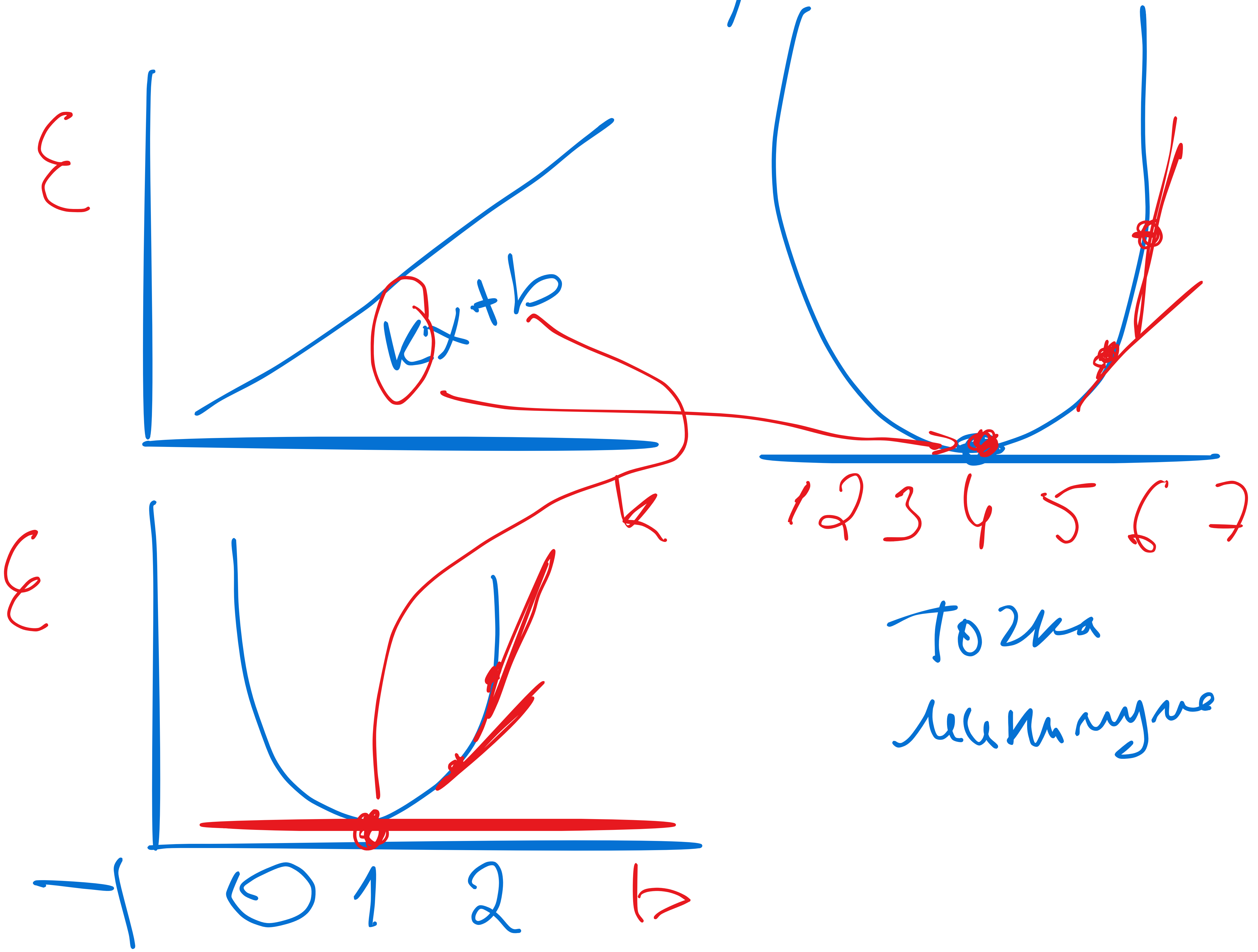
Что мы хотим сделать?

Мы хотим найти минимум функции (MSE) $(Y - \hat{Y})^2$

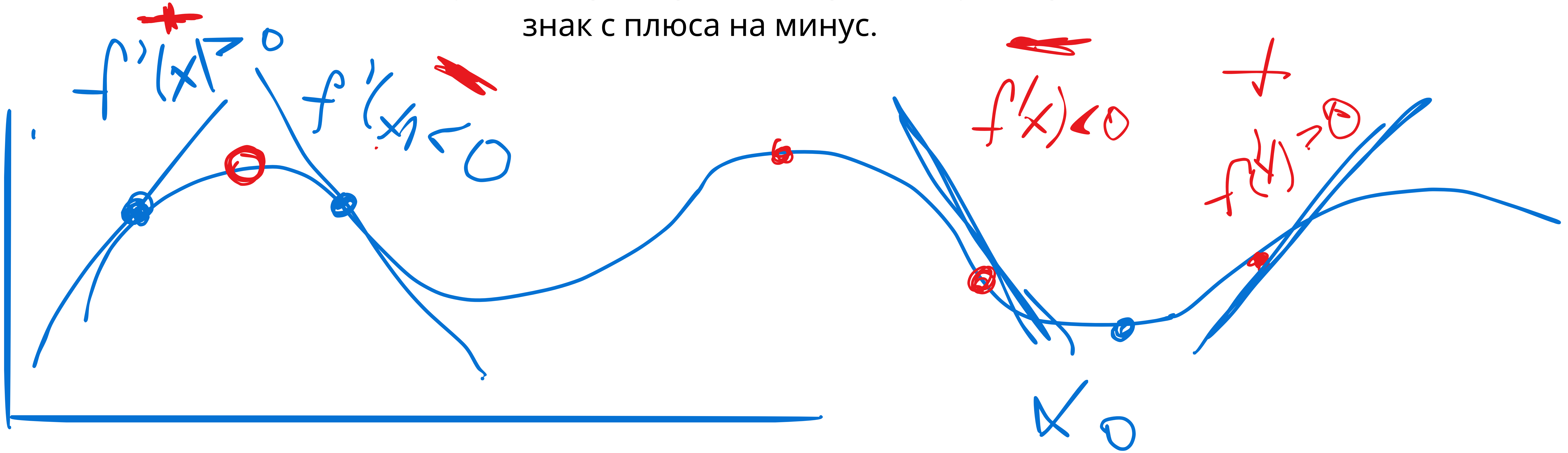
Градиентный спуск это алгоритм оптимизации



Теорема Ферма

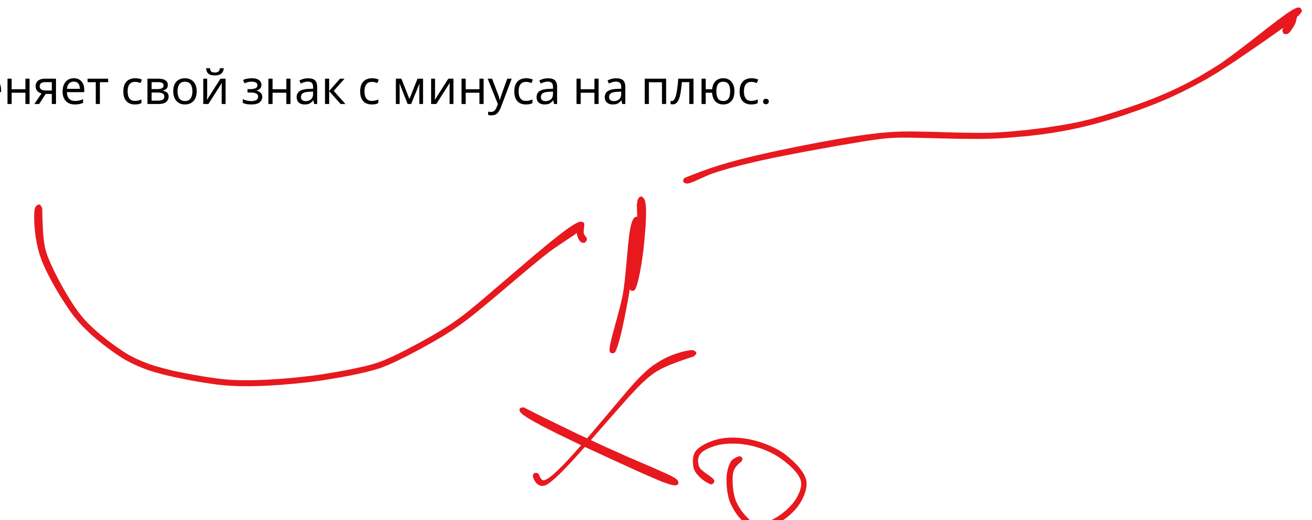
Теорема (Ферма). Пусть функция f(x) определена в некотором промежутке; имеет локальный экстремум во внутренней точке x0 этого промежутка.

Если x0 − точка локального максимума, то при переходе через точку x0 производная f′(x) меняет свой знак с плюса на минус.

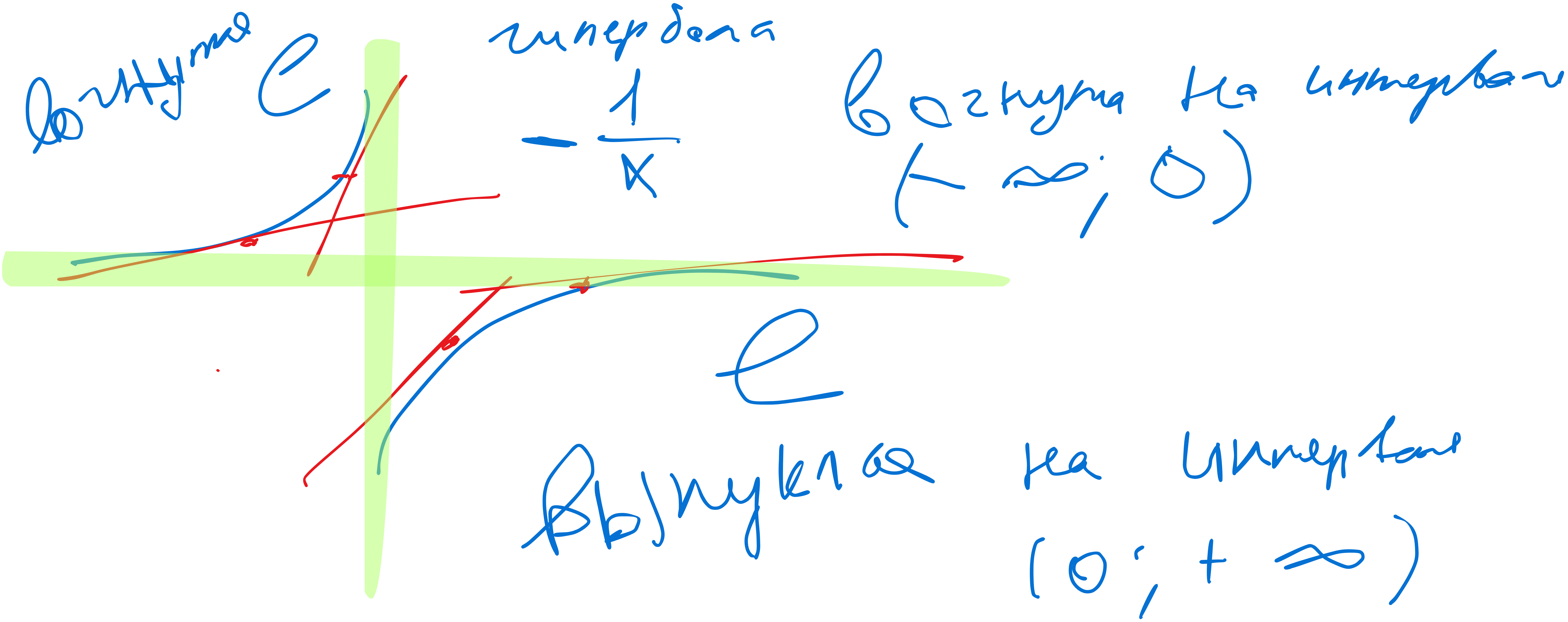


Если x0x0 − точка локального минимума, то при переходе через точку x0 производная f′(x) меняет свой знак с минуса на плюс.

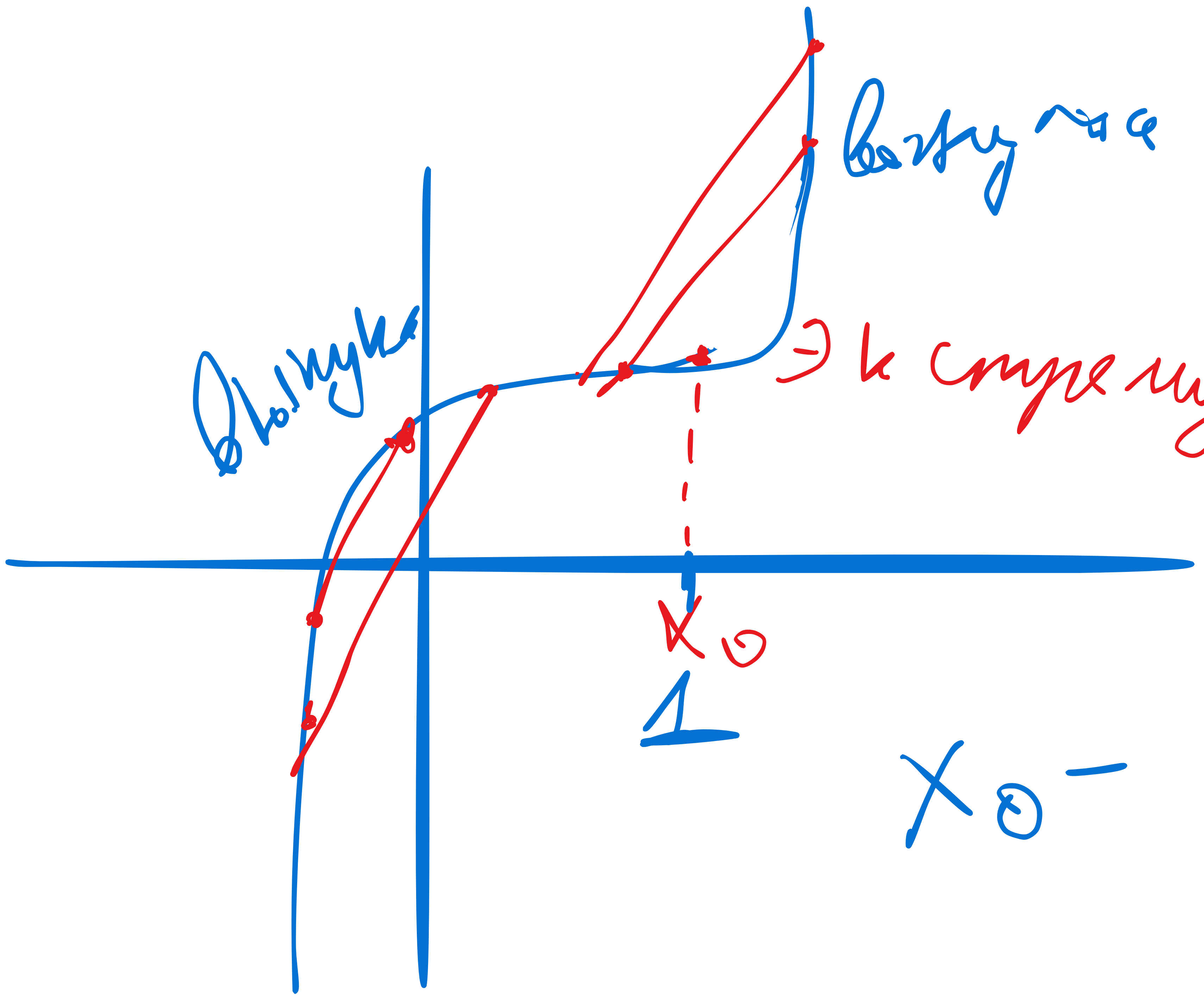
Если функция f(x) дифференцируема в точке x0, то f′(x0)=0



Точка графика, в которой он меняет выпуклость на вогнутость или вогнутость на выпуклость, называется точкой перегиба.



Выпуклый на интервале график расположен не выше касательной, проведённой к нему в произвольной точке данного интервала. Вогнутый же на интервале график − не ниже любой касательной на этом интервале.



Экстремума – критическая т

$\Rightarrow$ точка перегиба

x_0 – точка перегиба

График функции является выпуклым на некотором интервале, если он расположен не ниже любой хорды данного интервала

$(-\infty; 1)$

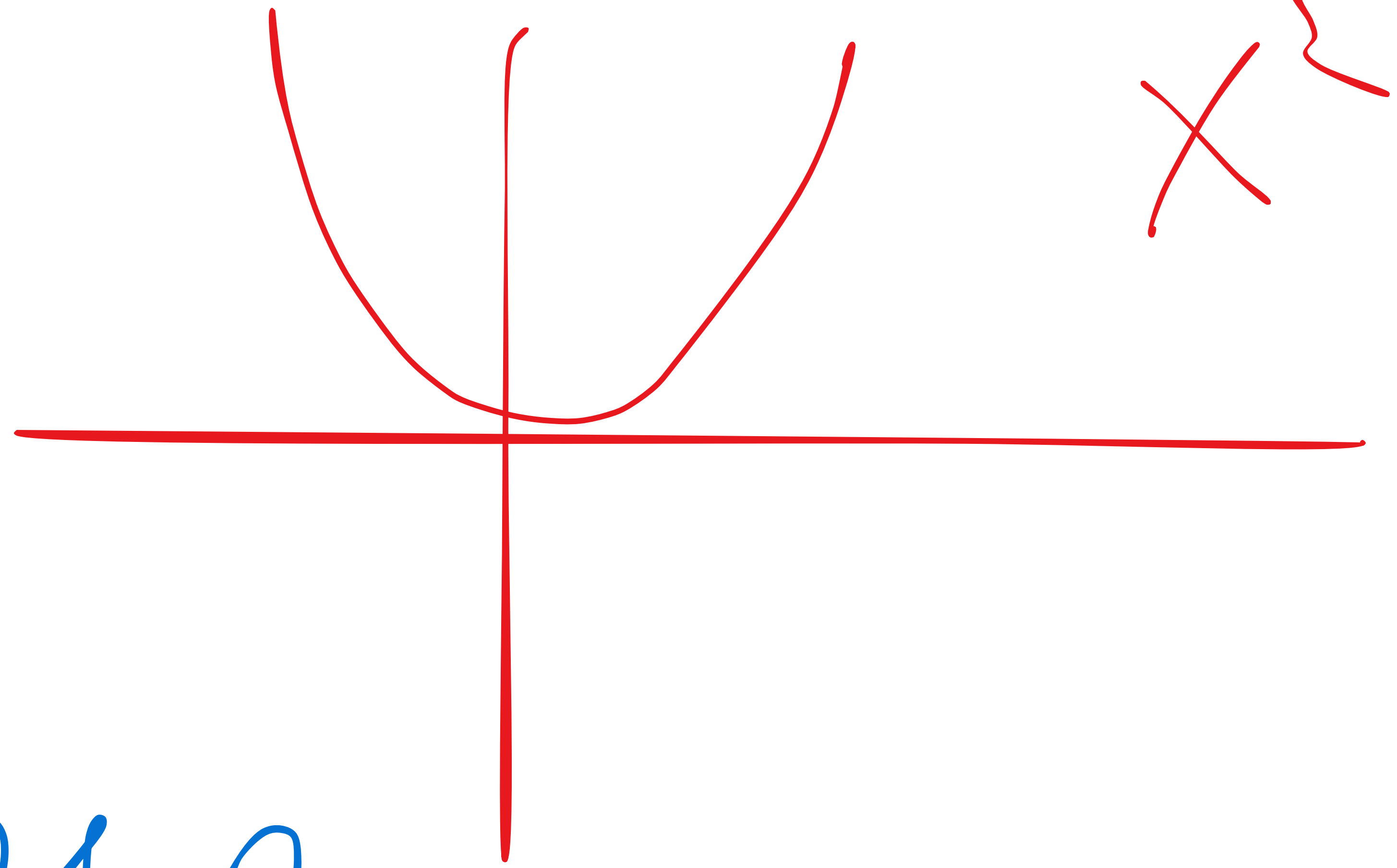
График функции являются вогнутым на интервале, если он расположен не выше любой хорды этого интервала.

$(1; +\infty)$

Необходимое условие перегиба
Если в точке есть перегиб графика функции , то:
либо значения не существует

$$f''(x_0) = 0 \qquad f''(x_0) \text{ — не сущ.}$$

$$n \cdot x^{n-1}$$



x^2

$$f' = x^2 = 2x^1$$

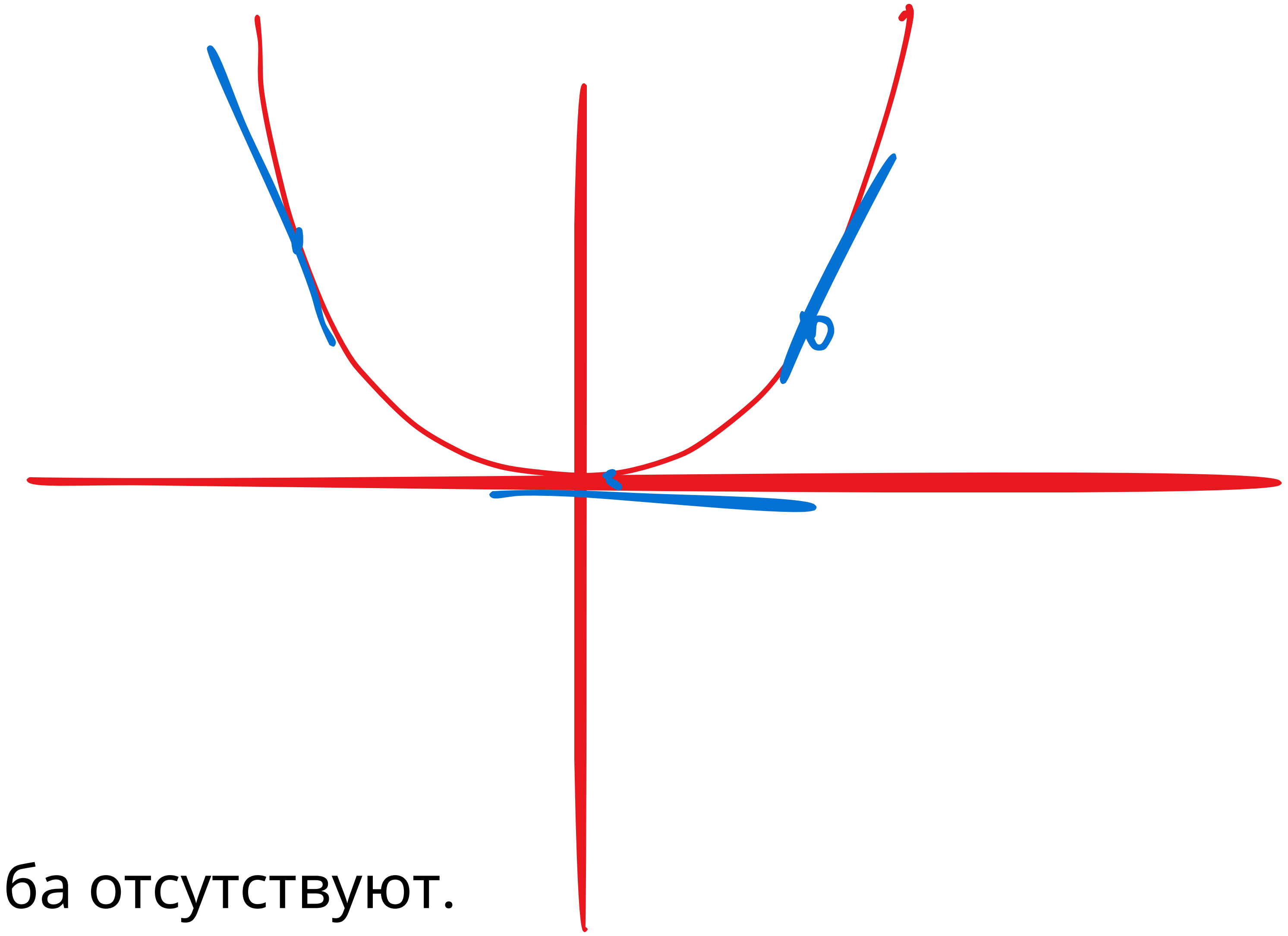
$$f''(2x) = 2(x)' = 2 \cdot 1 = 2$$

Достаточное условие перегиба

Если вторая производная при переходе через точку меняет знак, то в данной точке существует перегиб графика функции .

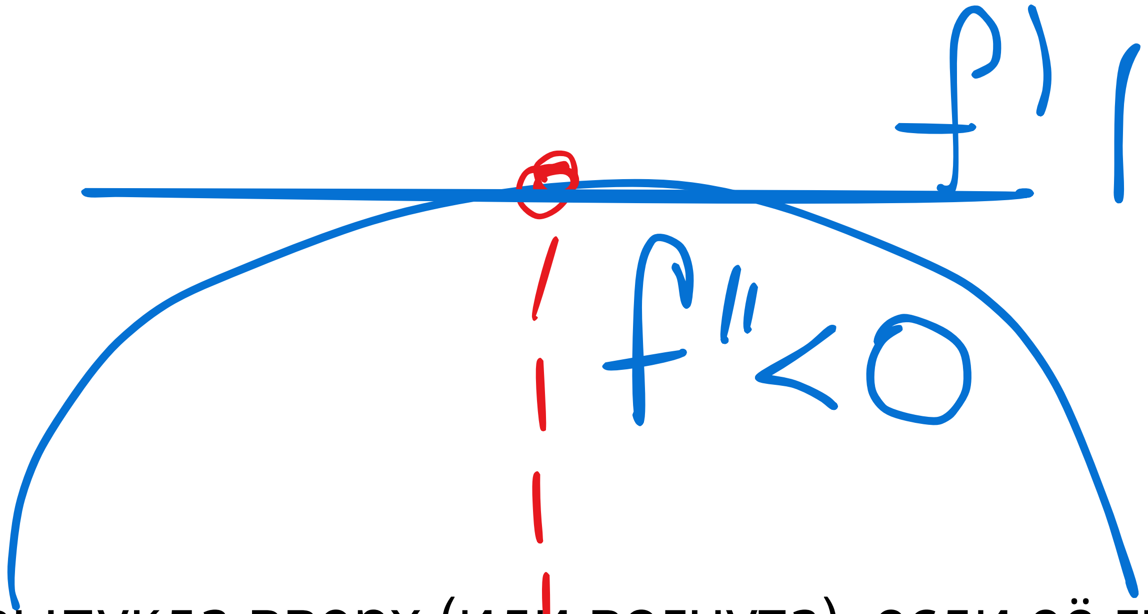
$$f' = x^2 = 2x > 0$$

$$f'' = 2 > 0$$



вогнута на всей области определения, точки перегиба отсутствуют.

Функция называется выпуклой вниз (или просто выпуклой), если её график «лежит ниже» любой прямой, соединяющей две произвольные точки на графике

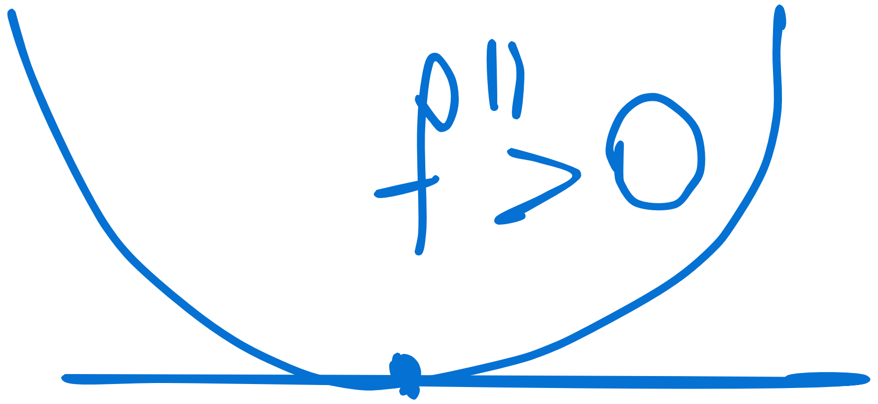


$f'(x_1) = 0$ $\qquad -x^2$

$f'' < 0$

функция выпукла вверх (или вогнута), если её график «лежит выше» такой прямой.

выпукла вниз, выпукла



$f'' > 0$

$f'(x_2) = 0$

$x^2 \Rightarrow f' = 2x = 0$

$f'' = 2 > 0$

Этапы оптимизации функций

1) Нахождение критических точек

вычисляем $f'(x)$

$\perp$ $f'(x) = 0 \vee f'(x)$ not $\exists$

крит точки

2) анализ характера крит. точек

$f''(x_0) > 0$ в (.) x_0 — лок мин

$f''(x_0) < 0$ в (.) x_0 — ~~лок~~ макс

$f''(x_0) = 0$ в (.) доп анализ исп выпуклость

Опр. интервалы $\nearrow$ $\searrow$

$\exists$ $f'(x_0) > 0$ на инт $\nearrow$

$\exists$ $f'(x_0) < 0$ на инт $\searrow$

анализ выпуклости

$f''(x) > 0$ выпукла вниз

$f''(x) < 0$ выпукла вверх

$$f'(x) = (x^2)' - 4(x)' + (3)'$$

$$f'(x) = 2x - (4)'$$

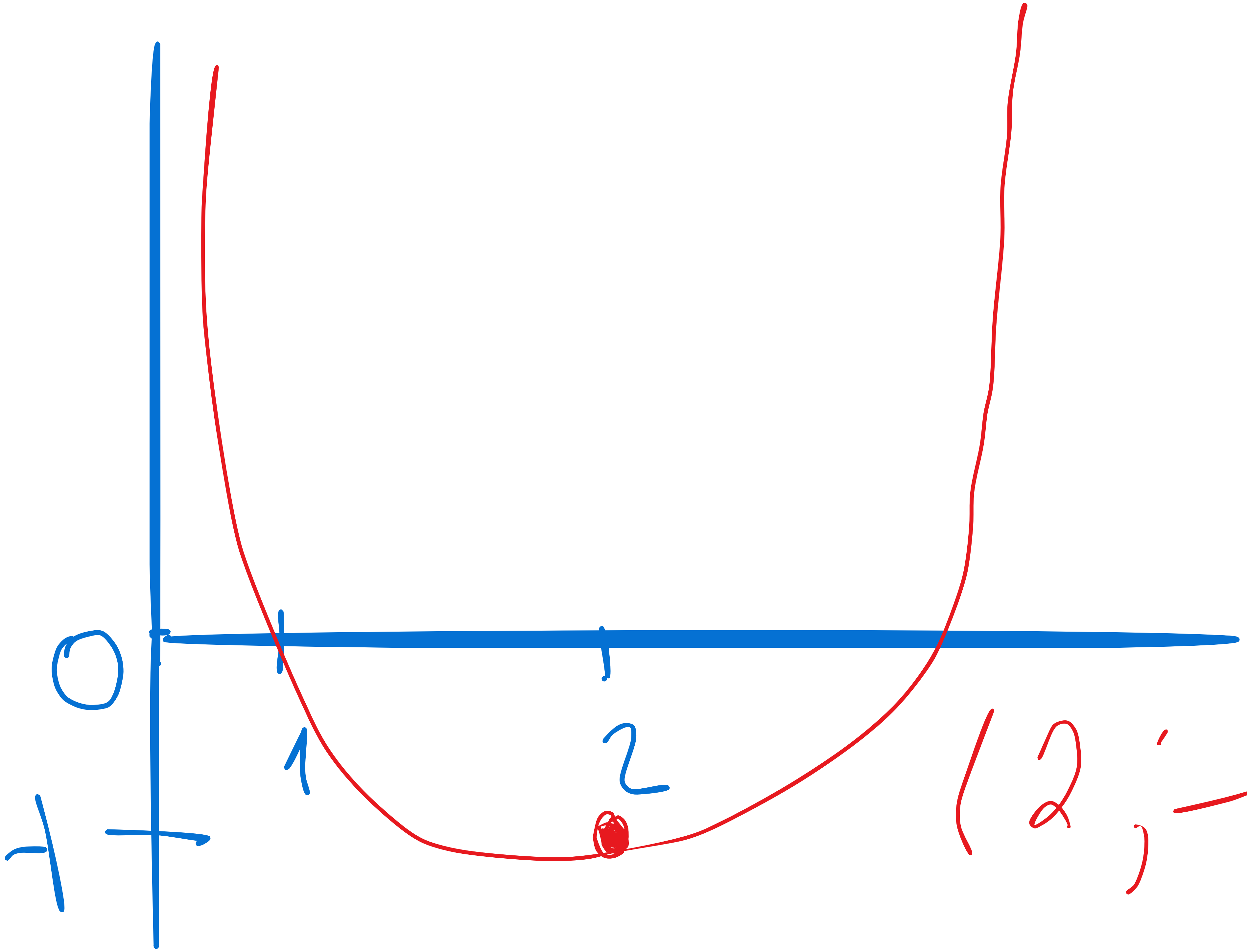
$$f''(x) = 2 > 0$$

$(\cdot)$ $x = 2$ локальный минимум

$$f(2) = 2^2 - 4\cdot 2 + 3 = -1$$

$f'(x) > 0$ $x > 2$ ↗

$f'(x) < 0$ $x < 2$ ↘

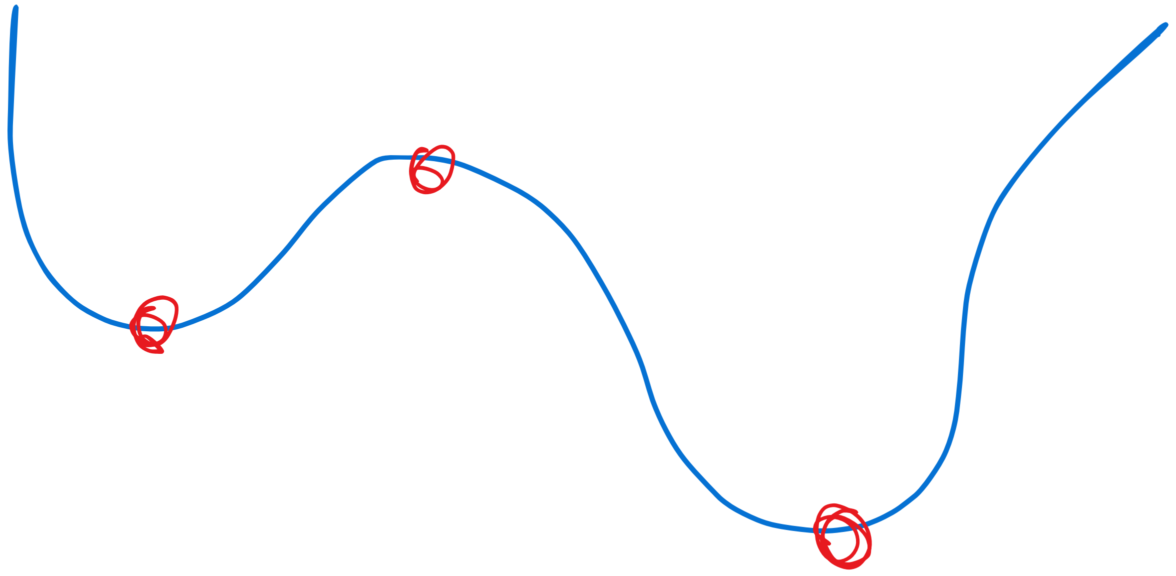


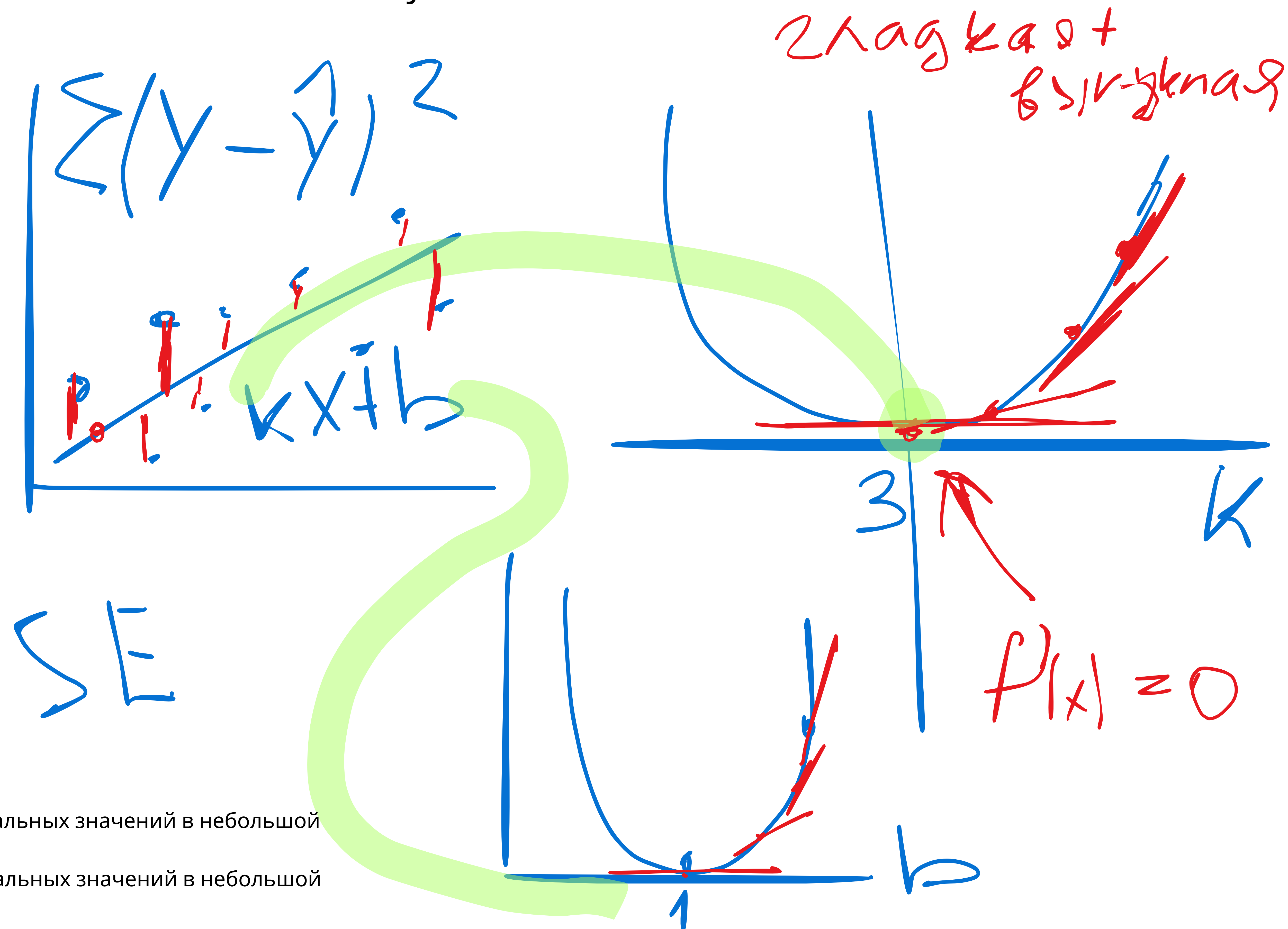
$(2; -1)$

$(-\infty; 2) \searrow$

$[2; +\infty) \nearrow$

основы оптимизации функций — процесса поиска экстремальных значений (минимумов и максимумов)





MSE

Локальный минимум — значение функции меньше всех остальных значений в небольшой окрестности точки.
Локальный максимум — значение функции больше всех остальных значений в небольшой окрестности точки.

Теорема Ролля   теорема о среднем значении

если функция принимает одинаковые значения f(a)=f(b) и она непрерывна и дифференцируема внутри отрезка, существует точка между a и b где функция должна иметь горизонтальную касательную, в которой производная равна нулю

Пусть функция f(x))

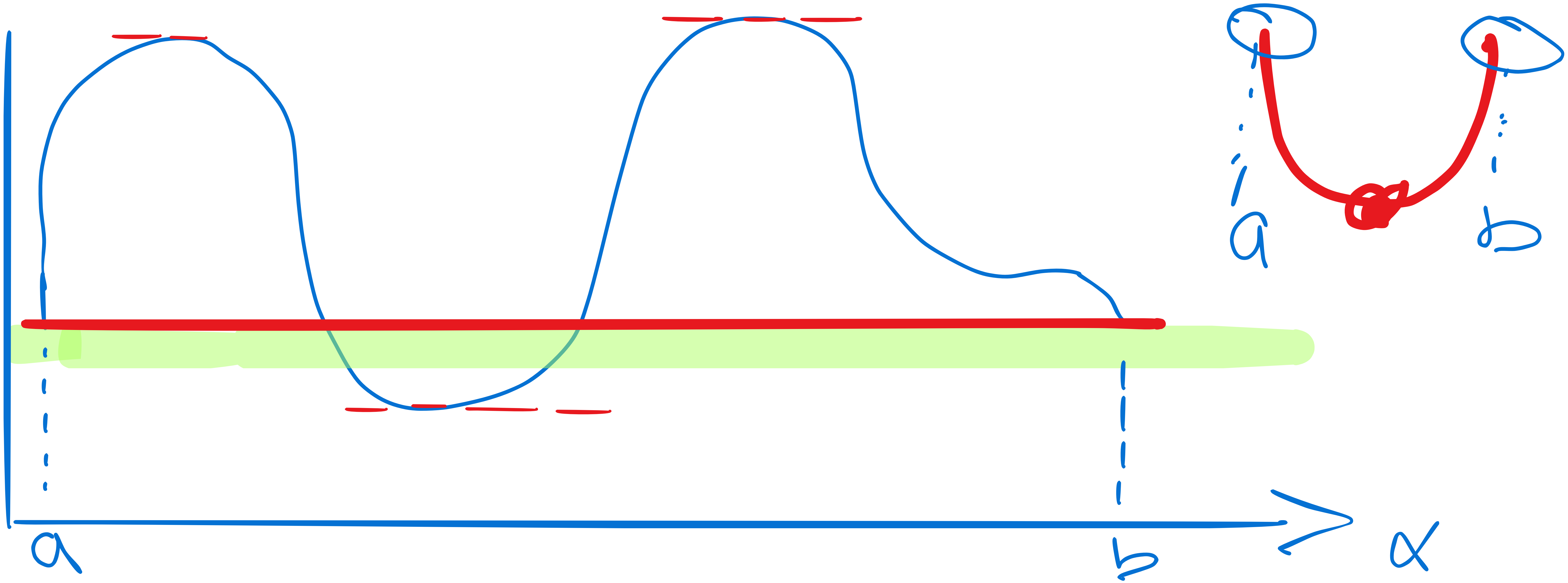
Непрерывна на [a,b]

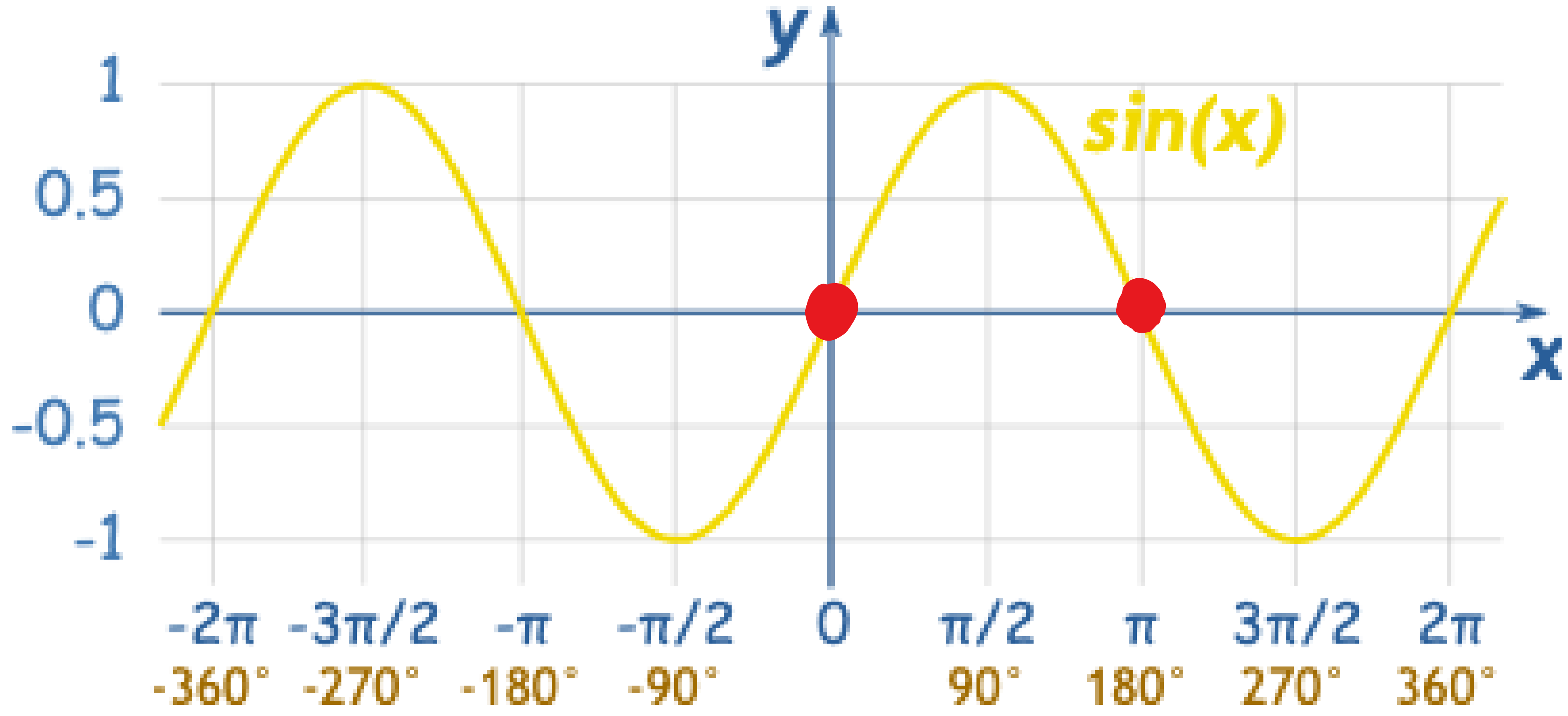
Дифференцируемость(a,b)

f(a)=f(b)

Тогда существует такая точка c∈(a,b) что

Гарант-Лок. Экст $f'(c)=0$





$f(x) = \sin(x)$

$[0; \pi]$

1) непрерывна $[0; \pi]$

2) дифференц. $(0; \pi)$

3) равенство значений на концах отрезка

$a \qquad b$

$f(a) = f(b)$

$f(0) = \sin(0) = 0$

$f(\pi) = \sin(\pi) = 0$

$f'(x) = \cos(x) = 0 \Rightarrow x = \frac{\pi}{2} \quad c = \frac{\pi}{2} \quad f'(c) =$

$$\frac{d}{dx}(\sin(x))$$

$$= \boxed{\cos(x)}$$

$$\sin' x = \lim_{h \to 0} \frac{\sin x \cos h + \cos x \sin h - \sin x}{h}$$

$$= \lim_{h \to 0} \frac{\sin x \cos h - \sin x + \cos x \sin h}{h}$$

$$= \lim_{h \to 0} \frac{\sin x(\cos h - 1) + \cos x \sin h}{h}$$

$$= \lim_{h \to 0} \sin x \cdot \frac{(\cos h - 1)}{h} + \lim_{h \to 0} \cos x \cdot \frac{\sin h}{h}$$

$$= \sin x \cdot 0 + \cos x \cdot 1 = \underline{\cos x}$$

Точка экстремума - это стационарная точка

Стационарные точки $f'(x) = 0$

Стационарная точка - это частный случай критической точки

Теорема Лагранжа (или теорема о среднем значении)

Если функция гладкая (без разрывов и "угловатостей"), то существует хотя бы одна точка на интервале, в которой касательная к графику функции будет иметь тот же наклон, что и секущая прямая, соединяющая точки (a,f(a)) (b,f(b))

1) непрерыв. [a, b]
2) диф. (a,b)

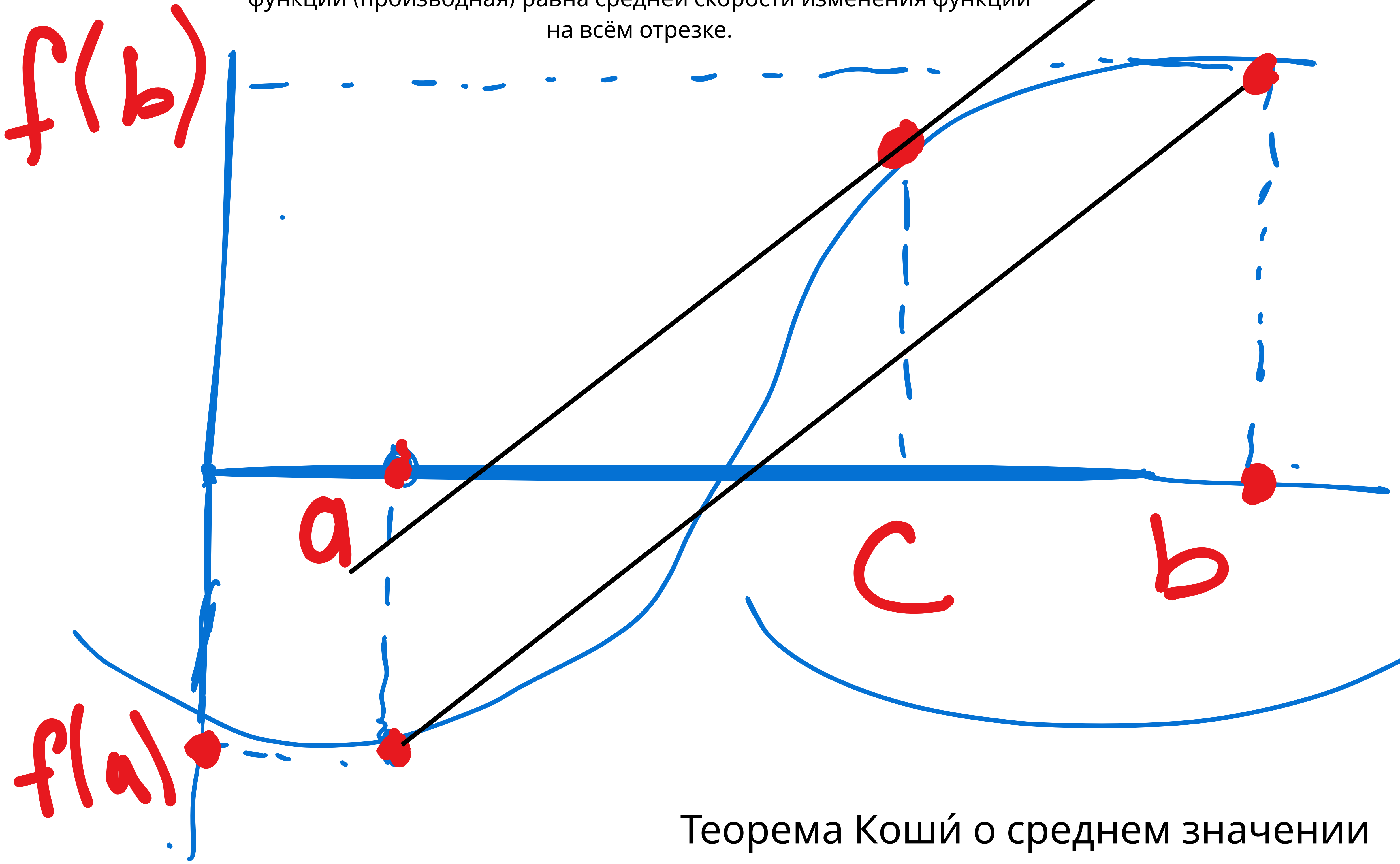
Теорема утверждает: найдётся точка c, где касательная к графику параллельна этой секущей.

Секущая прямая — это прямая, соединяющая концы графика на отрезке [a,b]

$c \in (a, b)$

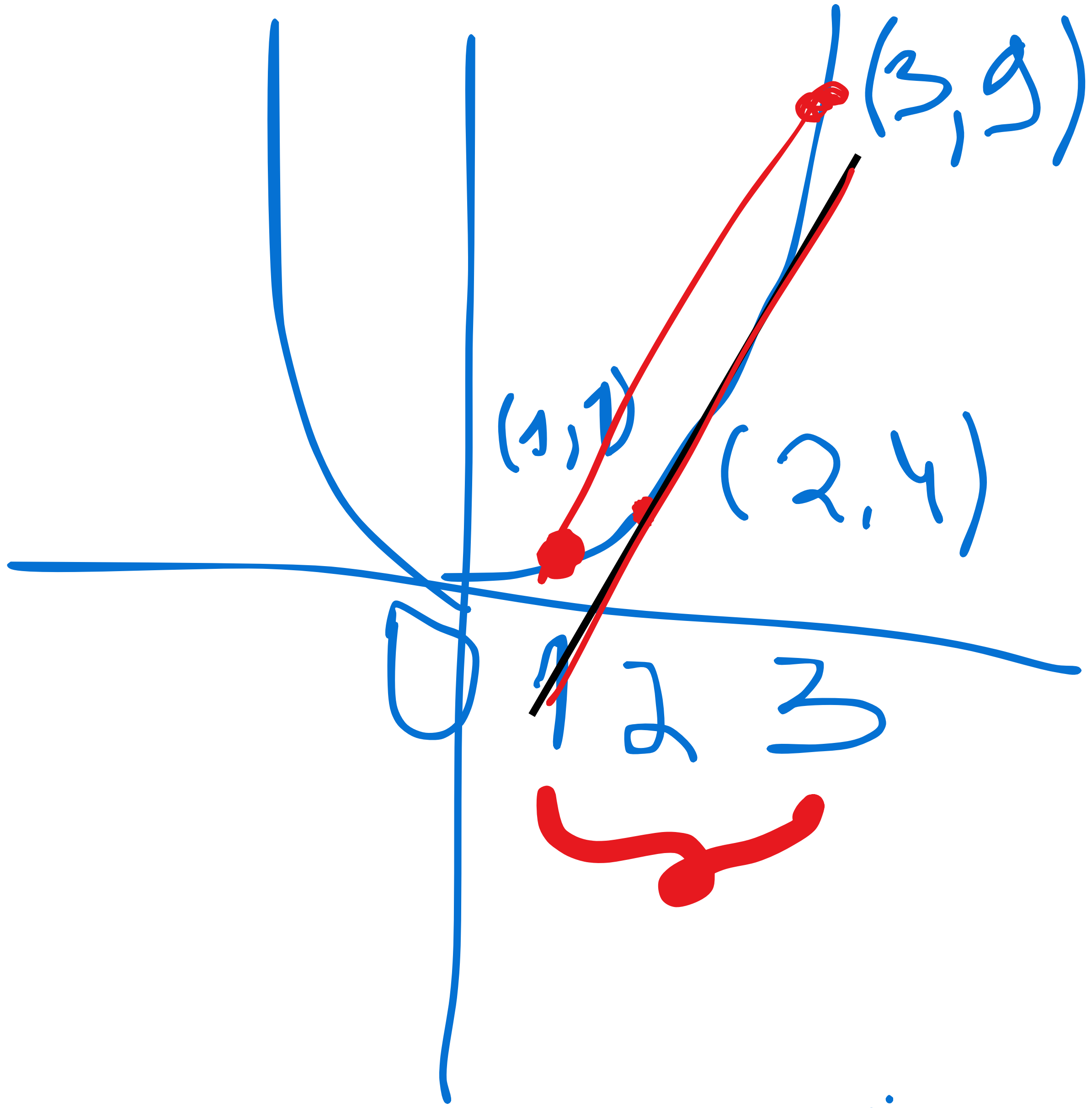
существует точка, в которой мгновенная скорость изменения функции (производная) равна средней скорости изменения функции на всём отрезке.

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$



$(a, f(a))$
$(b, f(b))$

Теорема Коши́ о среднем значении



$f(x) = x^2 \quad [1, 3]$

1) непрерывность

2) дифференц. $= f'(x) = 2x$

сущ для всех x, значит

функция дифференцируема

на $(1, 3)$

ср. скорость изм на отрезке $[1; 3]$

$\frac{f(3) - f(1)}{3-1} = \frac{9-1}{2} = \frac{8}{2} = 4$

По теореме Лагранжа

$\exists$ (·) $c \in (1,3) \mid f'(c) = 4$

$f'(x) = 2x \Rightarrow 2c = 4 \Rightarrow c = 2$

В (·) $c = 2$

мы обучаем модель линейной регрессии и используем функцию потерь L(θ)

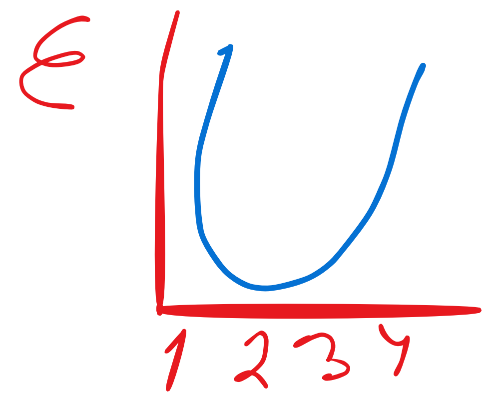
loss(θ)

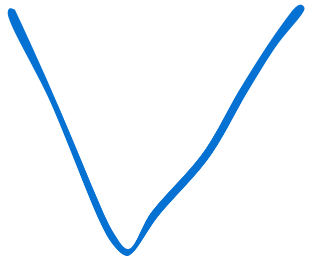
$\theta_0 + \theta_1 x_1 + \dots + \theta_n x_n$ → MSE, MAE

$\exists\, L(\theta)$

1) Непрерывна $[\theta_1, \theta_2]$

2) Дифференц-

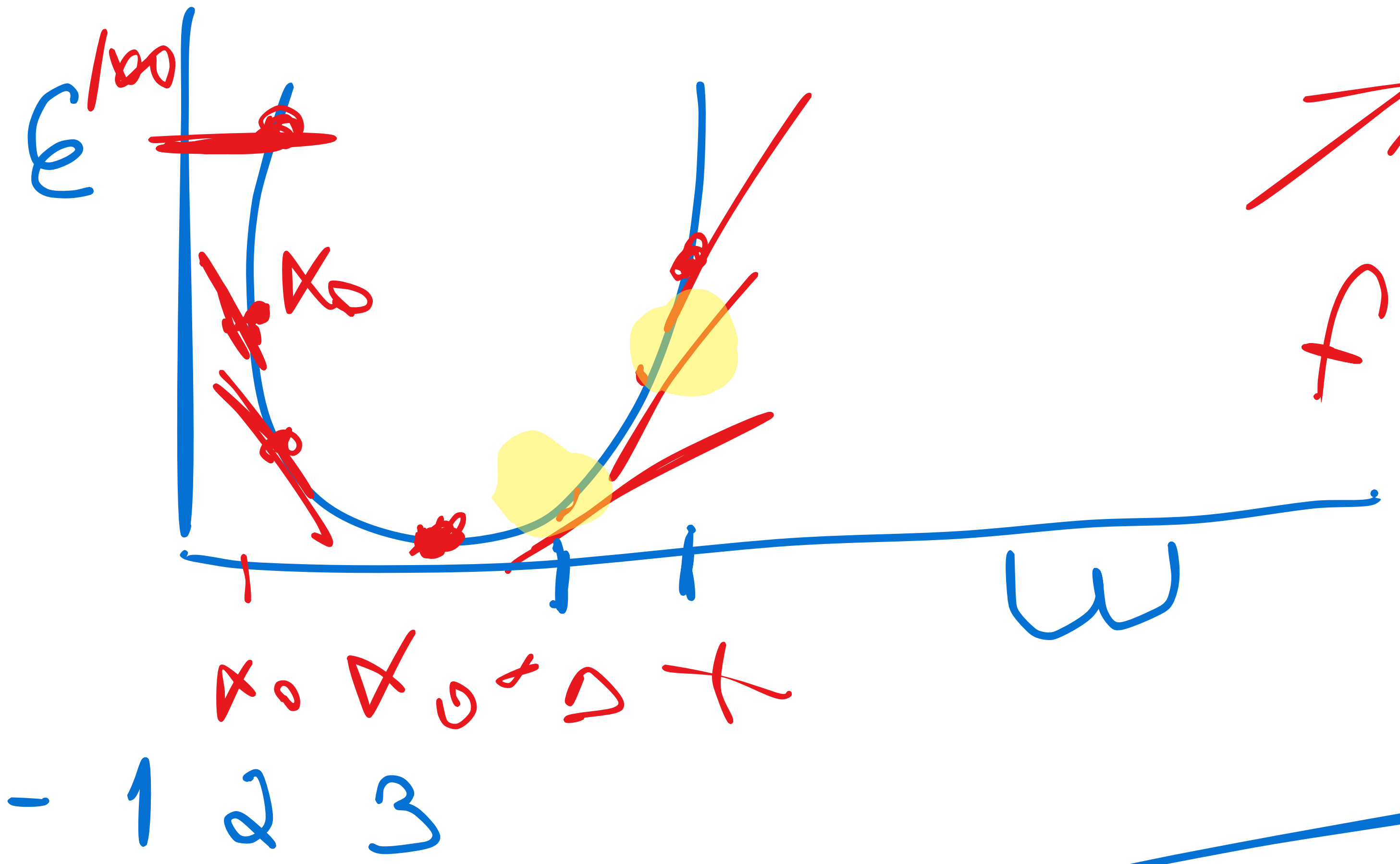




что $\theta = \theta^*$

3   2   θ x^2

$$L'(\theta) = \frac{L(\theta_2) - L(\theta_1)}{\theta_2 - \theta_1}$$



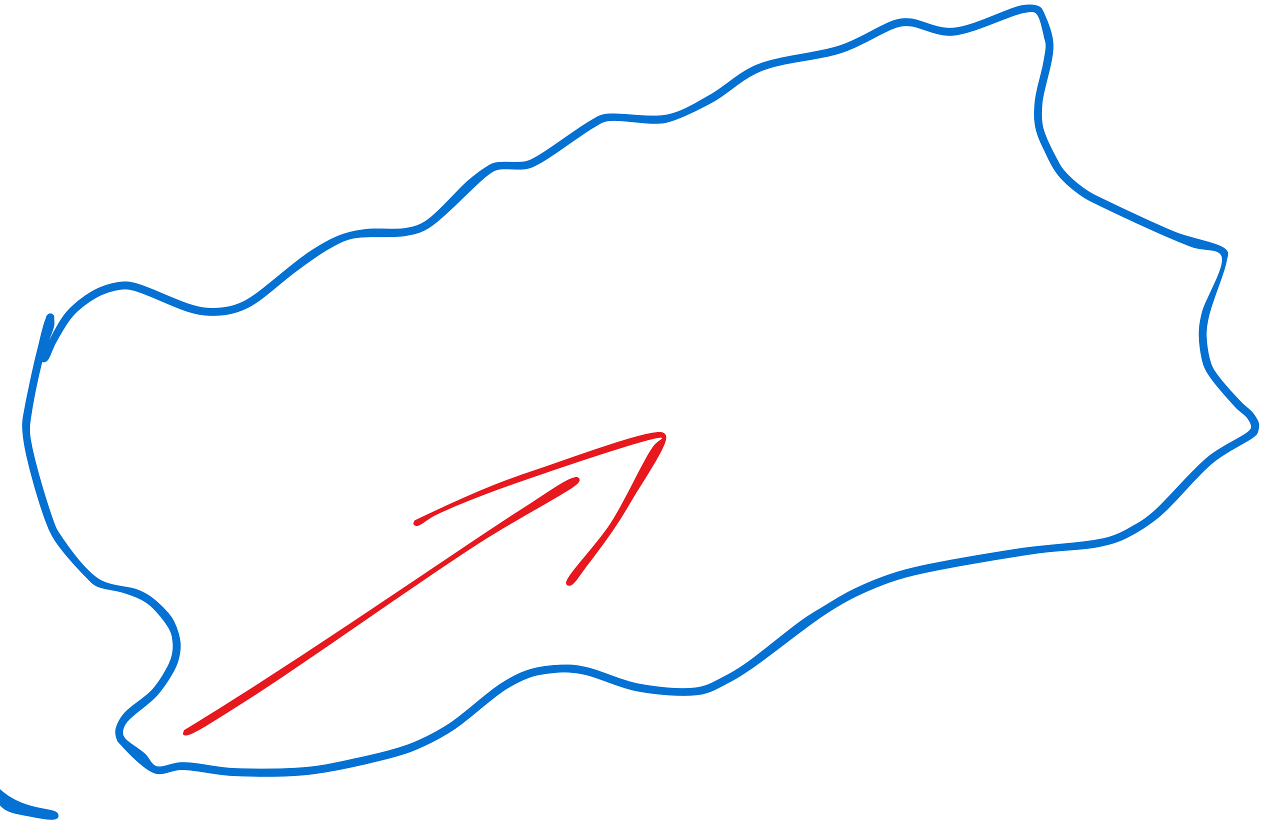
$\nearrow +$

производная = градиент

$f'(x) \cdot (-1)$ для ФНП

вектор $\nearrow \cdot (-1)$

$$\frac{J(w_2) - J(w_1)}{w_2 - w_1}$$



Если производная мала → изменения функции потерь малы → меняем шаг

Если производная велика → функция сильно падает → меньше шаг

$$w_1 = 1$$

$$w_2 = 2$$

$$J(w) = (w-3)^2$$

$$J(1) = (1-3)^2 = 4$$

$$J(2) = (2-3)^2 = 1$$

$$\frac{J(2) - J(1)}{2-1} = \frac{1-4}{1} = -3$$

$$J'(w) = 2(w-3) \Rightarrow$$

$$J'(c) = -3$$

$$2c = -3/2 = -1.5$$

Необходимые и достаточные условия экстремума функции одной переменной

Далее мы не будем различать локальные и глобальные экстремумы. Признаки будут срабатывать в обоих случаях.

Необходимое условие: если в точке x_0 экстремум, то либо $f'(x_0) = 0$, либо значения $f'(x_0)$ не существует. В первом случае мы фактически говорим о том, что касательная в данной точке должна быть параллельна оси OX. В втором случае несуществование производной в точке также может свидетельствовать об экстремуме: $f(x) = |x|, x_0 = 0$ - данная функция не дифференцируема в 0, однако имеет в этой точке глобальный минимум.

Это условие является необходимым, но не достаточным, поскольку, например, производная равна нулю в точках изменения кривизны, которые не являются экстремальными:

$f(x) = x^3, x_0 = 0$ - в точке 0 нет экстремума, но производная в ней равна 0.

Достаточное условие: при переходе через точку экстремума производная должна менять знак.

Два эти условия и описывает **теорема Ферма**, разобранная на предыдущем шаге.

Наибольшее и наименьшее значение функции

Предположим, что мы хотим найти наибольшее и наименьшее значение функции $f(x)$ на отрезке $[a, b]$. Заметим, что граничные точки в этом случае не могут являться точками экстремума, поскольку проверка перемены знака производной при переходе через точку требует существования функции в некоторой левой и правой полуокрестности этой точки. Но для граничных точек одна из полуокрестностей отсутствует - ведь мы не рассматриваем функцию вне отрезка . Поэтому глобальные экстремумы на отрезке не обязательно соответствуют максимальному/минимальному значению функции: **нужно отдельно проверить значение функции на концах отрезка и сравнить со значениями в экстремальных точках**.

🔍 Представим: ты настраиваешь линейную регрессию

У тебя есть **функция потерь MSE**, которая зависит от параметра (например, w — это наклон прямой).
Функция потерь показывает, **насколько плохо** модель предсказывает данные.
Ты хочешь **минимизировать** эту функцию — то есть **найти w, при котором ошибка минимальна**.

🎯 Что делает градиентный спуск?

Он говорит:

> "Чтобы уменьшить ошибку, пойду в сторону минуса производной"
> (то есть туда, где функция убывает)

Но... чтобы это делать **умно и быстро**, нужно понять:

- Насколько **сильно** меняется ошибка, если немного изменить w?
- Как **точно выбрать шаг изменения w?**

🔗 Здесь и вступает теорема Лагранжа:

Она говорит:

> "Между двумя точками параметра w_1 и w_2 **обязательно найдётся точка**, где производная (градиент) точно равна **среднему изменению функции потерь**"

Формула:

$$\frac{J(w_2) - J(w_1)}{w_2 - w_1} = J'(c) \quad \text{для некоторого } c \in (w_1, w_2)$$

🧠 Что это даёт?

1. **Ты можешь гарантировать**, что производная не исчезает "внезапно" — ты точно знаешь, **где-то между двумя значениями она существует**, и ты на неё опираешься при обновлении весов.
2. **Это помогает анализировать сходимость**:
   - Если производная мала → изменения функции потерь малы → нужно уменьшить шаг.
   - Если производная велика → функция сильно падает → шаг может быть больше.

🧪 Простой числовой пример:

Пусть у тебя веса:

- $w_1 = 1$
- $w_2 = 2$

И функция потерь:

$$J(w) = (w - 3)^2$$

Тогда:

- $J(1) = (1 - 3)^2 = 4$
- $J(2) = (2 - 3)^2 = 1$
- Среднее изменение:

$$\frac{J(2) - J(1)}{2 - 1} = \frac{1 - 4}{1} = -3$$

А производная:

$$J'(w) = 2(w - 3) \Rightarrow J'(c) = -3 \Rightarrow c = 1.5$$

Вот и точка c, в которой градиент точно равен этому среднему изменению.

✅ Вывод:

Теорема Лагранжа в линейной регрессии:

- Даёт **математическую гарантию**, что между двумя значениями весов градиент есть и соответствует изменению функции потерь.
- Помогает **понимать, насколько эффективен шаг градиентного спуска**
- Используется при **доказательствах сходимости** и в **оценке скорости обучения**

Хочешь, покажу это всё на графике с функцией потерь и касательной?

теорем Ролля,гарантирует наличие экстремума (минимума или максимума) при определённых условиях. В задачах минимизации оптимизации функции потерь именно эта теорема помогает установить условия существования экстремумов.

Как связать с теоремой Ролля?
Теорема Ролля важна теоретически — она объясняет, почему и где у функции (в том числе функции потерь!) может быть минимум.
Применение:
Функция потерь непрерывна.
Часто дифференцируема.
Мы ищем точки, где градиент равен нулю: ∇MSE=0\nabla \text{MSE} = 0∇MSE=0, как в теореме Ролля.
То есть: если у функции есть одинаковые значения в двух точках и она гладкая — внутри точно есть точка, где производная = 0.
А такие точки — кандидаты на минимум или максимум.
В градиентном спуске мы как раз ищем эти точки — но не зная заранее, где производная равна нулю, мы к ним шагаем по направлению градиента.
🧠 Почему важно понимать это?
Потому что:
Теорема Ролля и её обобщение — теорема Ферма и Лагранжа — обосновывают, что минимум достигается в точке, где производная = 0.
Это объясняет, зачем мы приравниваем градиент к нулю и ищем такие точки в оптимизации (в том числе при обучении моделей машинного обучения).

Стационарная точка
Это частный случай критической точки:
$f'(x0)=0$

То есть: если у функции есть одинаковые значения в двух точках и она гладкая — внутри точно есть точка, где производная = 0.

В линейной регрессии
Функция потерь — выпуклая (параболоид). Поэтому:
Есть единственная стационарная точка (где градиент равен нулю).
Она — глобальный минимум.
Именно к ней сходится градиентный спуск.

3. Формула градиентного спуска:

$$x_{\text{новое}} = x_{\text{старое}} - \eta \cdot f'(x_{\text{старое}})$$

Где:

- η — **шаг обучения** (learning rate), положительное число (например, 0.1)
- $f'(x)$ — производная в текущей точке
- Мы **отнимаем** градиент, потому что идём **вниз**

---

4. Пример

Пусть:

- $f(x) = x^2$
- Начнем с $x_0 = 3$
- Шаг $\eta = 0.1$

Тогда:

- $f'(3) = 2 \cdot 3 = 6$
- $x_1 = 3 - 0.1 \cdot 6 = 2.4$
- $f'(2.4) = 2 \cdot 2.4 = 4.8$
- $x_2 = 2.4 - 0.1 \cdot 4.8 = 1.92$
- и так далее...

Через несколько шагов ты приближаешься к **минимуму в точке** $x = 0$