Семинар 3

Функции нескольких переменных

Частные производные Пусть теперь у нас есть функция от нескольких переменных $f\colon \mathbb{R}^n \to \mathbb{R}$, то есть правило по вектору $x = (x_1, \ldots, x_n)$ ставящее число $f(x) = f(x_1, \ldots, x_n)$. Пусть теперь у нас задана точка $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ и мы хотим исследовать ее в окрестности этой точки. Первое, что мы можем сделать, давайте зафиксируем все координаты равными координатам точки a, а одну координату сделаем переменной. Например

$$\varphi_i(t) = f(a_1, \ldots, a_{i-1}, t, a_{i+1}, \ldots, a_n)$$

Тогда мы получим обычную функцию от переменной t. Давайте продифференцируем ее по t в точке a_i, результат называется частной производной f по переменной x_i в точке a, и обозначается все это вот так

$$\frac{\partial f}{\partial x_i}(a) = \varphi_i(a_i)'$$

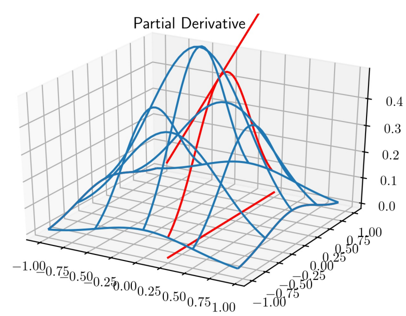
В терминах предела определение пишется так:

$$\frac{\partial f}{\partial x_i}(a) = \lim_{h \to 0} \frac{f(a_1, \ldots, a_{i-1}, a_i + h, a_{i+1}, \ldots, a_n) - f(a_1, \ldots, a_{i-1}, a_i, a_{i+1}, \ldots, a_n)}{h}$$

Пусть теперь у нас есть вектор $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, где единица стоит на i-ом месте. Тогда определение выше можно переписать так

$$\frac{\partial f}{\partial x_i}(a) = \lim_{h \to 0} \frac{f(a + he_i) - f(a)}{h}$$

Давайте поймем геометрический смысл проделанной работы на примере функции двух переменных $f\colon \mathbb{R}^2 \to \mathbb{R}$ по правилу $(x, y) \mapsto f(x, y)$. Пусть $a = (a_1, a_2)$ – некоторая точка, мы зафиксируем a_2, и рассмотрим $\varphi_1(t) = f(t, a_2)$. Заметим, что точки (t, a_2) лежат на прямой параллельной оси OX и пересекающей ось OY в точке a_2. Тогда $\varphi_1(t)$ – это ограничение функции $f(x, y)$ на эту прямую. То есть мы хотим узнать, а как себя ведет данная функция вдоль прямой параллельной OX. Таким образом $\partial f / \partial x(a)$ будет показывать наклон касательной в точке a вдоль прямой параллельной OX и проходящей через a.



Правила вычисления частных производных Если вам надо посчитать частную производную $\frac{\partial f}{\partial x_i}(x)$ от некоторой функции $f\colon \mathbb{R}^n \to \mathbb{R}$, то это делается ровно по тем же самым правилам, что и подсчет обычной производной. Тут надо думать так: вы смотрите на функцию $f(x_1, \ldots, x_i, \ldots, x_n)$ и считаете в ней все переменные кроме x_i какими-то числами и берете обычную производную по x_i как если бы вы дифференцировали по одной переменной.

Единственное полезное правило, которое стоит упомянуть в случае нескольких переменных – дифференцирование композиции. Предположим у нас есть несколько функций $f\colon \mathbb{R}^n \to \mathbb{R}$ по правилу $(y_1, \ldots, y_n) \mapsto f(y_1, \ldots, y_n)$ и $g_1, \ldots, g_n\colon \mathbb{R}^m \to \mathbb{R}$ все по правилам $(x_1, \ldots, x_m) \mapsto g_i(x_1, \ldots, x_m)$. Я специально выбрал разные имена переменным для удобства. Теперь рассмотрим функцию

$$\varphi(x_1, \ldots, x_m) = f(g_1(x_1, \ldots, x_n), \ldots, g_n(x_1, \ldots, x_n))$$

Как надо считать ее частную производную по x_i? Оказывается правило такое:

$$\frac{\partial\varphi}{\partial x_i}(x_1,\dots,x_m)=\sum_{k=1}^{n}\frac{\partial f}{\partial y_k}(g_1(x_1,\dots,x_n),\dots,g_n(x_1,\dots,x_n))\frac{\partial g_k}{\partial x_i}(x_1,\dots,x_m)$$

Или кратко, это правило записывать так

$$\frac{\partial f(g_1,\dots g_n)}{\partial x_i}=\sum_{k=1}^{n}\frac{\partial f}{\partial y_k}\frac{\partial g_k}{\partial x_i}$$

Вот пример, пусть у нас есть функция $f(y_1,y_2)=y_1^2+\sin y_2$ и пусть $g_1(x)=\cos x$ и $g_2(x)=\ln x$. Тогда $\varphi(x)=f(g_1(x),g_2(x))=(\cos x)^2+\sin(\ln x)$. Мы хотим найти производную φ по x. Нам никто не мешает это сделать в лоб и получить

$$\varphi(x)'=-2\cos x\sin x+\cos(\ln x)\frac{1}{x}$$

Однако, это можно сделать и по правилу дифференцирования композиции. Прежде чем считать, обратим внимание, так как функции g_1 и g_2 зависят только от одной переменной, то нет разницы между частной производной и обычной производной. А теперь посчитаем

$$\frac{\partial f}{\partial y_1}=2y_1,\quad \frac{\partial f}{\partial y_2}=\cos y_2,\quad \frac{dg_1}{dx}=-\sin x,\quad \frac{dg_2}{dx}=\frac{1}{x}$$

Если подставить эти выражения в формулу для дифференцирования сложной функции, то получится

$$\varphi(x)'=\frac{\partial f}{\partial y_1}(g_1,g_2)\frac{dg_1}{dx}+\frac{\partial f}{\partial y_2}(g_1,g_2)\frac{dg_2}{dx}=(2\cos x)(-\sin x)+(\cos\ln x)\left(\frac{1}{x}\right)$$

Обратите внимание, когда мы считаем частную производную $\frac{\partial f}{\partial y_1}$, то мы получаем выражение от y_1 и y_2, после этого мы туда вместо y_1 подставляем $g_1(x)$, а вместо y_2 подставляем $g_2(x)$. Аналогично и для другой частной производной.

Производная по направлению Если как и выше мы возьмем функцию $f\colon\mathbb{R}^n\to\mathbb{R}$, некоторую точку $a\in\mathbb{R}^n$ и некоторый вектор $v\in\mathbb{R}^n$, то можно определить производную по направлению следующим образом

$$\nabla_v f(a)=\lim_{h\to0}\frac{f(a+hv)-f(a)}{h}$$

Таким образом, частная производная – это лишь производная в направлении одного из векторов e_i. Смысл производной по направлению следующий. Если мы проведем через точку a прямую в направлении вектора v и ограничим на эту прямую функцию f, то получим функцию одной переменной $\varphi(t)=f(a+tv)$. Вот ее производная в нуле и будет производной по направлению. Это разумно, что мы хотим уметь брать производные в любом направлении, а не только вдоль координатных осей. Мало ли, что нам может пригодиться.

Градиент Как и выше у нас $f\colon\mathbb{R}^n\to\mathbb{R}$ – некоторая функцию и $a\in\mathbb{R}^n$ – произвольная точка. Определим градиент функции f в точке a, как вектор, составленный из частных производных в этой точке:[1]

$$\operatorname{grad}_a f=\left(\frac{\partial f}{\partial x_1}(a),\dots,\frac{\partial f}{\partial x_n}(a)\right)$$

Если я теперь снабжу $\mathbb{R}^n$ стандартным скалярным произведением $(v,u)=v^tu$, то вот есть какая полезная формула для вычисления производной по направлению. Пусть $v\in\mathbb{R}^n$ – произвольный вектор, тогда

$$\nabla_v f(a)=(\operatorname{grad}_a f,v)=\frac{\partial f}{\partial x_1}(a)v_1+\ldots+\frac{\partial f}{\partial x_n}(a)v_n=\sum_{i=1}^{n}\frac{\partial f}{\partial x_i}(a)v_i$$

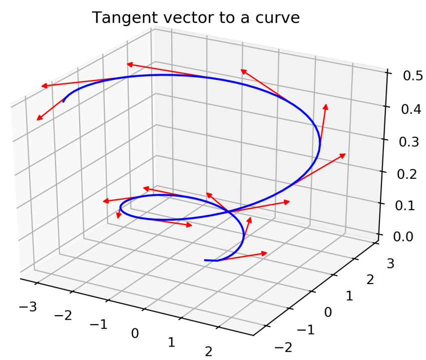
[1] Для тех кому интересно, градиент надо считать вектор-строкой. Почему, ну потому что по-хорошему он является не вектором, а линейным отображением $\operatorname{grad} f\colon\mathbb{R}^n\to\mathbb{R}$, но это уже отдельная песня. И мы на самом деле будем его трактовать то строкой, то столбцом в зависимости от удобства.

Вспомним, что геометрический смысл $\nabla_v f(a)$ – это скорость роста функции в направлении вектора v в точке a. С другой стороны

$$\nabla_v f(a) = (\operatorname{grad}_a f, v) = |\operatorname{grad}_a f||v|\cos\angle(\operatorname{grad}_a f, v)$$

Если я зафиксирую точку a и буду варьировать вектор v, то есть направление, в котором мы смотрим из точки a, то максимальное значение у выражения $\nabla_v f(a)$ будет достигнуто, когда вектор v сонаправлен градиенту. То есть градиент указывает в сторону наибольшего роста функции. Это самое важное свойство, которым обладает градиент. Он всегда говорит в какую сторону надо пойти, чтобы функция выросла. Однако, важно понимать, что градиент никогда не признается на сколько далеко надо пойти в указанном направлении. Он предпочитает оставить налет загадочности по этому вопросу.

Касательный вектор Пусть у нас наоборот задана функция $g\colon \mathbb{R} \to \mathbb{R}^n$ по правилу $t \mapsto x = g(t)$. Про нее можно думать, как про траекторию движения точки в пространстве $\mathbb{R}^n$. Моя задача определить вектор скорости для этой кривой, он же будет касательный вектор, который будет смотреть вдоль кривой, а его длина будет говорить на сколько мы быстро движемся в этот момент.



Идея такая же, как и с касательным вектором. Пусть у нас есть момент времени t_0, ему соответствует точка $x_0 = g(t_0)$. Возьмем следующий момент времени t_1, в этот момент мы будем находиться в точке $x_1 = g(t_1)$. Тогда за время $t_1 - t_0$ мы переместились из x_0 в x_1, а значит сдвинулись на вектор $x_1 - x_0$. Значит за этот момент времени наша средняя скорость равна отношению

$$v_{\text{средняя}} = \frac{x_1 - x_0}{t_1 - t_0} \in \mathbb{R}^n$$

Теперь надо время t_1 выбирать все ближе к времени t_0 и посмотреть предел средней скорости. Обратите внимание, что средняя скорость является вектором, потому, чтобы определить предел

$$v(t_0) = \lim_{t_1 \to t_0} \frac{x_1 - x_0}{t_1 - t_0}$$

нам надо понимать, что такое предел векторной последовательности. Тут можно поступить так: можно ввести норму на векторах и относительно нее определить предел, так как это делалось на прошлом занятии. Но я так же отмечал, что нормы можно выбирать разные, а понятие предела не изменится. Оказывается, можно выбрать норму так (например $\|x\|_\infty$), что сходимость будет просто покоординатной сходимостью. Последний предел так же обозначается производной по t, то есть скорость в момент времени t обозначается так $v(t) = g'(t)$.

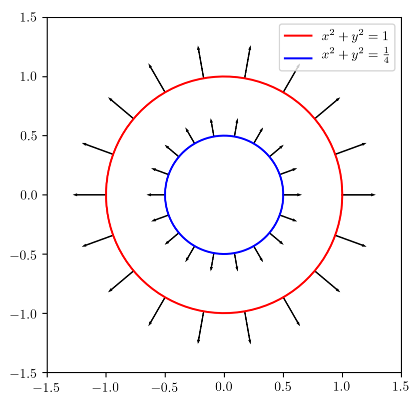
Давайте разберемся с примером. Пусть у нас есть кривая на плоскости $x(t) = (\cos t, \sin t)$. Тогда касательный вектор будет считаться, как производная по времени t. И оказывается эти операции можно сделать покоординатно, то есть $v(t) = ((\cos t)', (\sin t)') = (-\sin t, \cos t)$. То есть в точке $(\cos t, \sin t)$ касательный вектор будет $(-\sin t, \cos t)$. Аналогично все считается и в общем случае. Надо просто продифференцировать покоординатно вектор, задающий траекторию, и мы получим значение касательного вектора в данной точке. Его направление показывает куда в данный момент времени движется точка, а его длина показывает на сколько быстро мы движемся.

Поверхности и линии уровня Для функции $f\colon \mathbb{R}^n \to \mathbb{R}$ и произвольной константы $c \in \mathbb{R}$ можно нарисовать в пространстве $\mathbb{R}^n$ множество точек удовлетворяющих условию $f(x) = c$. Такую поверхность называют поверхностью уровня или изоповерхностью. На этой поверхности наша функция постоянна, а значит двигаясь по этой поверхности мы не меняем значение функции. В силу этого градиент оказывается всегда ортогонален поверхности уровня. Это можно понять так: можно запустить по изоповерхности произвольную кривую $x(t)$, тогда $f(x(t)) = c$ для любого момента времени t. Значит производная от этого выражения будет 0, так как константа, с другой стороны по правилам вычисления

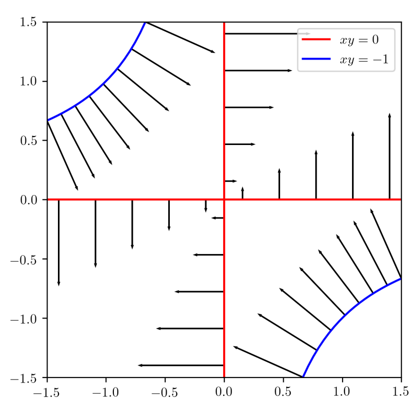
$$0 = f(x(t))' = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} x_i'(t) = (\operatorname{grad}_{x(t)} f, x'(t))$$

Здесь $x'(t)$ – это касательный вектор к $x(t)$ в момент времени t, то есть в точке $x(t)$ и градиент берется в этой самой точке $x(t)$. То есть мы видим, что любой касательный вектор ортогонален градиенту для любого пути по поверхности уровня.

- Пусть нам задана функция $f\colon \mathbb{R}^2 \to \mathbb{R}$ по правилу $(x, y) \mapsto x^2 + y^2$. Тогда ее линии уровня – это концентрические окружности с центром в начале координат. Ниже изображены две линии уровня и векторы градиента. Прошу обратить внимание, что для удобства векторы градиента нарочно укорочены в 4 раза.



- Пусть нам задана функция $f\colon \mathbb{R}^2 \to \mathbb{R}$ по правилу $(x, y) \mapsto xy$. Тогда ее линии уровня – это либо крест из осей координат, либо гиперболы. Ниже изображены две линии уровня и векторы градиента. Прошу обратить внимание, что векторы градиента нарочно укорочены, причем с разными коэффициентами на разных линиях уровня.



Касательная гиперповерхность Как и в случае функции одного переменного, мы можем провести касательную гиперповерхность к графику функции $f\colon \mathbb{R}^n \to \mathbb{R}$ в произвольной точке $a \in \mathbb{R}^n$. Мы будем считать, что у нас на $\mathbb{R}^n$ задано стандартное скалярное произведение. Теперь, если долго и пристально смотреть на случай функции одной переменной и определение градиента, то можно увидеть, что касательная плоскость будет задаваться в виде

$$y = (\operatorname{grad}_a f, x - a) + f(a) = f(a) + \frac{\partial f}{\partial x_1}(a)(x_1 - a_1) + \ldots + \frac{\partial f}{\partial x_n}(a)(x_n - a_n)$$

То есть если встать в точку $a \in \mathbb{R}^n$ и отойти на вектор $h \in \mathbb{R}^n$ в сторону от нее, то

$$f(a + h) = f(a) + (\operatorname{grad}_a f(f), h) + \rho(h), \text{ где } |\rho(h)| \leqslant C|h|^2.$$ [2, 3]

Таким образом, градиент в точке a – это еще и вектор, который задает угловые коэффициенты касательной гиперповерхности в точке a для данной функции.

Исследование на минимумы и максимумы Как и в случае функций одной переменной мы можем сказать, что такое минимум или максимум для функции $f\colon \mathbb{R}^n \to \mathbb{R}$, а так же, что такое локальный минимум и максимум. Если минимум и максимум определяются дословно, как точки, в которых значение функции самое маленькое (самое большое, соответственно), то для локальных минимумов и максимумов я должен уточнить в каком смысле понимается фраза «рядом с точкой».[4] Формально, точка $a \in \mathbb{R}^n$ называется локальным минимумом, если найдется шар радиуса r с центром в точке a, что для любой точки x из этого шара (то есть $|x - a| < r$) выполнено неравенство $f(a) \leqslant f(x)$. Аналогично дается определение локального максимума.

Скажем, что точка $a \in \mathbb{R}^n$ является критической для f, если $\operatorname{grad}_a f = 0$, то есть все частные производные f в точке a равны нулю. Как и в случае одной переменной верно следующее полезное утверждение, помогающее найти точки локального минимума и максимума.

Утверждение. *Пусть $f\colon \mathbb{R}^n \to \mathbb{R}$ – некоторая функция, $a \in \mathbb{R}^n$ – точка. Тогда, если a является точкой локального минимума или максимума, то она является критической точкой для f, то есть* $\operatorname{grad}_a f = 0$.

Гессиан Как и в случае функций одного переменного это условие является необходимым, но не достаточным. Какие-то из критических точек могут оказаться не точками минимума или максимума. Чтобы лучше понять, как ведет себя график функции в окрестности точки, используют трюк аналогичный случаю одной переменной – смотрят на приближение второго порядка.

Пусть как и раньше $f\colon \mathbb{R}^n \to \mathbb{R}$. Тогда для каждой точки $a \in \mathbb{R}^n$, мы можем определить $\partial f/\partial x_i(a)$. Таким образом получается новая функция $\partial f/\partial x_i\colon \mathbb{R}^n \to \mathbb{R}$ по правилу $x \mapsto \partial f/\partial x_i(x)$. А значит для этой новой функции тоже определены частные производные. В частности можно рассматривать частную производную по x_j от частной производной по x_i:

$$\frac{\partial \frac{\partial f}{\partial x_i}}{\partial x_j}(a)$$

Как вы видите получается безумно отвратительное выражение. Чтобы глаза не кровоточили лишний раз, подобное выражение сокращают так:

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(a) = \frac{\partial \frac{\partial f}{\partial x_i}}{\partial x_j}(a)$$

Еще полезно понимать, что для хороших функций, а на практике это означает для любых, не важно в каком порядке брать частные производные – результат будет один и тот же, то есть

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(a) = \frac{\partial^2 f}{\partial x_i \partial x_j}(a)$$

Теперь определим матрицу Гессе в точке $a \in \mathbb{R}^n$ для функции f, как матрицу составленную из частных производных:

$$H_a(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(a) & \ldots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(a) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(a) & \ldots & \frac{\partial^2 f}{\partial x_n \partial x_n}(a) \end{pmatrix}$$

[2]Здесь я так же немного обманываю с оценкой справа для упрощения жизни.

[3]Здесь $|h|$ означает длину вектора h относительно стандартного скалярного произведения.

[4]Тут опять, можно понимать понятие рядом в смысле какой-нибудь нормы. Самая популярная норма – евклидова длина, мы и будем ее использовать. Но вообще говоря, можно взять любую норму, определение от этого не изменится.

Тогда Гессианом называется следующая квадратичная форма

$$Q_{a,f}(x) = x^t H_a(f) x$$

Оказывается, что хорошие послушные функции раскладываются следующим образом

$$f(a+h) = f(a) + (\operatorname{grad}_a f, h) + \frac{1}{2}Q_{a,f}(h) + \rho(h), \text{ где } |\rho(h)| \leqslant C|h|^3$$ [5]

Обычно это равенство пишут в координатах в двух видах: в матричной форме и в координатной. Я напишу оба:

$$f(a+h) = f(a) + (\operatorname{grad}_a f)^t h + \frac{1}{2}h^t H_a(f) h + \rho(h)$$
$$f(a+h) = f(a) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a) h_i + \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(a) h_i h_j + \rho(h)$$

Таким образом, если мы хотим проверить функцию на локальный минимум или максимум, то вначале надо найти точки $a \in \mathbb{R}^n$ обнуляющие градиент, то есть с условием $\operatorname{grad}_a f = 0$. После этого, для каждой такой точки найти матрицу Гессе $H_a(f)$ и найти ее сигнатуру.[6] Если она положительно определена, то точка a будет точкой строгого минимума, а если отрицательно определена, то точкой строгого максимума. Строгость в данном случае означает, что в случае минимума, она будет строго меньше всех точек рядом, а в случае максимума, она будет строго больше всех точек рядом. Если оказалось, что матрица Гессе имеет и отрицательный и положительный индекс инерции, то a точно не точка минимума или максимума. А вот если индексы инерции неположительные или неотрицательные, то тут могут быть разные чудеса. Но как мы помним, в практике чудес не бывает.

Градиентный спуск Давайте отметим популярное приложение градиента для решения оптимизационных задач. Пусть у нас дана функция $f\colon \mathbb{R}^n \to \mathbb{R}$ и мы хотим найти ее точку локального минимума. Давайте в начале решим задачу чуть проще. Давайте найдем какую-нибудь критическую точку этой функции. Идея такая, возьмем любую случайную точку $a_1 \in \mathbb{R}^n$ и посчитаем градиент $\operatorname{grad}_{a_1} f$, если он ноль, то мы уже нашли критическую точку, если нет, то мы знаем, что если мы чуть-чуть отступим в сторону $-\operatorname{grad}_{a_1} f$, то значение функции должно уменьшиться. Выберем некоторое значение шага h_1 и шагнем в точку $a_2 = a_1 - h_1 \operatorname{grad}_{a_1} f$. Если мы были аккуратны и выбрали h_1 достаточно малым, то получим $f(a_2) < f(a_1)$. Таким образом мы можем двигаться против градиента подбирая скорость и надеяться, что в конце концов придем к некоторой критической точке.

Теперь в этом идейном алгоритме есть несколько непонятных вещей: придем ли мы когда-нибудь в критическую точку? окажется ли она точкой локального минимума? как выбирать длины шагов h_i в этом алгоритме. Ситуация с этими вопросами следующая. Если у функции есть точка локального минимума и мы попали в хорошую стартовую точку, то мы обязательно придем к точке локального минимума. Если кроме локального минимума присутствуют локальные максимумы и седловые точки, то в численных вычислениях они не создают проблем. От максимума мы и так будем уходить, двигаясь против градиента, а седло является картинкой неустойчивой, то есть если чуть-чуть сбиться с траектории мы с большой вероятностью попадем в его убывающую зону и уйдем от седловой точки. Есть одна беда, мы можем о-о-очень долго топтаться рядом с седлом. Куда хуже точки с неотрицательной сигнатурой, там могут происходить всякие неприятные эффекты. Но с практической точки зрения надо понимать, что этот метод вас всегда приведет к точке локального минимума, если шаги подобраны так, что $f(a_n) < f(a_{n-1})$.

По поводу подбора длины шага. Тут у меня для вас плохие новости. Не существует какого-то единого супер способа подобрать длину шага. Потому существует куча разных методов градиентного спуска, все они отличаются именно способом того, как подбирается длина шага на каждом этапе. Самый простой способ подбора длины шага – метод наискорейшего градиентного спуска. Когда мы выбираем шаг h_i так, чтобы выражение $f(a_i - h_i \operatorname{grad}_{a_i} f)$ было минимальным. Правда, бывает, что не всегда возможно решить явно эту задачу для данной функции f. Но и даже если мы ее решим и на каждом шаге будем находить оптимальное в этом смысле h_i, дела со скоростью стремления к минимуму у нас будут обстоять хреново. Со времен царя

[5] Как и прежде здесь $|h|$ – норма относительно стандартного скалярного произведения и я вас чуть-чуть обманываю в оценке справа.

[6] Сигнатуру можно проверять методом Якоби, симметричным Гауссом, а можно через поиск собственных значений матрицы Гессе. Данная задача как раз является той самой причиной, по которой нам важен метод определения сигнатуры.

Гороха люди накопили огромное количество всяких дурацких примеров, когда наискорейший градиентный спуск ведет себя плохо. Важный вывод, который надо сделать – градиентный спуск это просто идейно и сложно технически, когда речь заходит о подборе длины шага. Данным вопросом занимается целая область математики.