпуклая* функция $f : \mathbb{R}^d \to \mathbb{R}$. Тогда для любых $x, y \in \mathbb{R}^d$ выполнено

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2} \|x - y\|_2^2$$

и

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

Доказательство

Доказательство первого факта следует из выпуклости и предыдущего свойства гладкости.

Гладкость: свойства

Доказательство

Рассмотрим функцию $\phi(y) = f(y) - \langle \nabla f(x), y \rangle$. **Вопрос:** является ли она L_ϕ-гладкой? выпуклой?

Гладкость: свойства

Доказательство

Рассмотрим функцию $\phi(y) = f(y) - \langle \nabla f(x), y \rangle$. **Вопрос:** является ли она L_ϕ-гладкой? выпуклой? Да на оба вопроса и $L_\phi = L$ (проверка по определению).

Гладкость: свойства

Доказательство

Рассмотрим функцию $\phi(y) = f(y) - \langle \nabla f(x), y \rangle$. **Вопрос:** является ли она L_ϕ-гладкой? выпуклой? Да на оба вопроса и $L_\phi = L$ (проверка по определению). Также можно заметить, что $y^* = x$ – минимум.
Вопрос: почему?

Гладкость: свойства

Доказательство

Рассмотрим функцию $\phi(y) = f(y) - \langle \nabla f(x), y \rangle$. **Вопрос:** является ли она L_ϕ-гладкой? выпуклой? Да на оба вопроса и $L_\phi = L$ (проверка по определению). Также можно заметить, что $y^* = x$ – минимум.
Вопрос: почему? $\nabla \phi(y^*) = \nabla \phi(x) = 0$. Воспользуемся первым пунктом теоремы: $f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}\|x - y\|_2^2$ с $\left(y = y - \frac{1}{L}\nabla \phi(y), x = y, f = \phi \right)$. Тогда

$$\phi\left(y - \frac{1}{L}\nabla \phi(y)\right) - \phi(y) - \left\langle \nabla \phi(y), -\frac{1}{L}\nabla \phi(y) \right\rangle \leq \frac{1}{2L}\|\nabla \phi(y)\|_2^2$$

После небольшой перестановки:

$$\phi\left(y - \frac{1}{L}\nabla \phi(y)\right) \leq \phi(y) - \frac{1}{2L}\|\nabla \phi(y)\|_2^2$$

Гладкость: свойства

Доказательство

Тогда получаем, зная, что $y^* = x$ – минимум:

$$\phi(x) = \phi(y^*) \leq \phi\left(y - \frac{1}{L}\nabla\phi(y)\right) \leq \phi(y) - \frac{1}{2L}\|\nabla\phi(y)\|_2^2$$

Гладкость: свойства

Доказательство

Тогда получаем, зная, что $y^* = x$ – минимум:

$$\phi(x) = \phi(y^*) \leq \phi\left(y - \frac{1}{L}\nabla\phi(y)\right) \leq \phi(y) - \frac{1}{2L}\|\nabla\phi(y)\|_2^2$$

Подставляя ϕ:

$$f(x) - \langle \nabla f(x), x\rangle \leq f(y) - \langle \nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

Гладкость: свойства

Доказательство

Тогда получаем, зная, что $y^* = x$ – минимум:

$$\phi(x) = \phi(y^*) \leq \phi\left(y - \frac{1}{L}\nabla\phi(y)\right) \leq \phi(y) - \frac{1}{2L}\|\nabla\phi(y)\|_2^2$$

Подставляя ϕ:

$$f(x) - \langle \nabla f(x), x\rangle \leq f(y) - \langle \nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

Осталось переставить:

$$f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

Гладкость: свойства

Доказательство

Тогда получаем, зная, что $y^* = x$ – минимум:

$$\phi(x) = \phi(y^*) \leq \phi\left(y - \frac{1}{L}\nabla\phi(y)\right) \leq \phi(y) - \frac{1}{2L}\|\nabla\phi(y)\|_2^2$$

Подставляя ϕ:

$$f(x) - \langle\nabla f(x), x\rangle \leq f(y) - \langle\nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

Осталось переставить:

$$f(x) + \langle\nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

Вопрос: а пользовались вообще здесь выпуклость?

Гладкость: свойства

Доказательство

Тогда получаем, зная, что $y^* = x$ – минимум:

$$\phi(x) = \phi(y^*) \leq \phi\left(y - \frac{1}{L}\nabla\phi(y)\right) \leq \phi(y) - \frac{1}{2L}\|\nabla\phi(y)\|_2^2$$

Подставляя ϕ:

$$f(x) - \langle\nabla f(x), x\rangle \leq f(y) - \langle\nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

Осталось переставить:

$$f(x) + \langle\nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

Вопрос: а пользовались вообще здесь выпуклость? да, $\nabla\phi(y^*) = 0 \;\Rightarrow\; y^*$ – минимум.

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Стартуем аналогично:

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 =& \|x^k - \gamma_k \nabla f(x^k) - x^*\|_2^2 \\
=& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \\
=& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle \\
&+ \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2
\end{aligned}
$$

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Стартуем аналогично:

$$\begin{aligned}
\|x^{k+1} - x^*\|_2^2 =& \|x^k - \gamma_k \nabla f(x^k) - x^*\|_2^2 \\
=& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \\
=& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle \\
& + \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2
\end{aligned}$$

Но сделаем тоньше. Сильная выпуклость в виде:
$-\langle \nabla f(x), x - y \rangle \leq -\left(\frac{\mu}{2}\|x - y\|_2^2 + f(x) - f(y)\right)$:

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Стартуем аналогично:

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 =& \|x^k - \gamma_k \nabla f(x^k) - x^*\|_2^2 \\
=& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \\
=& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k), x^k - x^* \rangle \\
&+ \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2
\end{aligned}
$$

Но сделаем тоньше. Сильная выпуклость в виде:
$-\langle \nabla f(x), x - y \rangle \leq -\left(\frac{\mu}{2}\|x - y\|_2^2 + f(x) - f(y)\right)$:

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 \leq& \|x^k - x^*\|_2^2 - 2\gamma_k \left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*)\right) \\
&+ \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2
\end{aligned}
$$

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Дальше гладкость, но в виде:
$\|\nabla f(x^k) - \nabla f(x^*)\|_2^2 \leq 2L\left(f(x^k) - f(x^*)\right)$. **Вопрос:** все ли верно в этом свойстве?

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Дальше гладкость, но в виде:
$\|\nabla f(x^k) - \nabla f(x^*)\|_2^2 \leq 2L\left(f(x^k) - f(x^*)\right)$. **Вопрос:** все ли верно в этом свойстве? Да, использовано, что $\nabla f(x^*) = 0$. Получаем

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*)\right) \\
& + 2\gamma_k^2 L(f(x^k) - f(x^*)) \\
= & (1 - \gamma_k \mu)\|x^k - x^*\|_2^2 + 2\gamma_k(\gamma_k L - 1)(f(x^k) - f(x^*))
\end{aligned}
$$

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Дальше гладкость, но в виде:
$\|\nabla f(x^k) - \nabla f(x^*)\|_2^2 \leq 2L\left(f(x^k) - f(x^*)\right)$. **Вопрос:** все ли верно в этом свойстве? Да, использовано, что $\nabla f(x^*) = 0$. Получаем

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*)\right) \\
& + 2\gamma_k^2 L(f(x^k) - f(x^*)) \\
= & (1 - \gamma_k \mu)\|x^k - x^*\|_2^2 + 2\gamma_k(\gamma_k L - 1)(f(x^k) - f(x^*))
\end{aligned}
$$

Вопрос: что осталось?

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

Дальше гладкость, но в виде:
$\|\nabla f(x^k) - \nabla f(x^*)\|_2^2 \leq 2L\left(f(x^k) - f(x^*)\right)$. **Вопрос:** все ли верно в этом свойстве? Да, использовано, что $\nabla f(x^*) = 0$. Получаем

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 - 2\gamma_k\left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*)\right) \\
& + 2\gamma_k^2 L(f(x^k) - f(x^*)) \\
= & (1 - \gamma_k\mu)\|x^k - x^*\|_2^2 + 2\gamma_k(\gamma_k L - 1)(f(x^k) - f(x^*))
\end{aligned}
$$

Вопрос: что осталось? $(\gamma_k L - 1) \leq 0$. А значит $\gamma_k \leq \frac{1}{L}$.

$$
\|x^{k+1} - x^*\|_2^2 \leq (1 - \gamma_k\mu)\|x^k - x^*\|_2^2
$$

Сходимость: L-гладкие и μ-сильно выпуклые функции

Доказательство

С предыдущего слайда:

$$\|x^{k+1} - x^*\|_2^2 \leq (1 - \gamma_k \mu)\|x^k - x^*\|_2^2$$

Запускаем рекурсию:

$$\|x^K - x^*\|_2^2 \leq \prod_{k=0}^{K-1} (1 - \gamma_k \mu)\|x^0 - x^*\|_2^2$$

С постоянным щагом $\gamma_k = \gamma = \frac{1}{L}$:

$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^K \|x^0 - x^*\|_2^2$$

Сходимость: L-гладкие и μ-сильно выпуклые функции

Теорема сходимость градиентного спуска для L-гладких и μ-сильно выпуклых функций

Пусть задача безусловной оптимизации (1) с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью градиентного спуска. Тогда справедлива следующая оценка сходимости

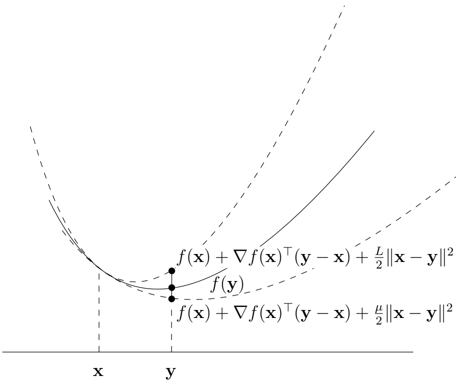
$$\|x^K - x^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^K \|x^0 - x^*\|_2^2.$$

Более того, чтобы добиться точности ε по аргументу, необходимо

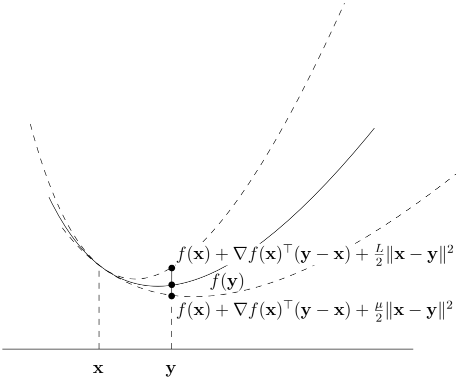
$$K = O\left(\frac{L}{\mu} \log \frac{\|x^0 - x^*\|_2}{\varepsilon}\right) = \tilde{O}\left(\frac{L}{\mu}\right) \text{ итераций.}$$

Мы будем использовать O-нотацию, чтобы "убирать" численные фактор и $\tilde{O}$-нотацию, чтобы убирать еще и log-факторы.

Немного интуиции доказательства



Немного интуиции доказательства



Шагаем, исходя из свойств верхней границы (L) – чтобы гарантированно не "улететь", и перемещаемся в худшем случае, исходя из свойств нижней границы (μ).

Сходимость

	μ-сильно выпуклая	выпуклая	невыпуклая
L-гладкая	$O\left(\frac{L}{\mu}\log\frac{\|x^0-x^*\|_2}{\varepsilon}\right)$	$O\left(\frac{L\|x^0-x^*\|_2^2}{\varepsilon}\right)$	$O\left(\frac{L(f(x^0)-f^*)}{\varepsilon^2}\right)$
M-липшицева	$O\left(\frac{M^2}{\mu^2\varepsilon}\right)$	$O\left(\frac{M^2\|x^0-x^*\|_2^2}{\varepsilon^2}\right)$	2 лекция

- В сильно выпуклом случае по аргументу: $\|x - x^*\|_2 \leq \varepsilon$,
- В выпуклом случае по функции (решение x^* может быть не единственно): $f(x) - f^* \leq \varepsilon$,
- В невыпуклом случае (сходимость к какой-то стационарной точке): $\|\nabla f(x)\|_2 \leq \varepsilon$.

Сходимость

	μ-сильно выпуклая	выпуклая	невыпуклая
L-гладкая	$O\left(\frac{L}{\mu}\log\frac{\|x^0-x^*\|_2}{\varepsilon}\right)$	$O\left(\frac{L\|x^0-x^*\|_2^2}{\varepsilon}\right)$	$O\left(\frac{L(f(x^0)-f^*)}{\varepsilon^2}\right)$
M-липшицева	$O\left(\frac{M^2}{\mu^2\varepsilon}\right)$	$O\left(\frac{M^2\|x^0-x^*\|_2^2}{\varepsilon^2}\right)$	2 лекция

- В сильно выпуклом случае по аргументу: $\|x - x^*\|_2 \leq \varepsilon$,
- В выпуклом случае по функции (решение x^* может быть не единственно): $f(x) - f^* \leq \varepsilon$,
- В невыпуклом случае (сходимость к какой-то стационарной точке): $\|\nabla f(x)\|_2 \leq \varepsilon$.
- Градиентный спуск оптимален (**вопрос:** что это значит?) в негладком случае, а также в гладком невыпуклом.
- Наш анализ градиентного спуска в сильно выпуклом случае неулучшаем с точностью до численных множителей.
- В гладком выпуклом и сильно выпуклом случаях возможны улучшения, но для этого нужен другой метод (4 лекция).

Подбор шага: Поляк-Шор

Уже получали

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*)\right) \\ & + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \\ \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(f(x^k) - f(x^*)\right) + \gamma_k^2 \|\nabla f(x^k)\|_2^2\end{aligned}$$

Подбор шага: Поляк-Шор

Уже получали

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*)\right) \\ & + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \\ \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(f(x^k) - f(x^*)\right) + \gamma_k^2 \|\nabla f(x^k)\|_2^2\end{aligned}$$

Вопрос: как можно подобрать γ_k оптимально в этой ситуации?

Подбор шага: Поляк-Шор

Уже получали

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*) \right) \\ & + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \\ \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(f(x^k) - f(x^*) \right) + \gamma_k^2 \|\nabla f(x^k)\|_2^2\end{aligned}$$

Вопрос: как можно подобрать γ_k оптимально в этой ситуации?
$\arg\min_{\gamma_k} \left(-2\gamma_k \left(f(x^k) - f(x^*) \right) + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \right)$?

Подбор шага: Поляк-Шор

Уже получали

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*)\right) \\ & + \gamma_k^2\|\nabla f(x^k)\|_2^2 \\ \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(f(x^k) - f(x^*)\right) + \gamma_k^2\|\nabla f(x^k)\|_2^2\end{aligned}$$

Вопрос: как можно подобрать γ_k оптимально в этой ситуации? $\arg\min_{\gamma_k} \left(-2\gamma_k \left(f(x^k) - f(x^*)\right) + \gamma_k^2\|\nabla f(x^k)\|_2^2\right)$?

$$\gamma_k = \frac{f(x^k) - f(x^*)}{\|\nabla f(x^k)\|_2^2}$$

Вопрос: какие видите проблемы?

Подбор шага: Поляк-Шор

Уже получали

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(\frac{\mu}{2}\|x^k - x^*\|_2^2 + f(x^k) - f(x^*) \right) \\
& + \gamma_k^2 \|\nabla f(x^k)\|_2^2 \\
\leq & \|x^k - x^*\|_2^2 - 2\gamma_k \left(f(x^k) - f(x^*) \right) + \gamma_k^2 \|\nabla f(x^k)\|_2^2
\end{aligned}
$$

Вопрос: как можно подобрать γ_k оптимально в этой ситуации? $\arg\min_{\gamma_k} \left(-2\gamma_k \left(f(x^k) - f(x^*)\right) + \gamma_k^2 \|\nabla f(x^k)\|_2^2\right)$?

$$
\gamma_k = \frac{f(x^k) - f(x^*)}{\|\nabla f(x^k)\|_2^2}
$$

Вопрос: какие видите проблемы? $f(x^*)$ – иногда известно, а иногда можно оценить.

Подбор шага

- Шаг Поляка-Шора:

$$\gamma_k = \frac{f(x^k) - f(x^*)}{\alpha \|\nabla f(x^k)\|_2^2}, \quad \alpha \geq 1 \quad \text{(надо подбирать)}$$

Подбор шага

- Шаг Поляка-Шора:

$$\gamma_k = \frac{f(x^k) - f(x^*)}{\alpha \|\nabla f(x^k)\|_2^2}, \quad \alpha \geq 1 \quad \text{(надо подбирать)}$$

- Наискорейший спуск:

$$\gamma_k = \arg\min_{\gamma} f(x^k - \gamma \nabla f(x^k))$$

Подбор шага

- Шаг Поляка-Шора:

$$\gamma_k = \frac{f(x^k) - f(x^*)}{\alpha \|\nabla f(x^k)\|_2^2}, \quad \alpha \geq 1 \quad \text{(надо подбирать)}$$

- Наискорейший спуск:

$$\gamma_k = \arg\min_{\gamma} f(x^k - \gamma \nabla f(x^k))$$

Вопрос: как решать?

Подбор шага

- Шаг Поляка-Шора:

$$\gamma_k = \frac{f(x^k) - f(x^*)}{\alpha \|\nabla f(x^k)\|_2^2}, \quad \alpha \geq 1 \quad \text{(надо подбирать)}$$

- Наискорейший спуск:

$$\gamma_k = \arg\min_{\gamma} f(x^k - \gamma \nabla f(x^k))$$

Вопрос: как решать? Иногда есть явная формула, а так нужно решать одномерную задачу.

Подбор шага

- Шаг Поляка-Шора:

$$\gamma_k = \frac{f(x^k) - f(x^*)}{\alpha \|\nabla f(x^k)\|_2^2}, \quad \alpha \geq 1 \quad \text{(надо подбирать)}$$

- Наискорейший спуск:

$$\gamma_k = \arg\min_{\gamma} f(x^k - \gamma \nabla f(x^k))$$

Вопрос: как решать? Иногда есть явная формула, а так нужно решать одномерную задачу.

- Правила Армихо, Вульфа и Гольдстейна.
- Адаптивный подбор, например, онлайн оценка локальной константы L.