START ML

KARPOV.COURSES

ЛИНЕЙНАЯ РЕГРЕССИЯ OLS: МАТРИЧНАЯ ФОРМА

Тогда задача поиска оптимальных коэффициентов сводится к минимизации произведения двух матриц. Можно взять матричный дифференциал и приравнять его к нулю. Этот шаг аналогичен нахождению частных производных и решению системы для поиска критических точек.

$$Q = \frac{1}{n} \cdot (X \cdot B - Y)^T (X \cdot B - Y) \to min$$

$$dQ(\beta^*) = 0$$

Взятие матричного дифференциала рассматривать не будем, запишем сразу результат:

$$\beta^* \quad = \quad (X^T \cdot X)^{-1} \cdot X^T \cdot Y$$

КАК НА САМОМ ДЕЛЕ РАБОТАЕТ МО?

ПОД ОСТАЛЬНЫЕ ЗАДАЧИ ТОЖЕ ЕСТЬ ФОРМУЛЫ?

На самом деле, не во всех задачах существует универсальная формула, описывающее единственное и лучшее решение. Нужен более гибкий подход к минимизации ошибки.

Так же хотелось бы уметь легко распределять вычисления.

ГРАДИЕНТНЫЙ СПУСК

Функция одной переменной $f(x) = x^2$

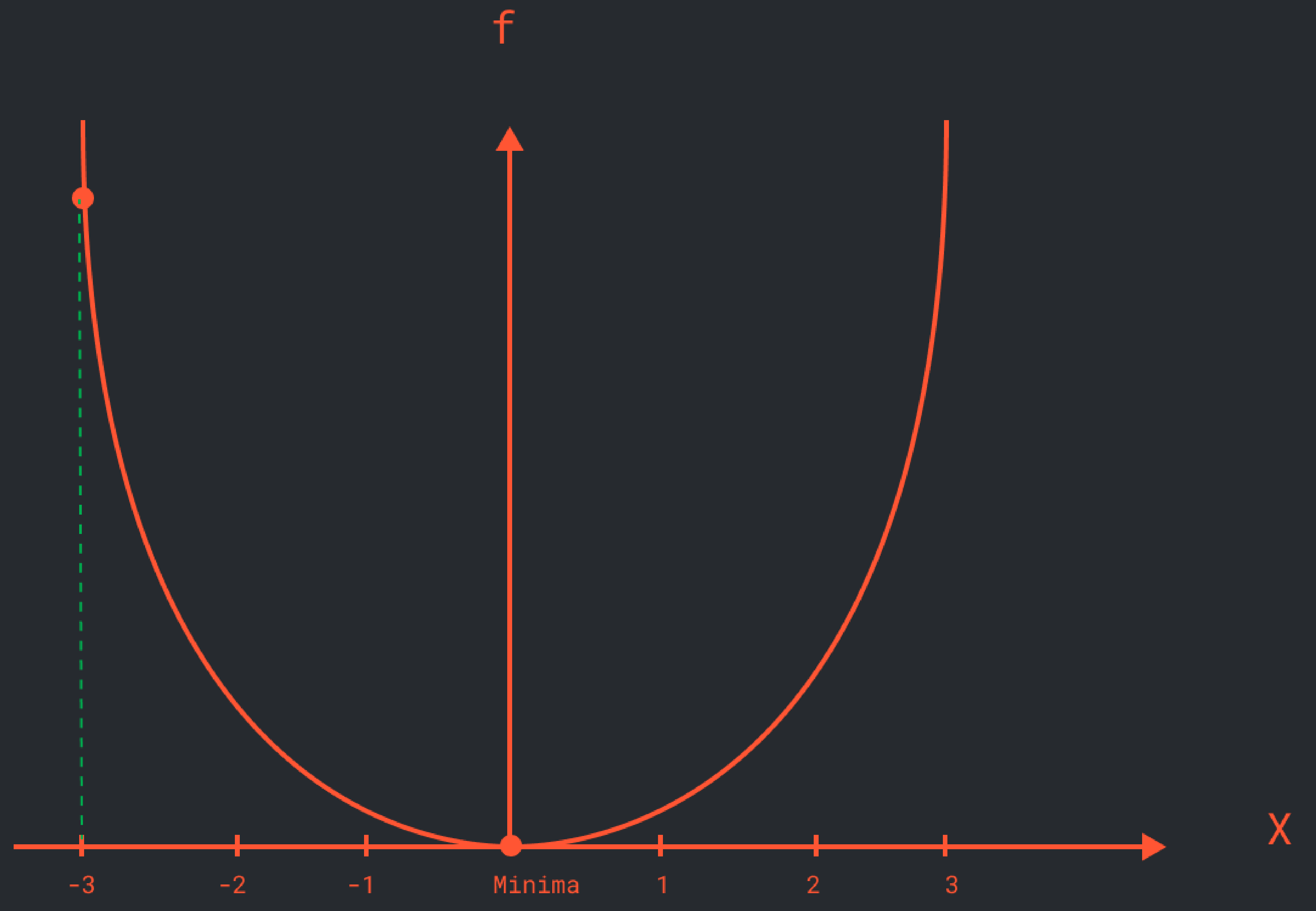
— Инициализируемся в случайной точке X_{start}

— Например, $X_1 = -3$

— $f'(-3) = 2 \cdot x = 2 \cdot (-3) = -6$

— Производная показывает, в каком направлении стоит двигаться, чтобы расти в значении функции!

— Давайте тогда шагнем в обратную сторону



ГРАДИЕНТНЫЙ СПУСК

Функция одной переменной $f(x) = x^2$

— Инициализируемся в случайной точке X_{start}

— Например, $X_1 = -3$

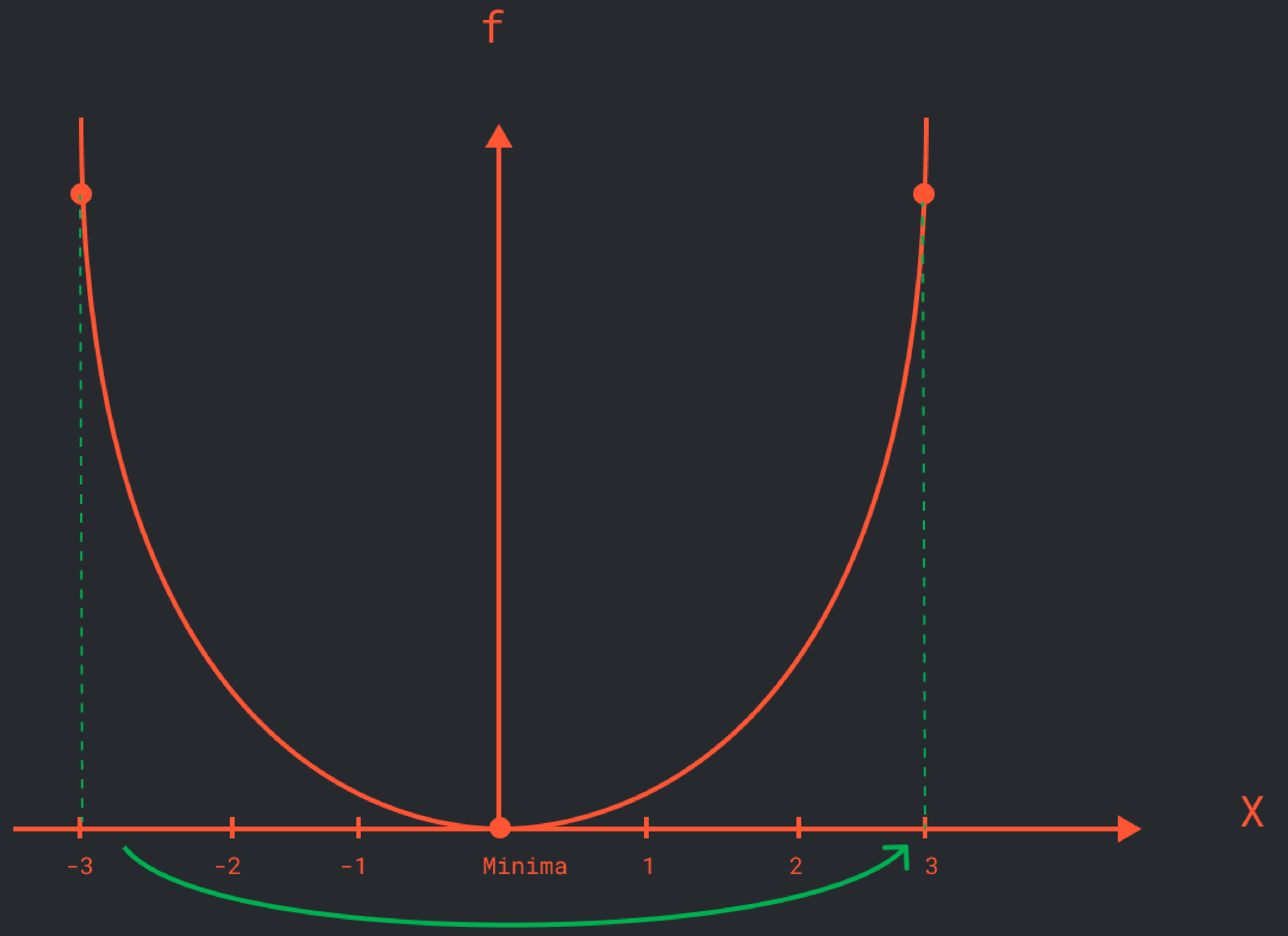
— $f'(-3) = 2 \cdot x = 2 \cdot (-3) = -6$

— Производная показывает, в каком направлении стоит двигаться, чтобы расти в значении функции!

— Давайте тогда шагнем в обратную сторону

— $X_2 = X_1 - f'(X_1) = -3 - (-6) = 3$

— Кажется, немного перескочили!



ГРАДИЕНТНЫЙ СПУСК

Функция одной переменной $f(x) = x^2$

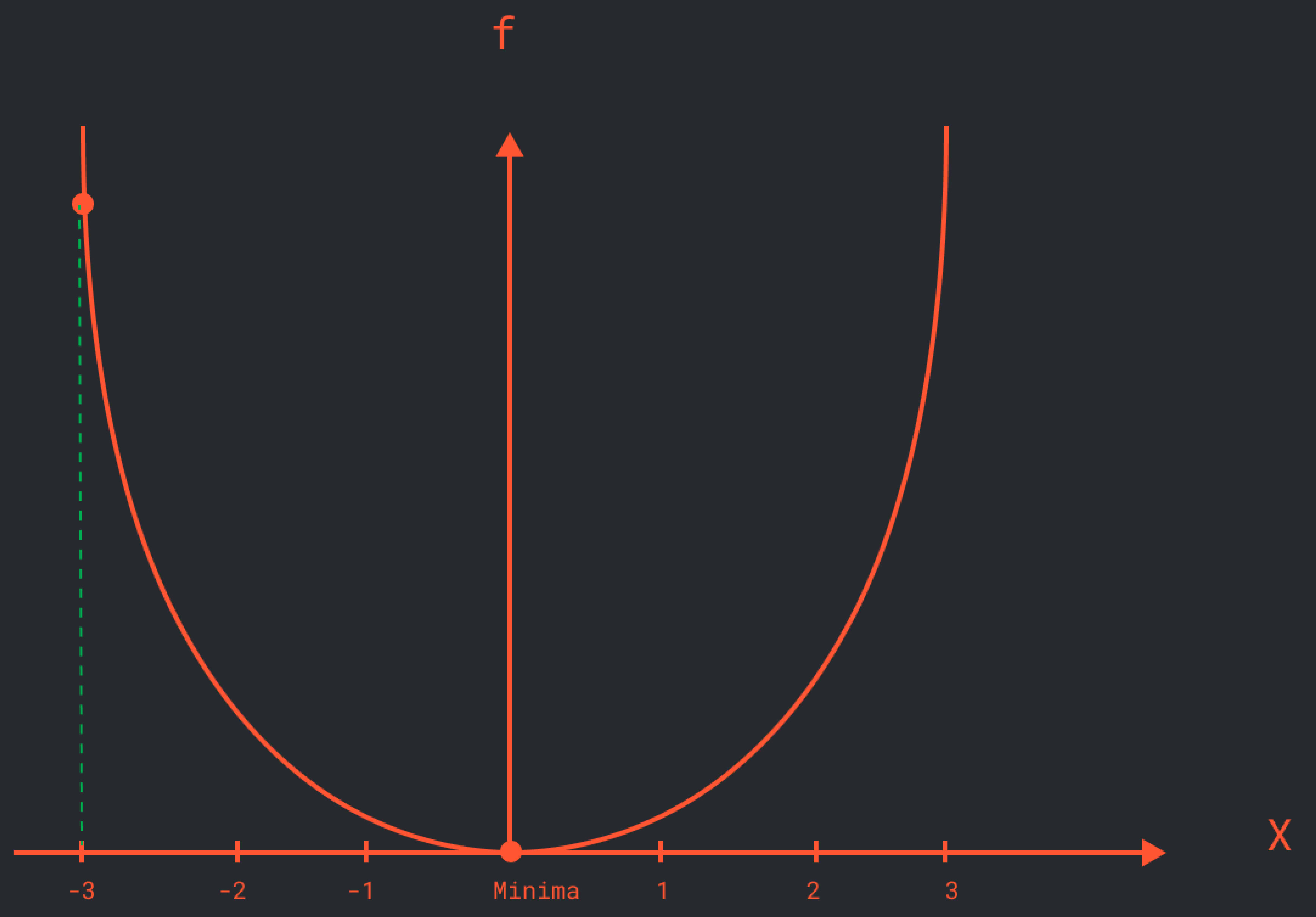
— $X_1 = -3$

— $f'(X_1) = -6$

learning rate

— Нормируем производную на η (к примеру, $\frac{1}{3}$)

— $X_2 = X_1 - \eta \cdot f'(X_1) = -3 - \frac{1}{3} \cdot (-6) = -1$



ГРАДИЕНТНЫЙ СПУСК

Функция одной переменной $f(x) = x^2$

— $X_1 = -3$

— $f'(X_1) = -6$

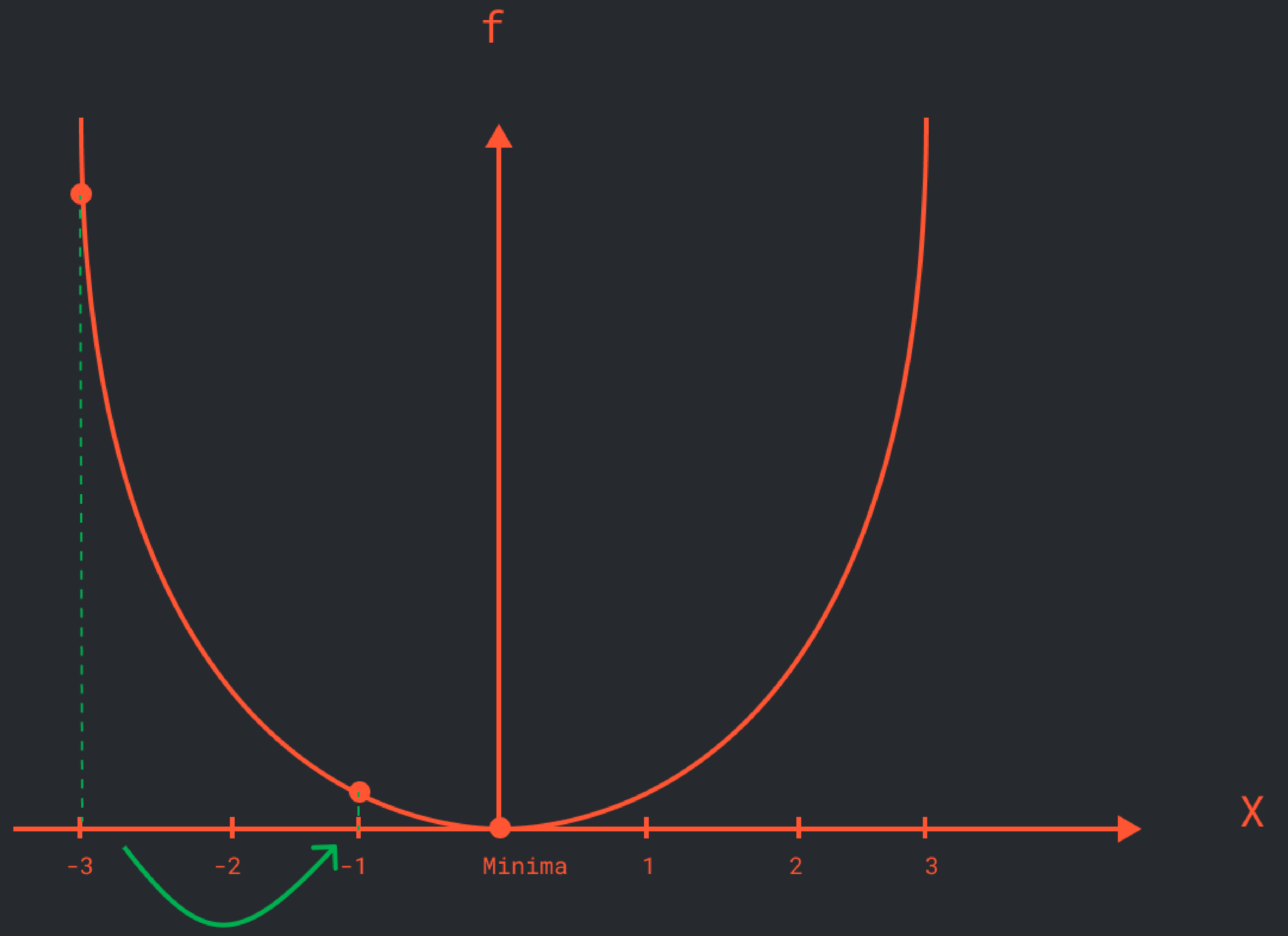
learning rate

— Нормируем производную на η (к примеру, $\frac{1}{3}$)

— $X_2 = X_1 - \eta \cdot f'(X_1) = -3 - \frac{1}{3} \cdot (-6) = -1$

— $f'(X_2) = 2 \cdot (-1) = -2$

— $X_3 = X_2 - \eta \cdot f'(X_2) = -1 - \frac{1}{3} \cdot (-2) = -\frac{1}{3}$



ГРАДИЕНТНЫЙ СПУСК

Функция одной переменной $f(x) = x^2$

— $X_1 = -3$

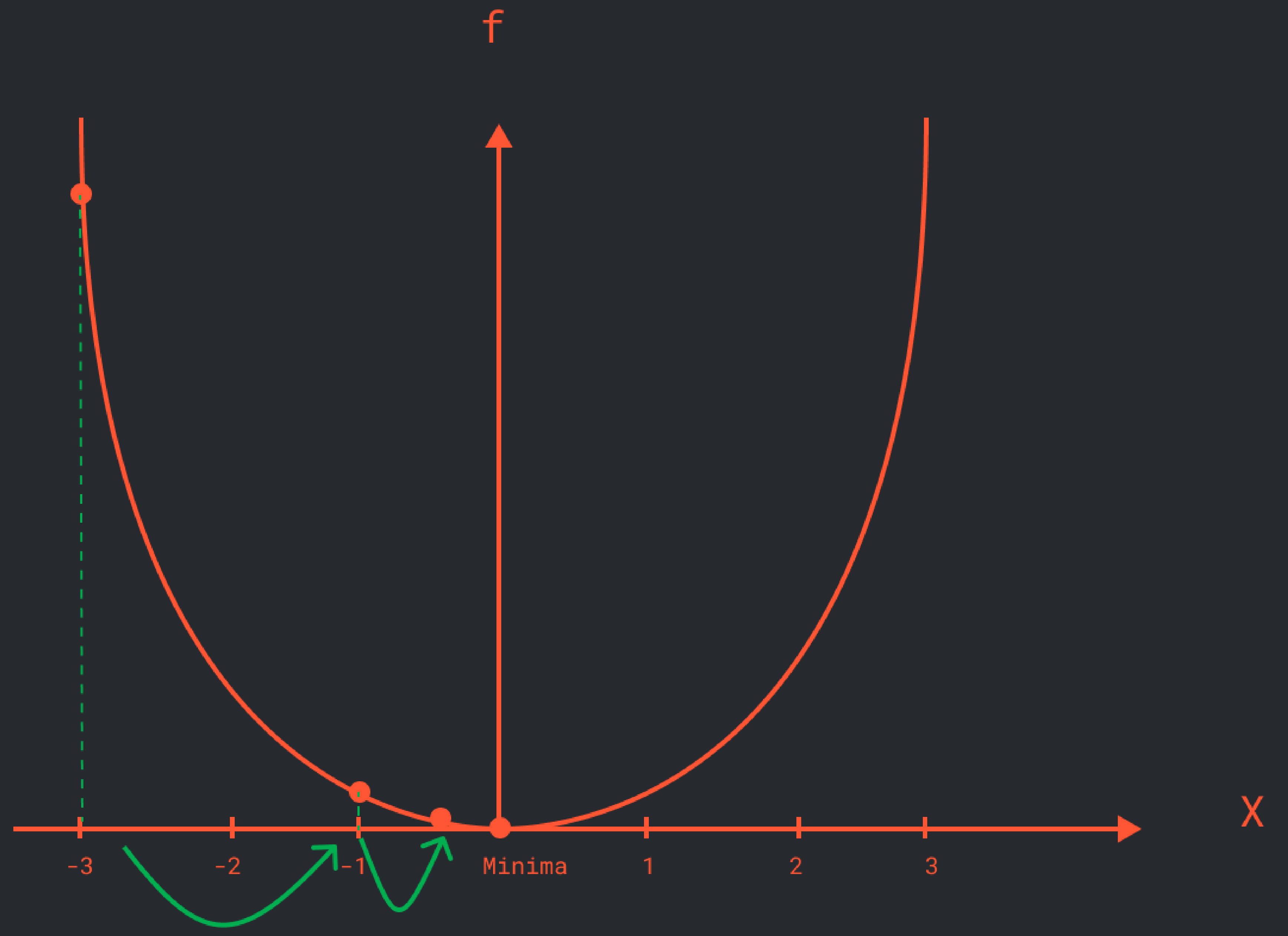
— $f'(X_1) = -6$ $\quad$ *learning rate*

— Нормируем производную на η (к примеру, $\frac{1}{3}$)

— $X_2 = X_1 - \eta \cdot f'(X_1) = -3 - \frac{1}{3} \cdot (-6) = -1$

— $f'(X_2) = 2 \cdot (-1) = -2$

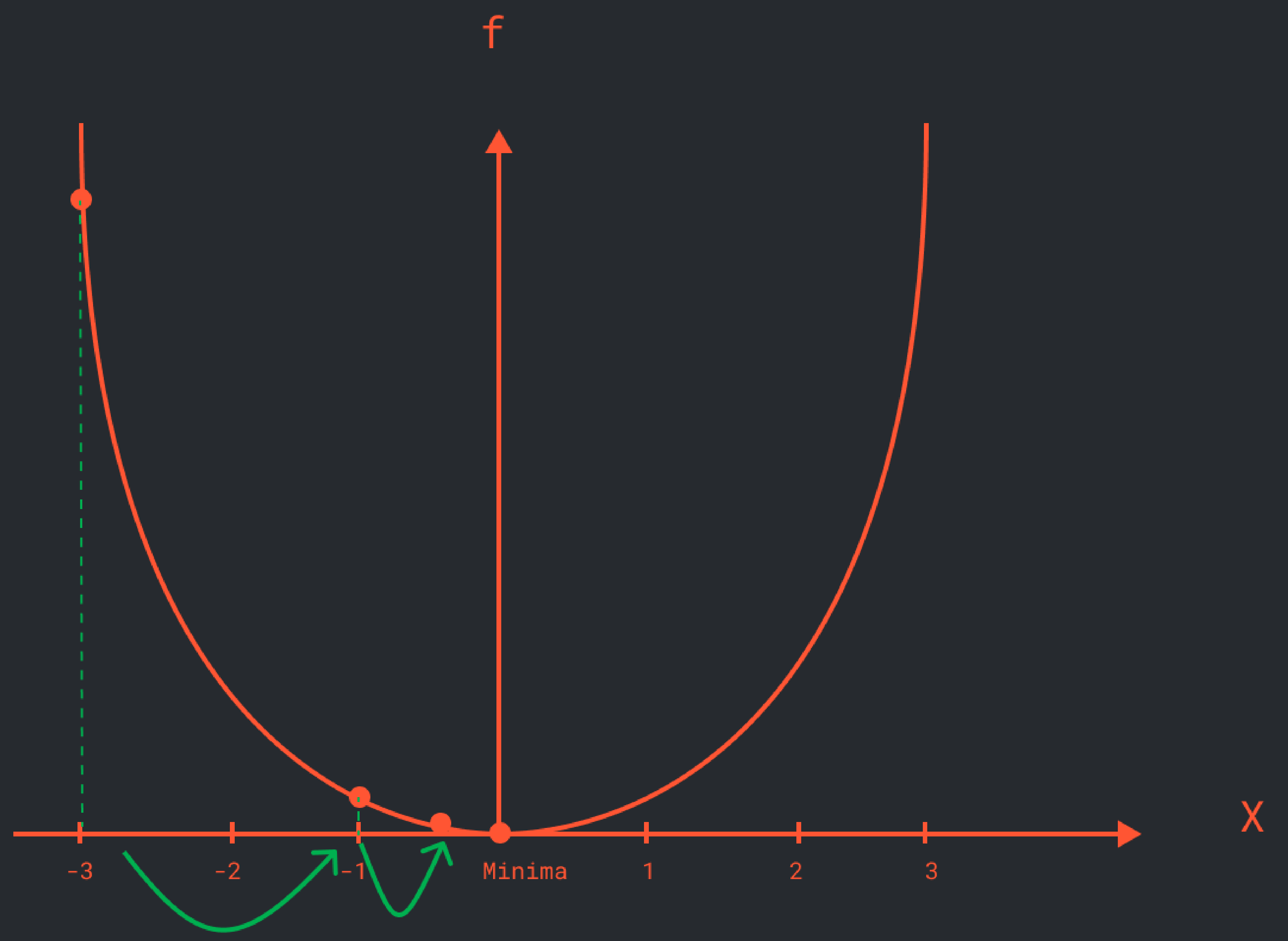
— $X_3 = X_2 - \eta \cdot f'(X_2) = -1 - \frac{1}{3} \cdot (-2) = -\frac{1}{3}$



ГРАДИЕНТНЫЙ СПУСК

Функция одной переменной $f(x) = x^2$

- $X_1 = -3$
- $f'(X_1) = -6$ *learning rate*
- Нормируем производную на η (к примеру, $\frac{1}{3}$)
- $X_2 = X_1 - \eta \cdot f'(X_1) = -3 - \frac{1}{3} \cdot (-6) = -1$
- $f'(X_2) = 2 \cdot (-1) = -2$
- $X_3 = X_2 - \eta \cdot f'(X_2) = -1 - \frac{1}{3} \cdot (-2) = -\frac{1}{3}$
- Можем продолжить!



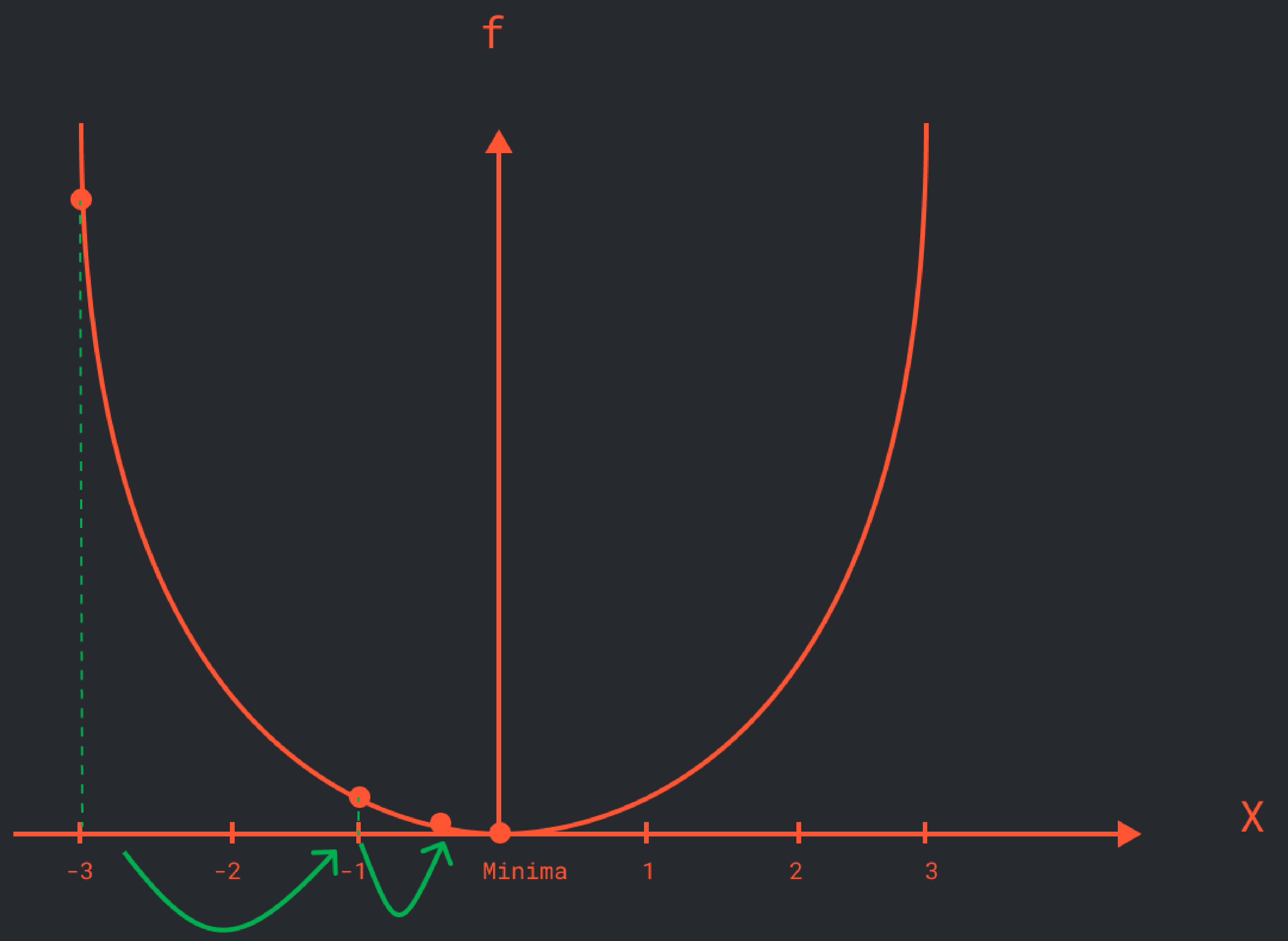
ГРАДИЕНТНЫЙ СПУСК

Функция одной переменной $f(x) = x^2$

- $X_1 = -3$
- $f'(X_1) = -6$

learning rate

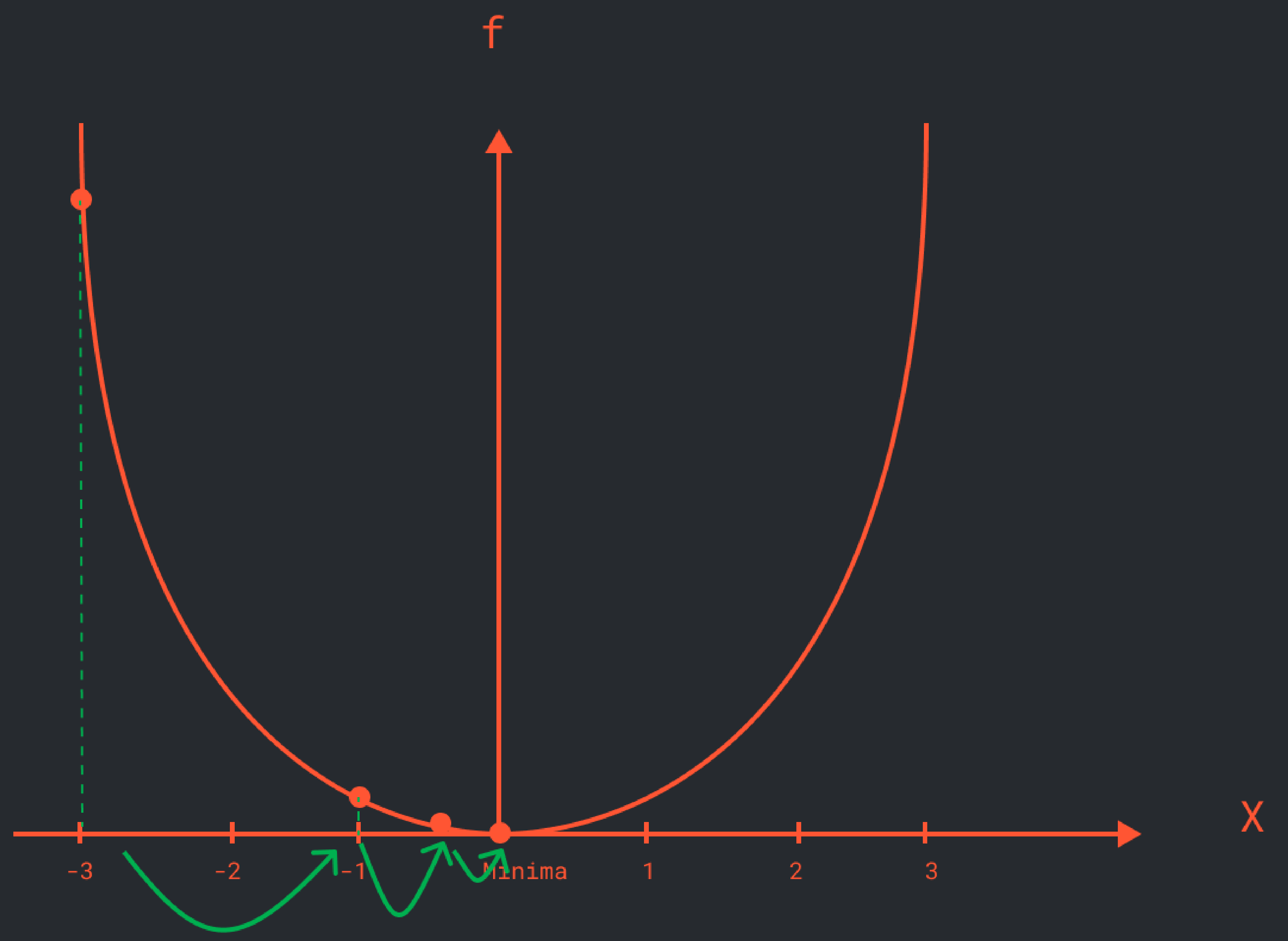
- Нормируем производную на η (к примеру, $\frac{1}{3}$)
- $X_2 = X_1 - \eta \cdot f'(X_1) = -3 - \frac{1}{3} \cdot (-6) = -1$
- $f'(X_2) = 2 \cdot (-1) = -2$
- $X_3 = X_2 - \eta \cdot f'(X_2) = -1 - \frac{1}{3} \cdot (-2) = -\frac{1}{3}$
- Можем продолжить!
- $X_4 = X_3 - \eta \cdot f'(X_3) = -\frac{1}{3} - \frac{1}{3} \cdot \left(-\frac{2}{3}\right) = -\frac{1}{9}$
- $X_5 = X_4 - \eta \cdot f'(X_4) = -\frac{1}{9} - \frac{1}{3} \cdot \left(-\frac{2}{9}\right) = -\frac{1}{27}$



ГРАДИЕНТНЫЙ СПУСК

Функция одной переменной $f(x) = x^2$

- $X_4 = X_3 - \eta \cdot f'(X_3) = -\frac{1}{3} - \frac{1}{3} \cdot \left(-\frac{2}{3}\right) = -\frac{1}{9}$
- $X_5 = X_4 - \eta \cdot f'(X_4) = -\frac{1}{9} - \frac{1}{3} \cdot \left(-\frac{2}{9}\right) = -\frac{1}{27}$
- …
- $X_9 = X_8 - \eta \cdot f'(X_8) = -\frac{1}{2187} \approx 0$
- $X_{10} = X_9 - \eta \cdot f'(X_9) = -\frac{1}{6561} \approx 0$
- Придумаем критерий, когда останавливаться!



ГРАДИЕНТНЫЙ СПУСК

Функция одной переменной $f(x) = x^2$

— $X_9 = X_8 - \eta \cdot f'(X_8) = -\frac{1}{2187} \approx 0$

— $X_{10} = X_9 - \eta \cdot f'(X_9) = -\frac{1}{6561} \approx 0$

— Придумаем критерии, когда останавливаться!

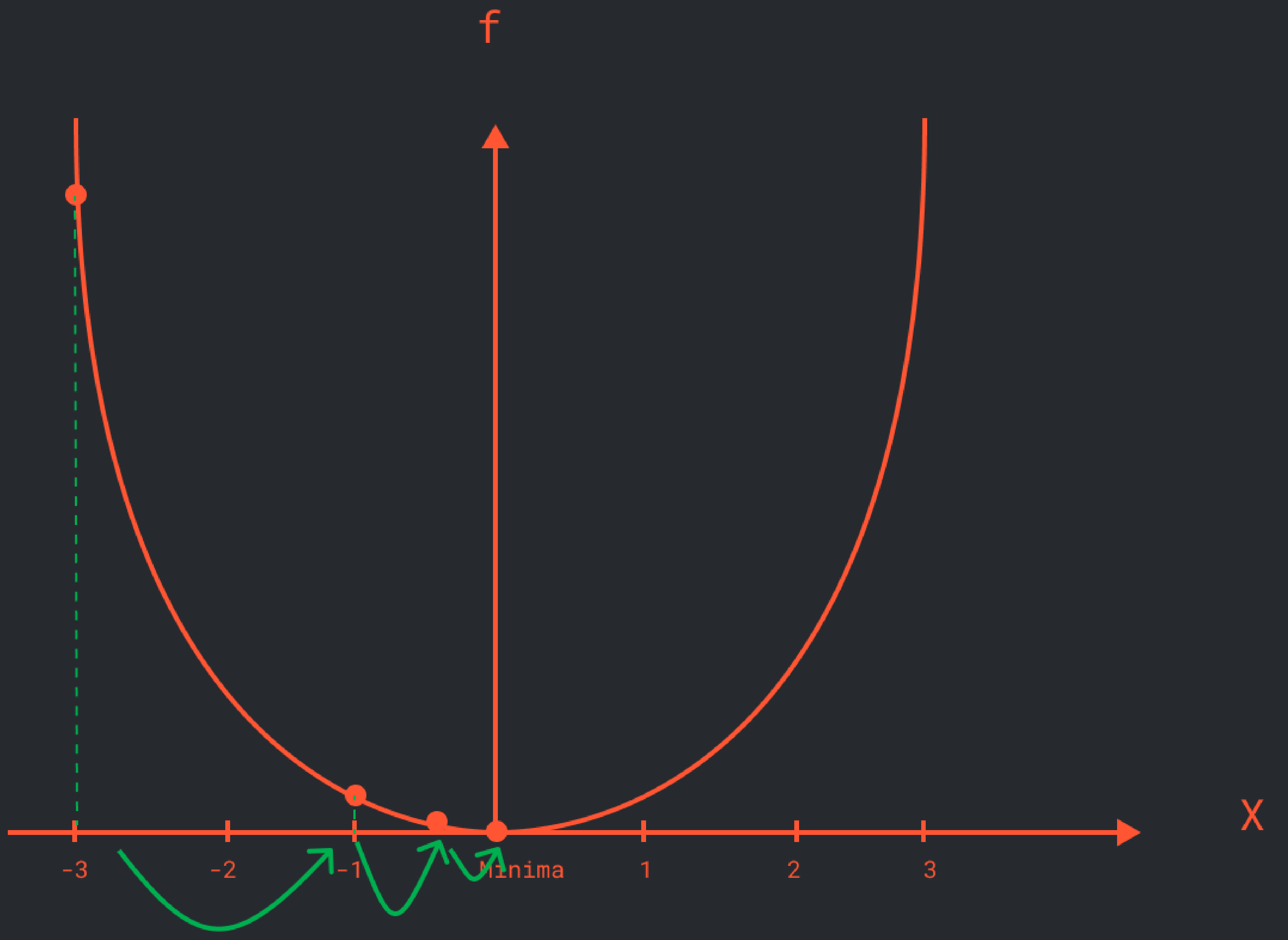
— Вариант 1: маленькое значение производной

— $|f'(X_i)| < \xi$

— $|f'(X_{10})| = |2 \cdot \left(-\frac{1}{6561}\right)| \approx 0{,}0003$

— Пусть наш порог $\xi = 0{,}0003$

— Тогда на 10 – ой итерации СТОП!



ГРАДИЕНТНЫЙ СПУСК

Функция одной переменной $f(x) = x^2$

— $X_9 = X_8 - \eta \cdot f'(X_8) = -\frac{1}{2187} \approx 0$

— $X_{10} = X_9 - \eta \cdot f'(X_9) = -\frac{1}{6561} \approx 0$

— Придумаем критерии, когда останавливаться!

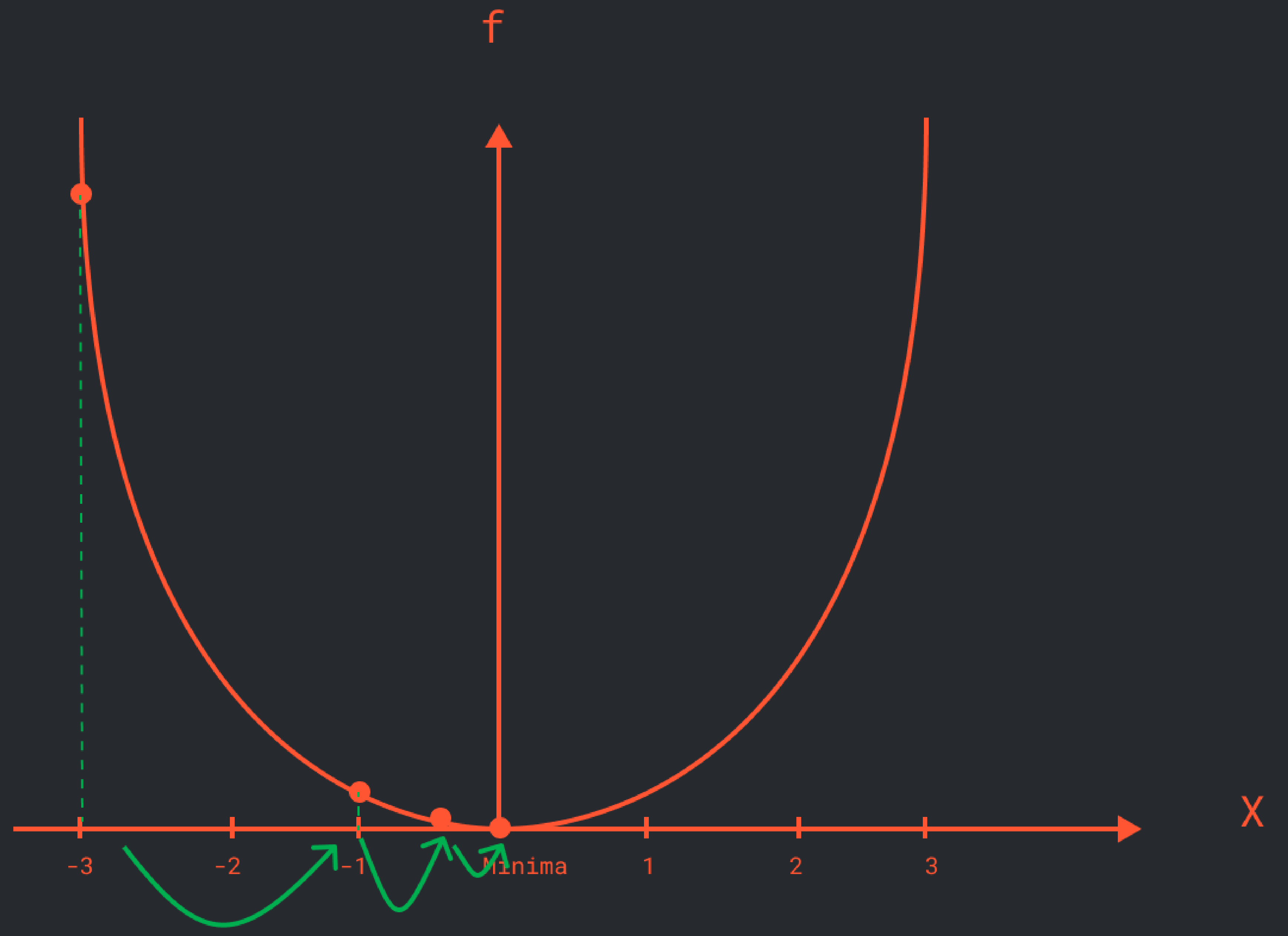
— Вариант 2: маленький шаг

— $|X_{i+1} - X_i| < \xi$

— $|X_{10} - X_9| = \left| -\frac{1}{6561} - \left(-\frac{1}{2187}\right) \right| \approx 0{,}0003$

— Пусть наш порог $\xi = 0{,}0003$

— Тогда на 10 – ой итерации СТОП!



ГРАДИЕНТНЫЙ СПУСК

Функция одной переменной $f(x) = x^2$

— $X_9 = X_8 - \eta \cdot f'(X_8) = -\frac{1}{2187} \approx 0$

— $X_{10} = X_9 - \eta \cdot f'(X_9) = -\frac{1}{6561} \approx 0$

— Придумаем критерии, когда останавливаться!

— Вариант 3: маленькое изменение в $f(x)$

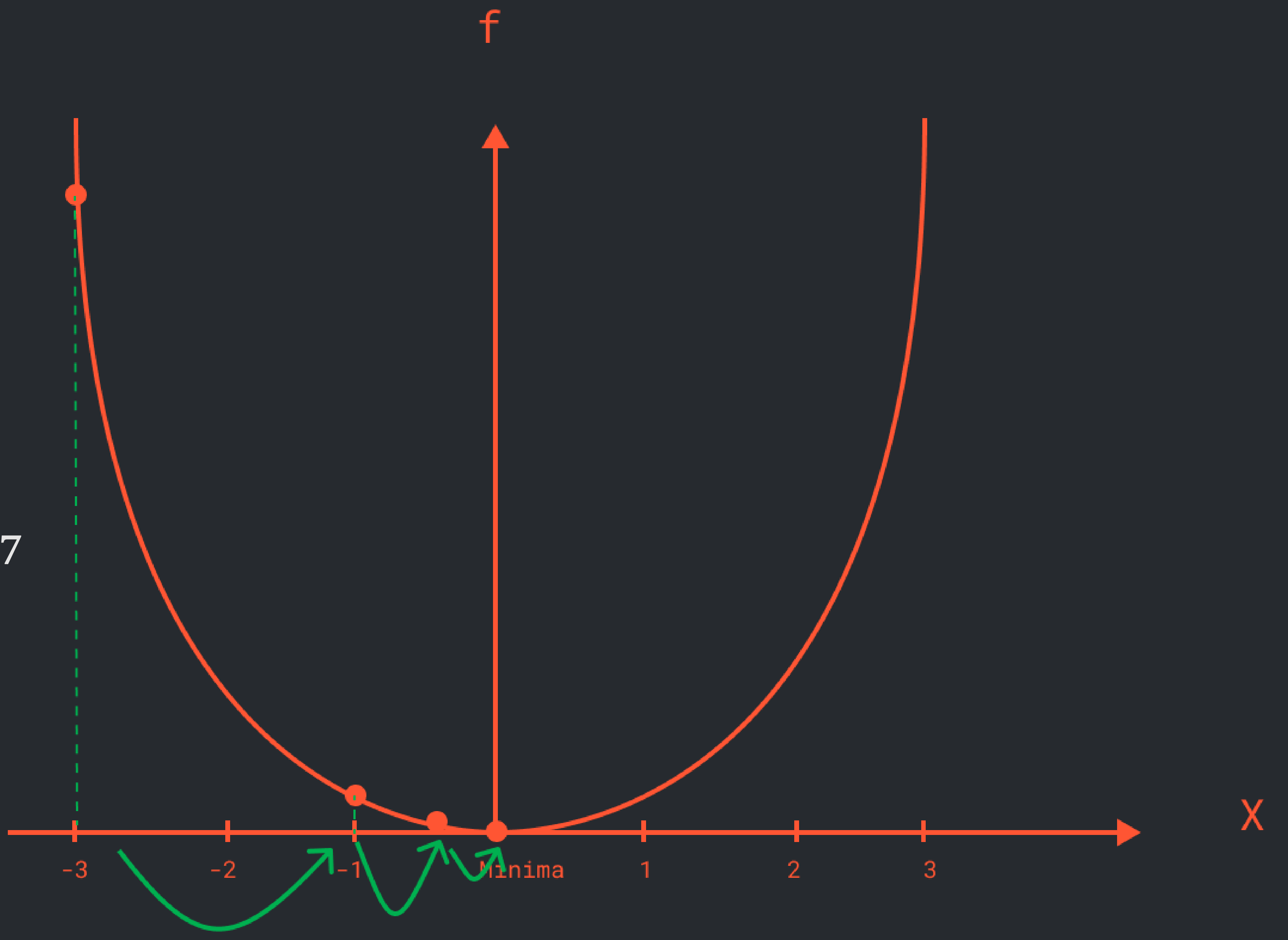
— $|f(X_{i+1}) - f(X_i)| < \xi$

— $|f(X_{10}) - f(X_9)| = \left|\left(-\frac{1}{6561}\right)^2 - \left(-\frac{1}{2187}\right)^2\right| \approx 10^{-7}$

— Пусть наш порог опять какое-то маленькое число.

— Допустим, $0{,}001 = 10^{-3}$

— Тогда уж точно останавливаемся.



ГРАДИЕНТНЫЙ СПУСК

Функция одной переменной f(x)

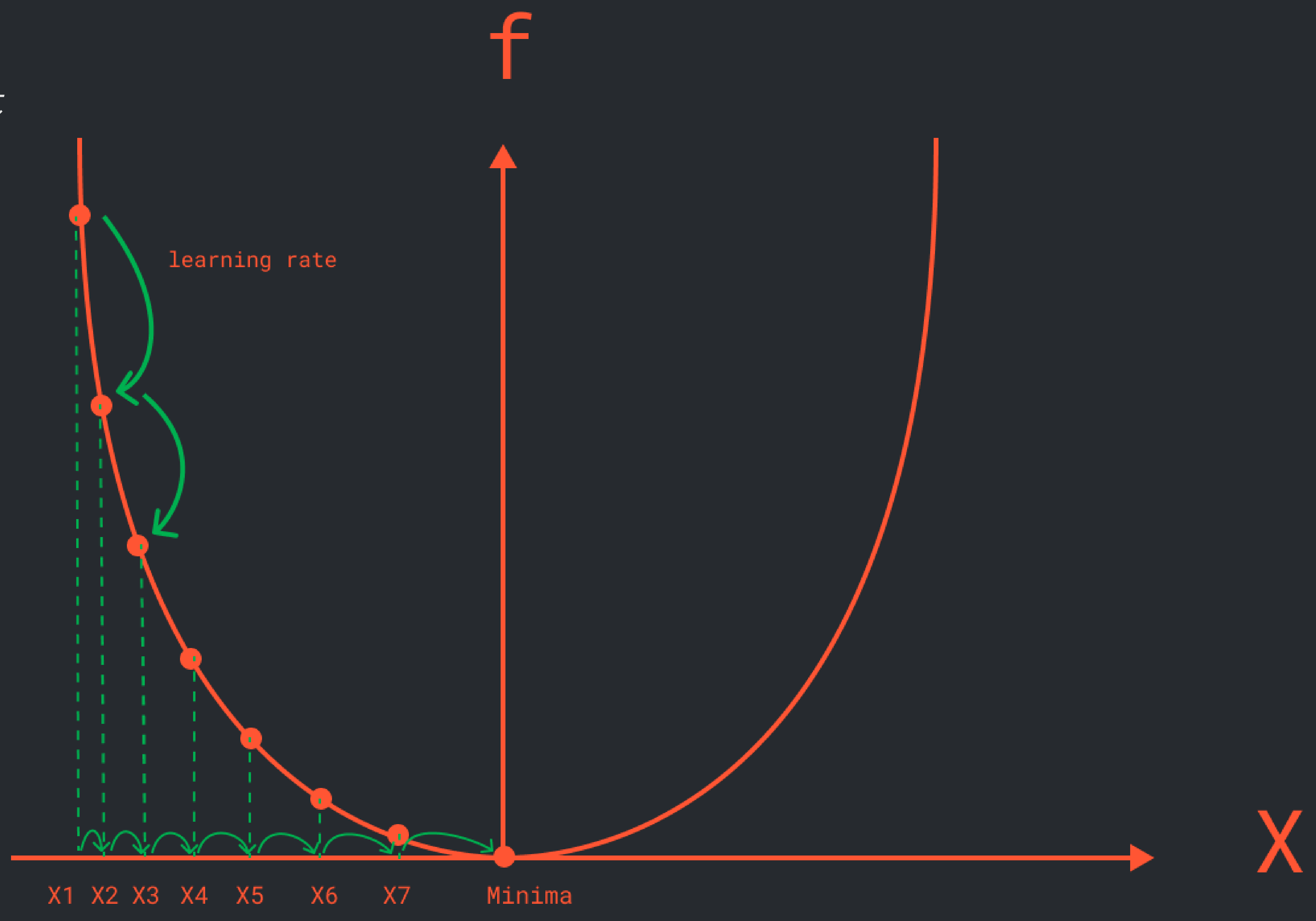
— Инициализируемся в случайной точке X_{start}

— До сходимости:

$$\text{step} = f'(X_{start})$$
$$X_{next} = X_{start} - \eta_i \cdot step$$
$$X_{start} = X_{next}$$

— Три варианта порога (threshold):

$$|f'(\mathrm{X_{start}})| \leq \xi$$
$$|f(X_{next}) - f(X_{start})| \leq \xi$$
$$|X_{start} - X_{next}| \leq \xi$$



РЕЗЮМЕ

— Узнали, как работает градиентный спуск

— Изучили несколько вариантов критерия останова

— Умеем находить локальные минимумы произвольных функций одной переменной!

— А что с функциями нескольких переменных?

— Нам понадобится знать, что такое линия уровня и вектор-градиент

ЛИКБЕЗ №1: ЛИНИИ УРОВНЯ

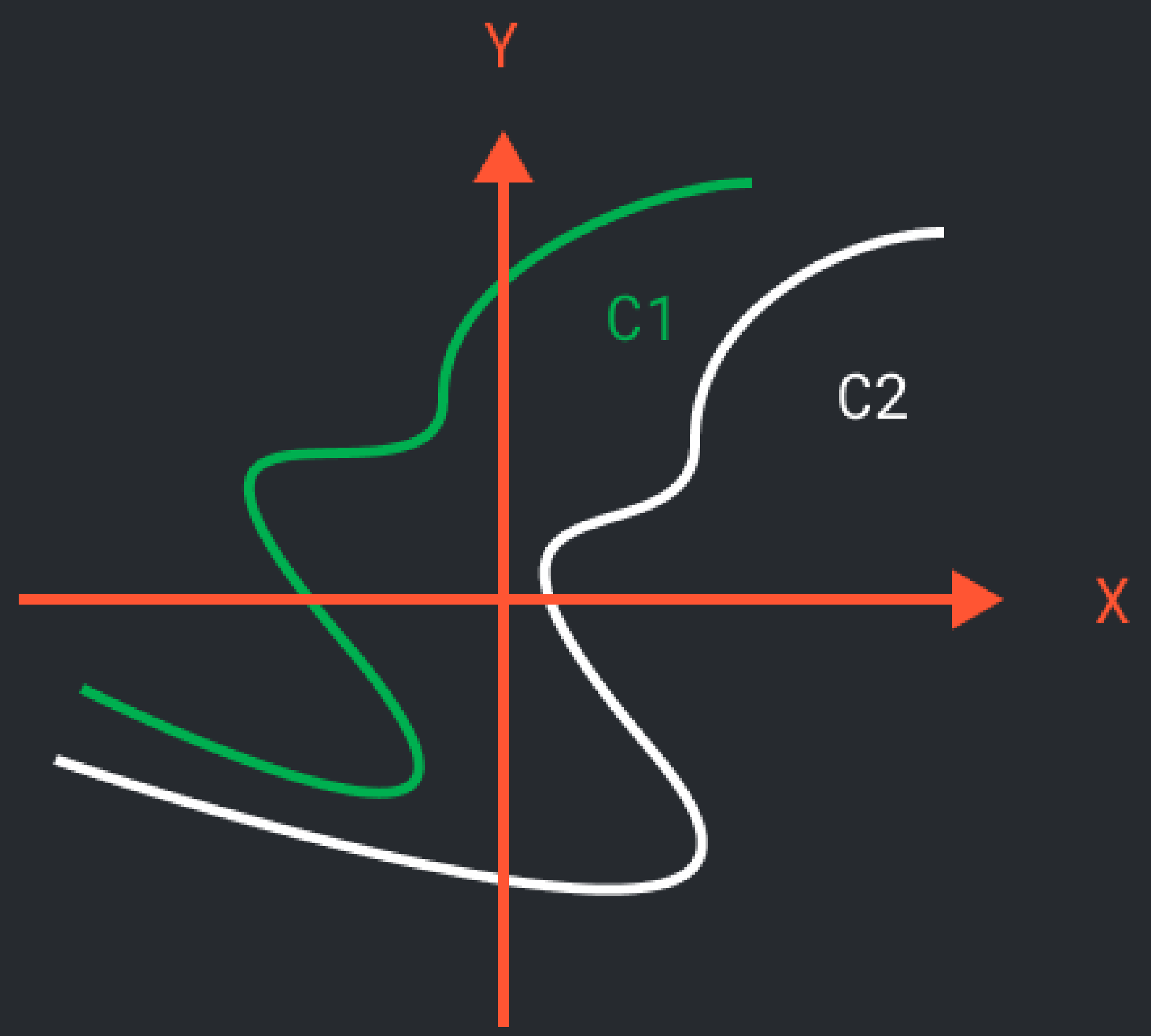
— Обычно единое значение C некоторой функции дают сразу много точек

— Можно зафиксировать C и найти это множество

— Можно его даже изобразить!

— Будем называть это линией уровня $z(x, y) = C$

— Очевидно, они не могут пересекаться



ЛИКБЕЗ №1: ЛИНИИ УРОВНЯ, ПРИМЕР

- Пусть имеем $z(x, y) = x^2 + y^2$
- $C = 16$: $x^2 + y^2 = 16$

ЛИКБЕЗ №1: ЛИНИИ УРОВНЯ, ПРИМЕР

— Пусть имеем $z(x, y) = x^2 + y^2$

— $C = 16$: $x^2 + y^2 = 16$

— $C = 1$: $x^2 + y^2 = 1$

ЛИКБЕЗ №1: ЛИНИИ УРОВНЯ, ПРИМЕР

- Пусть имеем $z(x,y) = x^2 + y^2$
- $C = 16$: $x^2 + y^2 = 16$
- $C = 1$: $x^2 + y^2 = 1$
- $C = 0$: $x^2 + y^2 = 0$

ЛИКБЕЗ №1: ЛИНИИ УРОВНЯ, ПРИМЕР

- Пусть имеем $z(x, y) = x^2 + y^2$
- $C = 16$: $x^2 + y^2 = 16$
- $C = 1$: $x^2 + y^2 = 1$
- $C = 0$: $x^2 + y^2 = 0$
- Ищем экстремумы – значит перепрыгиваем линии

ЛИКБЕЗ №1: ЛИНИИ УРОВНЯ, ПРИМЕР

- Пусть имеем $z(x, y) = x^2 + y^2$
- $C = 16$: $x^2 + y^2 = 16$
- $C = 1$: $x^2 + y^2 = 1$
- $C = 0$: $x^2 + y^2 = 0$
- Ищем экстремумы – значит перепрыгиваем линии
- Как, находясь на какой-то линии уровня, понять, в какую сторону сделать шаг для поиска минимума?

ЛИКБЕЗ №1: ГРАДИЕНТ ФУНКЦИИ

— Градиент – это просто вектор из производных 1го порядка

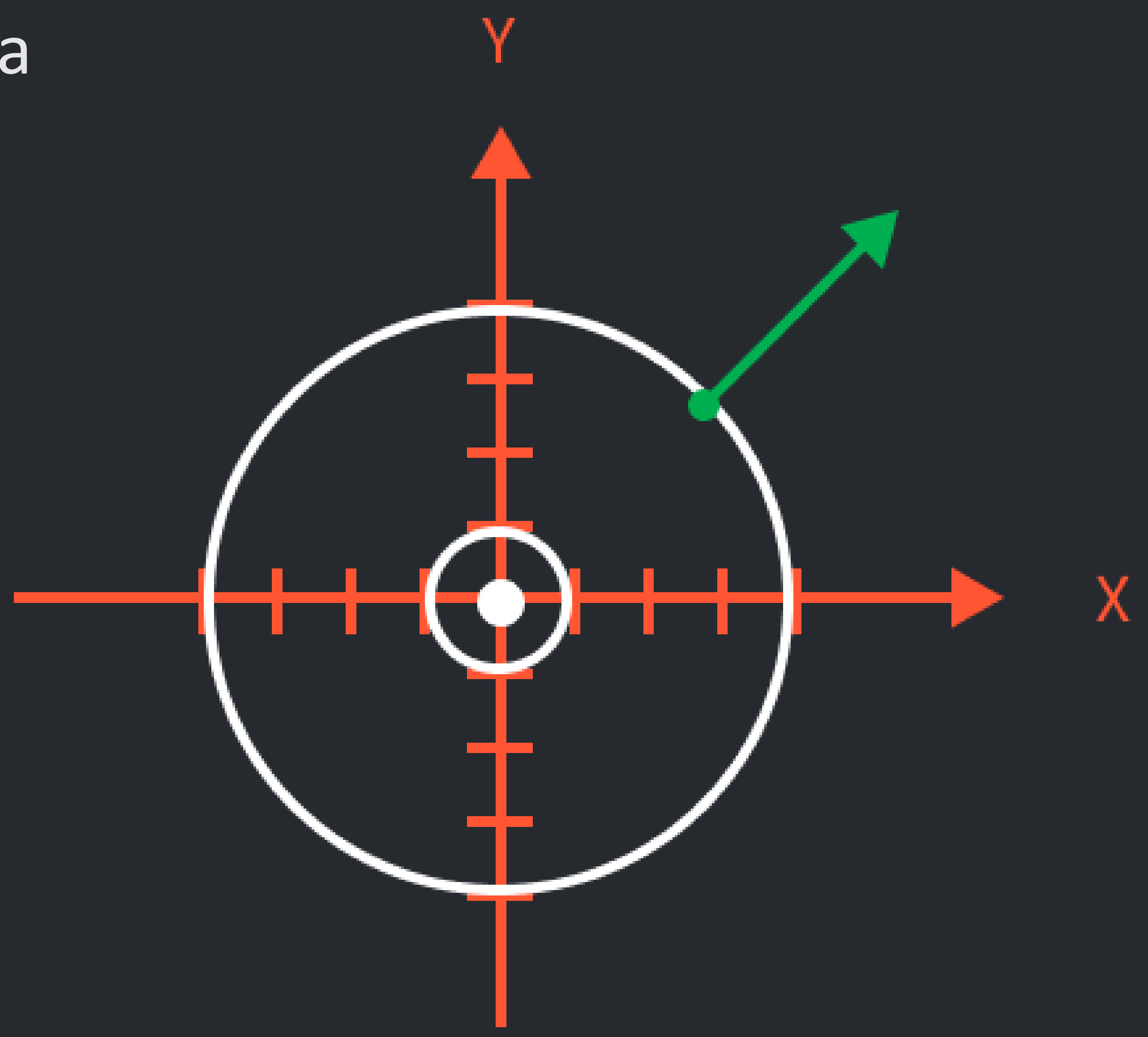
— $z = x^2 + y^2$

— $\nabla z = (2 \cdot x \quad 2 \cdot y)$

— $\nabla z(2\sqrt{2}; 2\sqrt{2}) = (4\sqrt{2} \quad 4\sqrt{2})$

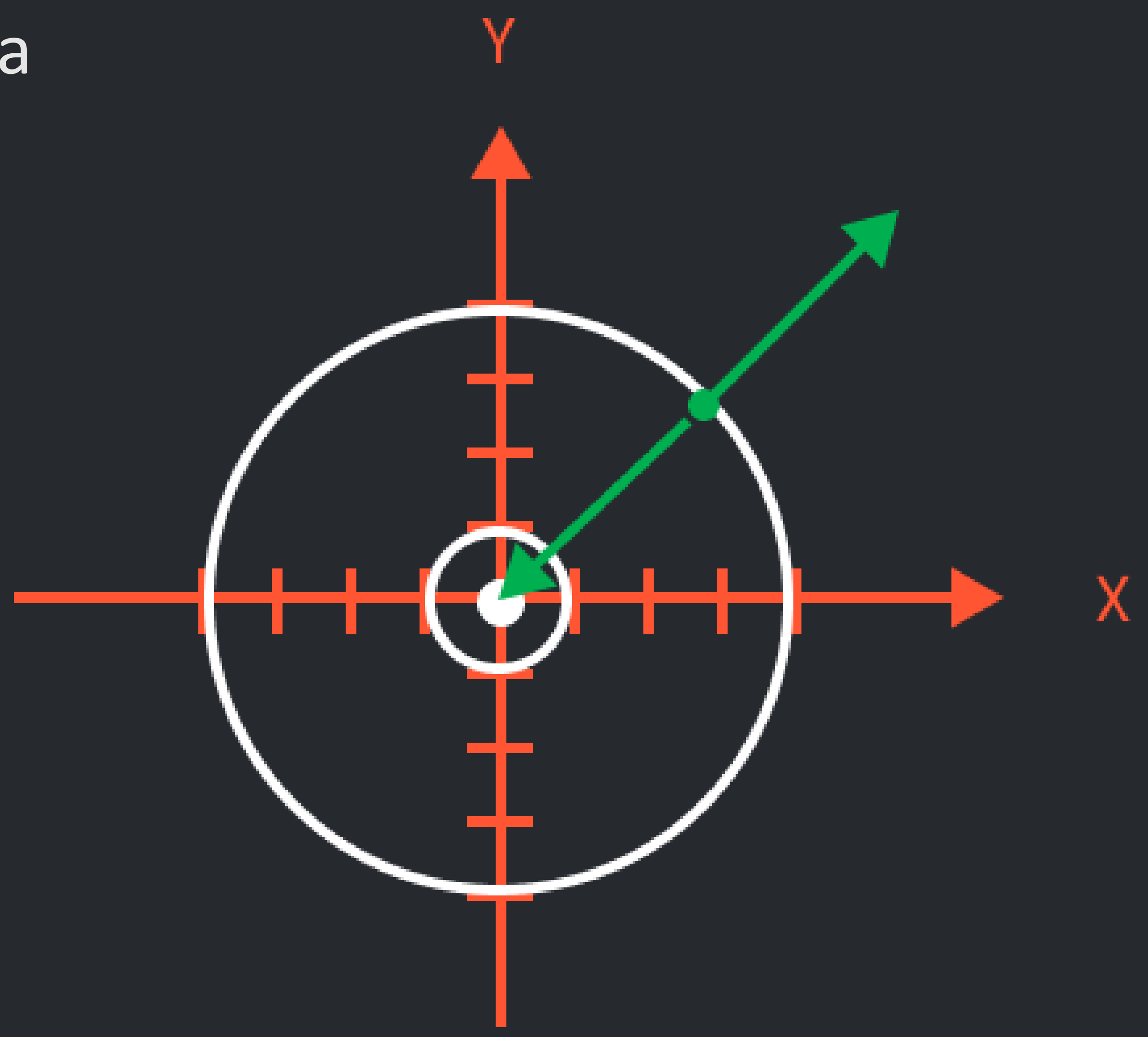
ЛИКБЕЗ №1: ГРАДИЕНТ ФУНКЦИИ

- Градиент – это просто вектор из производных 1го порядка
- $z = x^2 + y^2$
- $\nabla z = (2 \cdot x \quad 2 \cdot y)$
- $\nabla z(2\sqrt{2}; 2\sqrt{2}) = (4\sqrt{2} \quad 4\sqrt{2})$
- Показывает направление быстрейшего роста
- $-\nabla z(2\sqrt{2}; 2\sqrt{2}) = (-4\sqrt{2} \quad -4\sqrt{2})$
- Антиградиент – направление наискорейшего убывания



ЛИКБЕЗ №1: ГРАДИЕНТ ФУНКЦИИ

- Градиент – это просто вектор из производных 1го порядка
- $z = x^2 + y^2$
- $\nabla z = (2 \cdot x \quad 2 \cdot y)$
- $\nabla z(2\sqrt{2}; 2\sqrt{2}) = (4\sqrt{2} \quad 4\sqrt{2})$
- Показывает направление быстрейшего роста
- $-\nabla z(2\sqrt{2}; 2\sqrt{2}) = (-4\sqrt{2} \quad -4\sqrt{2})$
- Антиградиент – направление наискорейшего убывания
- Градиент $\perp$ касательной к линии уровня



ЛИКБЕЗ №1: ГРАДИЕНТ ФУНКЦИИ

— Градиент – это просто вектор из производных 1го порядка

— $z = x^2 + y^2$

— $\nabla z = (2 \cdot x \quad 2 \cdot y)$

— $\nabla z(2\sqrt{2}; 2\sqrt{2}) = (4\sqrt{2} \quad 4\sqrt{2})$

— Показывает направление быстрейшего роста

— $-\nabla z(2\sqrt{2}; 2\sqrt{2}) = (-4\sqrt{2} \quad -4\sqrt{2})$

— Антиградиент – направление наискорейшего убывания

— Градиент $\perp$ касательной к линии уровня

ЛИКБЕЗ №1: ГРАДИЕНТ ФУНКЦИИ

— Пример итерации градиентного спуска:

— $z = x^2 + y^2$

— $X_{start} = (2\sqrt{2}; 2\sqrt{2})$

— $\nabla z = (2 \cdot x \quad 2 \cdot y)$

— $\nabla z\left(2\sqrt{2}; 2\sqrt{2}\right) = \left(4\sqrt{2} \quad 4\sqrt{2}\right)$

— $\mathrm{X_{next}} = \mathrm{X_{start}} - \eta \cdot \nabla z(X_{start})$

— $\mathrm{X_{next}} = \left(2\sqrt{2}; 2\sqrt{2}\right) - \frac{1}{4} \cdot \left(4\sqrt{2} \quad 4\sqrt{2}\right) = (\sqrt{2}; \sqrt{2})$

— И так до срабатывания критерия остановы!

— Как они выглядят теперь?

ГРАДИЕНТНЫЙ СПУСК

Функция многих переменных $z(x) = x^2 + y^2$

— Вариант 1: маленькая длина градиента

— $|\nabla z(X_i)| < \xi$

— $|\nabla z(X_{start})| = |(4\sqrt{2} \quad 4\sqrt{2})| = \sqrt{\left(4\sqrt{2}\right)^2 + \left(4\sqrt{2}\right)^2} = 8$

— Пусть наш порог $\xi = 0{,}01$

— $8 > 0{,}01$

— Тогда продолжаем!

ГРАДИЕНТНЫЙ СПУСК

Функция многих переменных $z(x) = x^2 + y^2$

— Вариант 2: маленькая длина шага

— $|\mathrm{X}_{\mathrm{start}} - \mathrm{X}_{next}| < \xi$

— $|\mathrm{X}_{\mathrm{start}} - \mathrm{X}_{next}| = |(2\sqrt{2}\,; 2\sqrt{2}) - (\sqrt{2};\ \sqrt{2})|$

$$|\sqrt{2}; \sqrt{2}| = \sqrt{(\sqrt{2})^2 + (\sqrt{2})^2} = \sqrt{4}$$

— Пусть наш порог $\xi = 0{,}01$

— $\sqrt{4} > 0{,}01$

— Тогда продолжаем!

ГРАДИЕНТНЫЙ СПУСК

Функция многих переменных $z(x) = x^2 + y^2$

— Вариант 3: маленькое изменении в значении функции

— $|f(\mathrm{X}_{\mathrm{start}}) - f(\mathrm{X}_{next})| < \xi$

— $f(\mathrm{X}_{\mathrm{start}}) - f(\mathrm{X}_{next})| = \left|f\left(2\sqrt{2}\,; 2\sqrt{2}\right) - f\left(\sqrt{2}; \sqrt{2}\right)\right| =$

$$\left|\left(\left(2\sqrt{2}\right)^2 + \left(2\sqrt{2}\right)^2\right) - \left(\left(\sqrt{2}\right)^2 + \left(\sqrt{2}\right)^2\right)\right| =$$

$|16 - 4| = 12$

— Пусть наш порог $\xi = 0{,}01$

— $12 > 0{,}01$

— Тогда продолжаем!

ГРАДИЕНТНЫЙ СПУСК

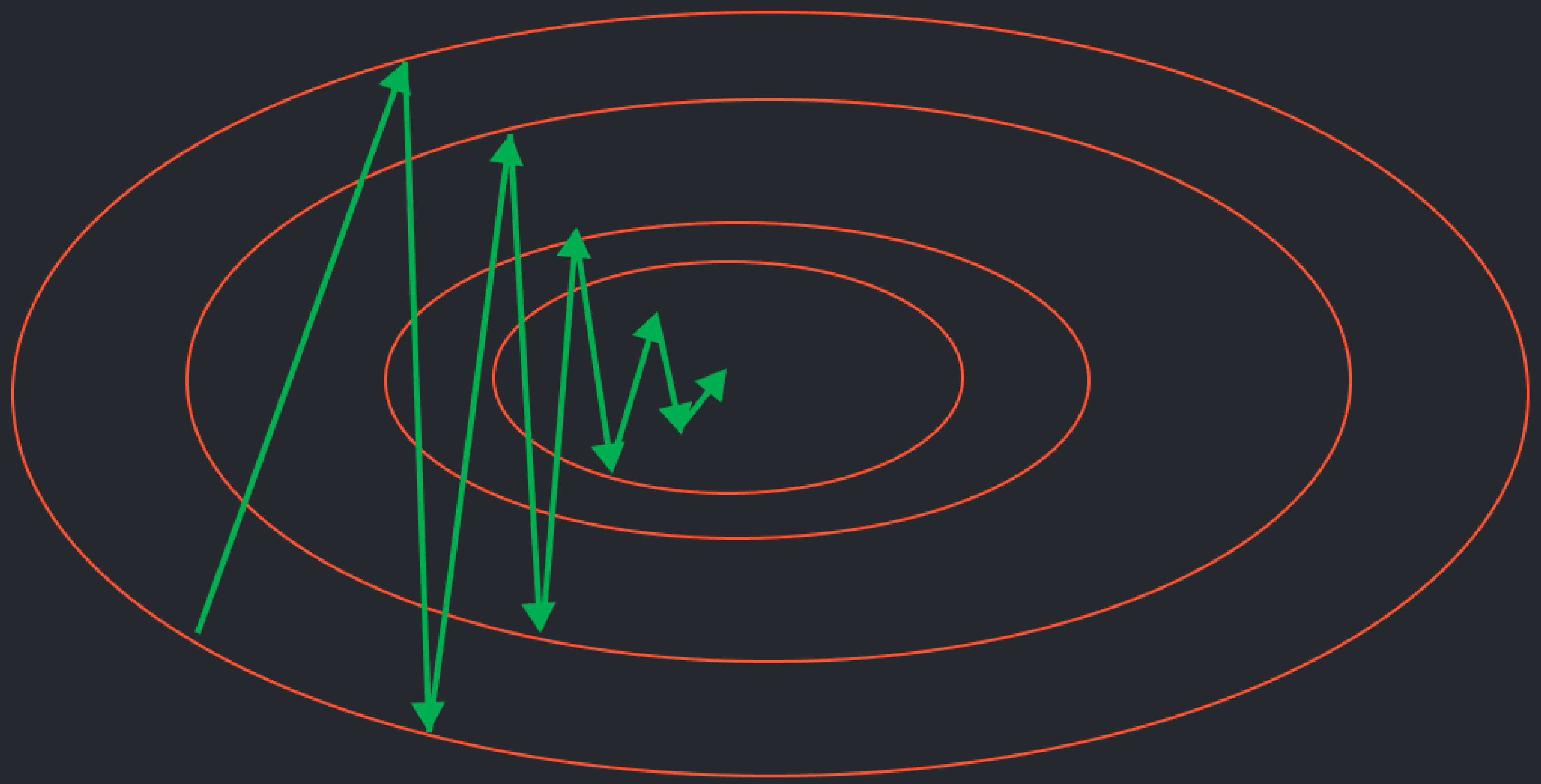
— Инициализируемся в случайной точке X_{start}

— До сходимости:

$$\text{step} = \nabla z'(X_{start})$$
$$X_{next} = X_{start} - \eta_i \cdot step$$
$$X_{start} = X_{next}$$

— Три варианта порога (threshold):

$$||\nabla z(\mathrm{X_{start}})|| \leq \xi$$
$$|f(X_{next}) - f(X_{start})| \leq \xi$$
$$|X_{start} - X_{next}| \leq \xi$$



ГРАДИЕНТНЫЙ СПУСК: ЛИНЕЙНАЯ РЕГРЕССИЯ

- Пусть имеем n объектов и m признаков
- $Q = \frac{1}{n}\sum_i^n (a(x_i) - y_i)^2 = \frac{1}{n}\sum_i^n \left(\beta_1 \cdot d_1^i + \dots + \beta_m \cdot d_m^i - y_i\right)^2$
- $Q'_{\beta_1} = \frac{2}{n} \cdot \sum_i^n d_1^i \cdot (\beta_1 \cdot d_1^i + \dots + \beta_m \cdot d_m^i - y_i)$
- …
- $Q'_{\beta_m} = \frac{2}{n} \cdot \sum_i^n d_m^i \cdot (\beta_1 \cdot d_1^i + \dots + \beta_m \cdot d_m^i - y_i)$
- $\nabla Q = (Q'_{\beta_1} \dots Q'_{\beta_m})$
- $X_{next} = X_{start} - \eta \cdot \nabla Q$

ПРИМЕР

$$
\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array}
\begin{array}{c} \begin{array}{cc} d_1 & d_2 \end{array} \\ \begin{pmatrix} 23 & 0{,}5 \\ 35 & 1 \\ 18 & 0 \end{pmatrix} \end{array}
\qquad
\begin{array}{c} y_1 \\ y_2 \\ y_3 \end{array}
\begin{pmatrix} 55 \\ 100 \\ 45 \end{pmatrix}
$$

ПРИМЕР

- $Q'_{\beta_1} = \frac{1}{3} \cdot [23 \cdot 2 \cdot (\beta_1 \cdot 23 + \beta_2 \cdot 0.5 + \beta_0 - 55) + 35 \cdot 2 \cdot (\beta_1 \cdot 35 + \beta_2 \cdot 1 + \beta_0 - 100) + 18 \cdot 2 \cdot (\beta_1 \cdot 18 + \beta_0 - 45)]$
- $Q'_{\beta_2} = \frac{1}{3} \cdot [0.5 \cdot 2 \cdot (\beta_1 \cdot 23 + \beta_2 \cdot 0.5 + \beta_0 - 55) + 1 \cdot 2 \cdot (\beta_1 \cdot 35 + \beta_2 \cdot 1 + \beta_0 - 100)]$
- $Q'_{\beta_0} = \frac{1}{3} \cdot [1 \cdot 2 \cdot (\beta_1 \cdot 23 + \beta_2 \cdot 0.5 + \beta_0 - 55) + 1 \cdot 2 \cdot (\beta_1 \cdot 35 + \beta_2 \cdot 1 + \beta_0 - 100) + 1 \cdot 2 \cdot (\beta_1 \cdot 18 + \beta_0 - 45)]$

- Пусть инициализируемся в точке $\beta_{start} = \left(\beta_{start_1};\ \beta_{start_2};\ \beta_{start_0}\right) = (0;\ 0;\ 0)$
- $Q'_{\beta_1} = \frac{1}{3} \cdot [23 \cdot 2 \cdot (0 \cdot 23 + 0 \cdot 0.5 + 0 - 55) + 35 \cdot 2 \cdot (0 \cdot 35 + 0 \cdot 1 + 0 - 100) + 18 \cdot 2 \cdot (0 \cdot 18 + 0 - 45)] = -3716\frac{1}{3}$
- $Q'_{\beta_2} = \frac{1}{3} \cdot [0.5 \cdot 2 \cdot (0 \cdot 23 + 0 \cdot 0.5 + 0 - 55) + 1 \cdot 2 \cdot (0 \cdot 35 + 0 \cdot 1 + 0 - 100)] = -85$
- $Q'_{\beta_0} = \frac{1}{3} \cdot [1 \cdot 2 \cdot (0 \cdot 23 + 0 \cdot 0.5 + 0 - 55) + 1 \cdot 2 \cdot (0 \cdot 35 + 0 \cdot 1 + 0 - 100) + 1 \cdot 2 \cdot (0 \cdot 18 + 0 - 45)] = -133\frac{1}{3}$

ПРИМЕР

- Тогда $\nabla Q(\beta_{start}) = (\,-3716\frac{1}{3} \qquad -85 \qquad -133\frac{1}{3}\,)$
- $\beta_{next} = \beta_{start} - \eta \cdot \nabla Q(\beta_{start})$
- $\beta_{next} = (0;\ \ 0;\ \ 0) - 0{,}01 \cdot \left(-3716\frac{1}{3} \qquad -85 \qquad -133\frac{1}{3}\right) = (37\frac{1}{6};\ \ 0{,}85;\ \ 1\frac{1}{3})$
- Мы сделали градиентный шаг!
- И так далее, только уже теперь $\beta_{start} = (37\frac{1}{6};\ \ 0{,}85;\ \ 1\frac{1}{3})$
- $\beta_{next} = \beta_{start} - \eta \cdot \nabla Q(\beta_{start})$
- …
- Пока не сработает выбранный критерий останова!

РЕЗЮМЕ

— Узнали, что такое градиент функции

— Поняли, почему метод градиентного спуска так называется

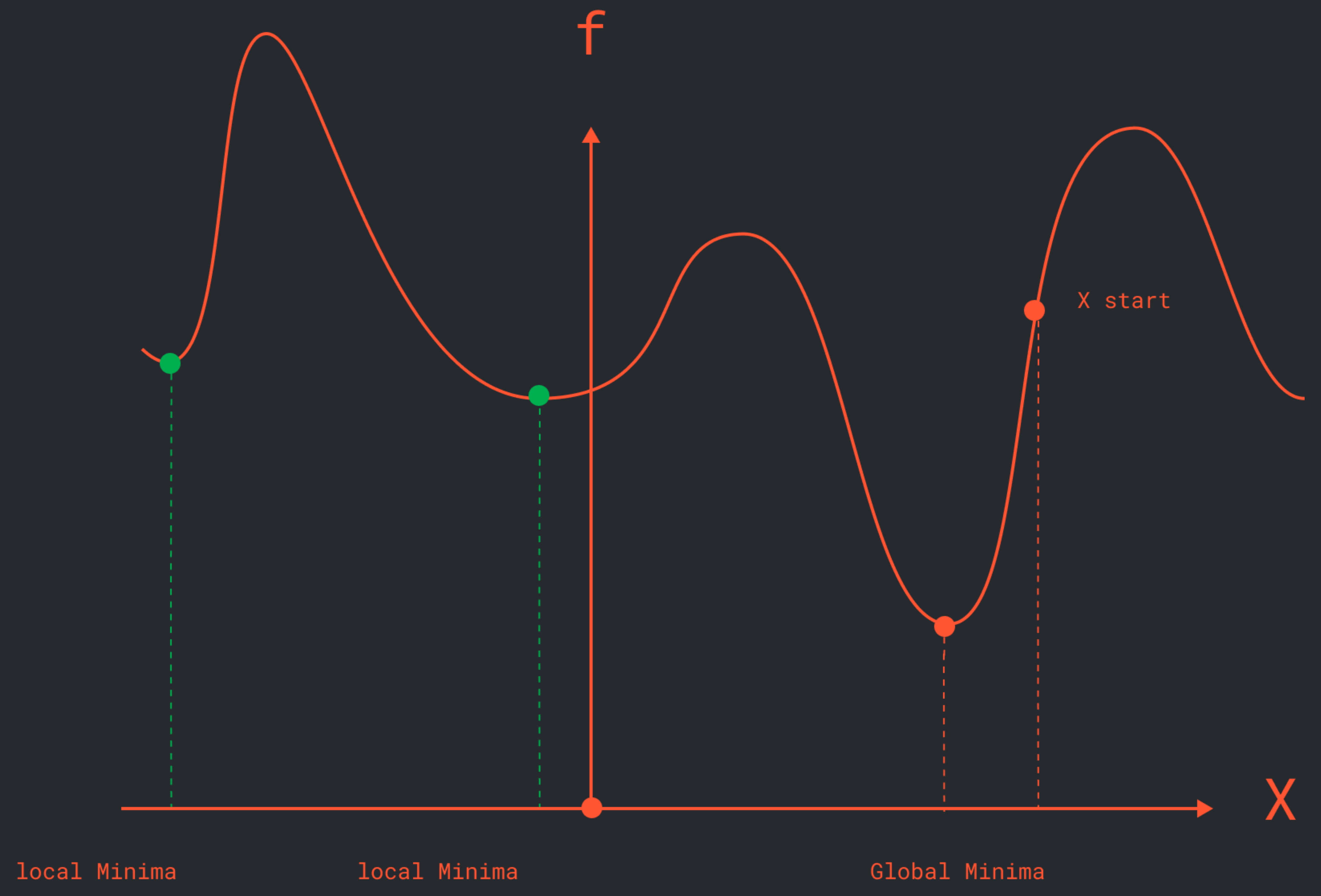
— Научились рисовать линии уровня и визуализировать спуск для ФМП

— Узнали, как выглядят старые критерии останова для многомерного случая

— Какие η и ξ выбирать для спуска? Какие могут быть с этим проблемы?

ГРАДИЕНТНЫЙ СПУСК: БОЛЬШАЯ ДЛИНА ШАГА

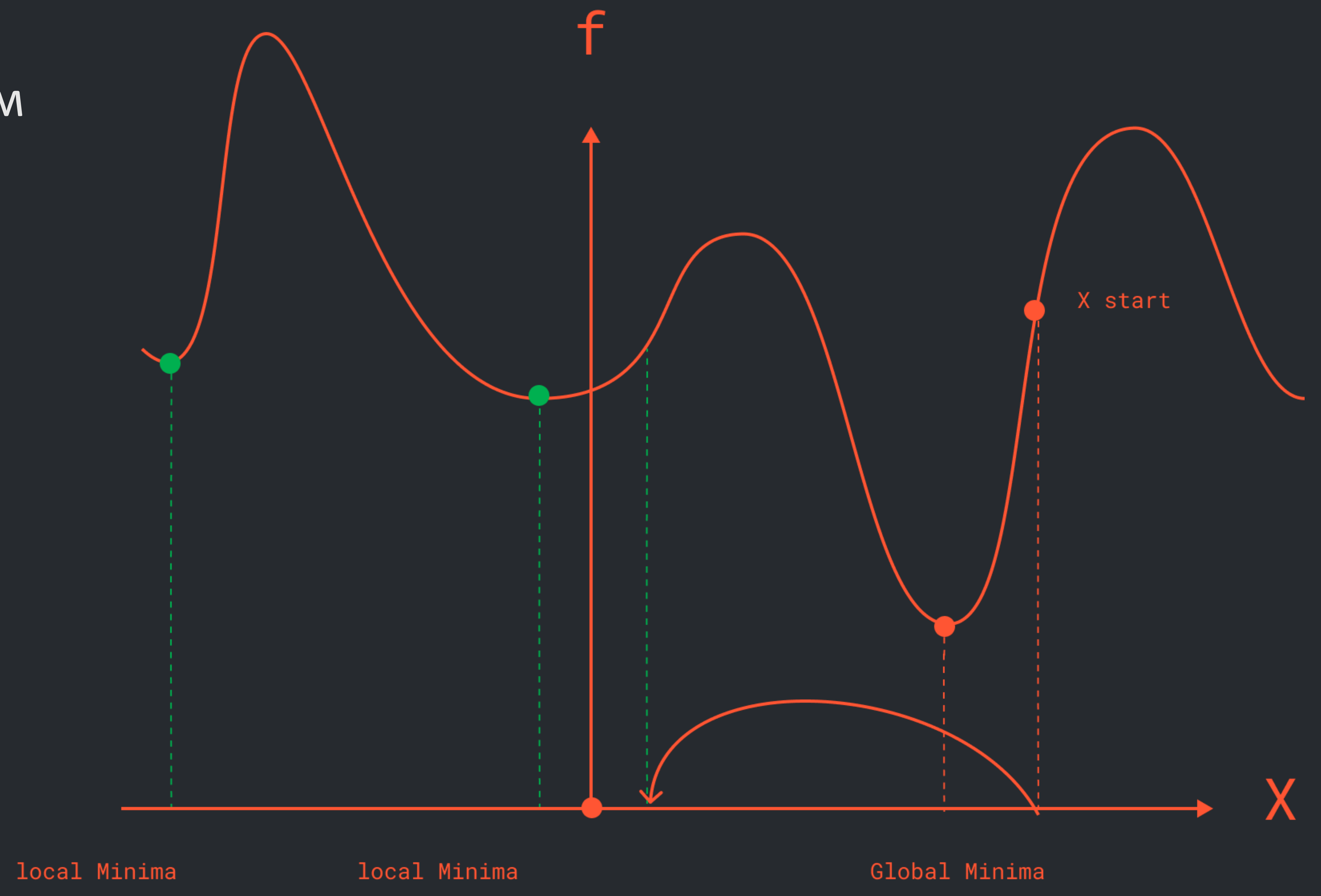
Функция одной переменной f(x)



ГРАДИЕНТНЫЙ СПУСК: БОЛЬШАЯ ДЛИНА ШАГА

Функция одной переменной f(x)

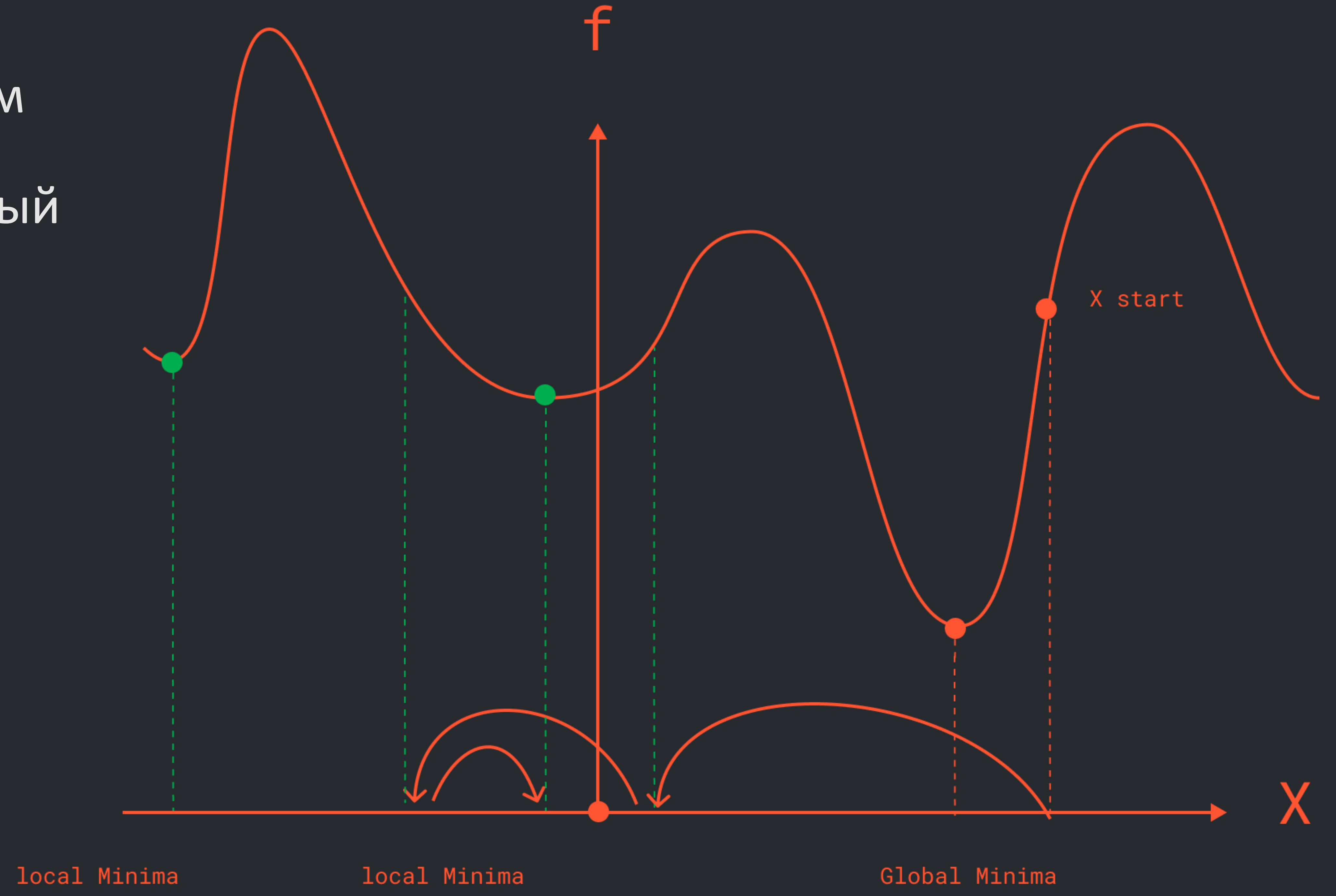
— Можем перескочить желанный минимум



ГРАДИЕНТНЫЙ СПУСК: БОЛЬШАЯ ДЛИНА ШАГА

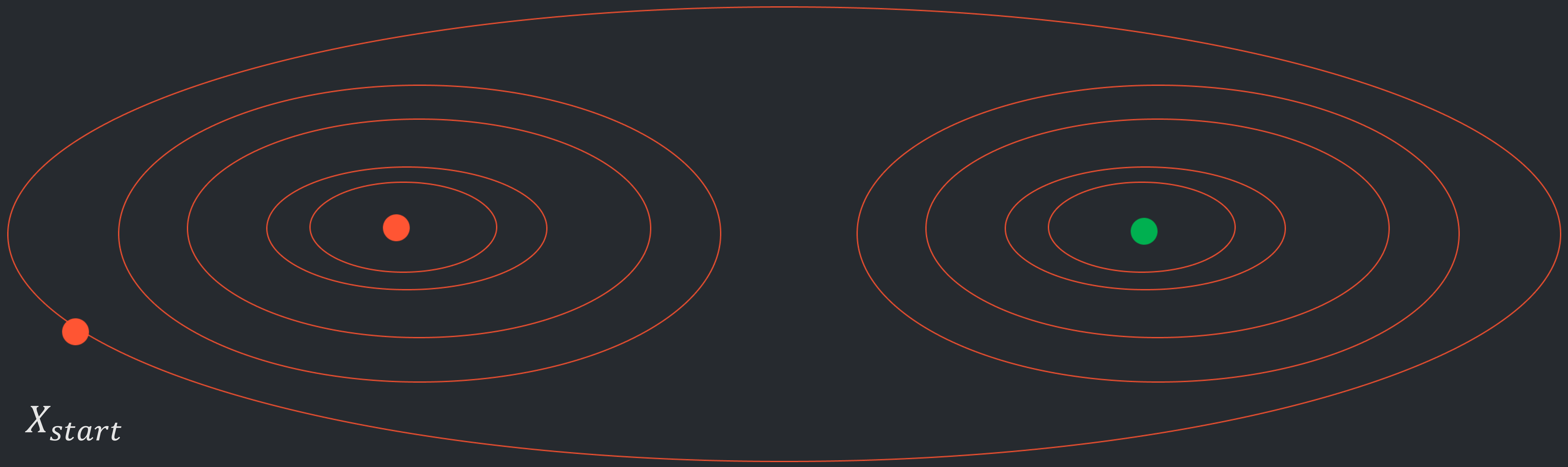
Функция одной переменной f(x)

- Можем перескочить желанный минимум
- И попасть в итоге более плохой локальный



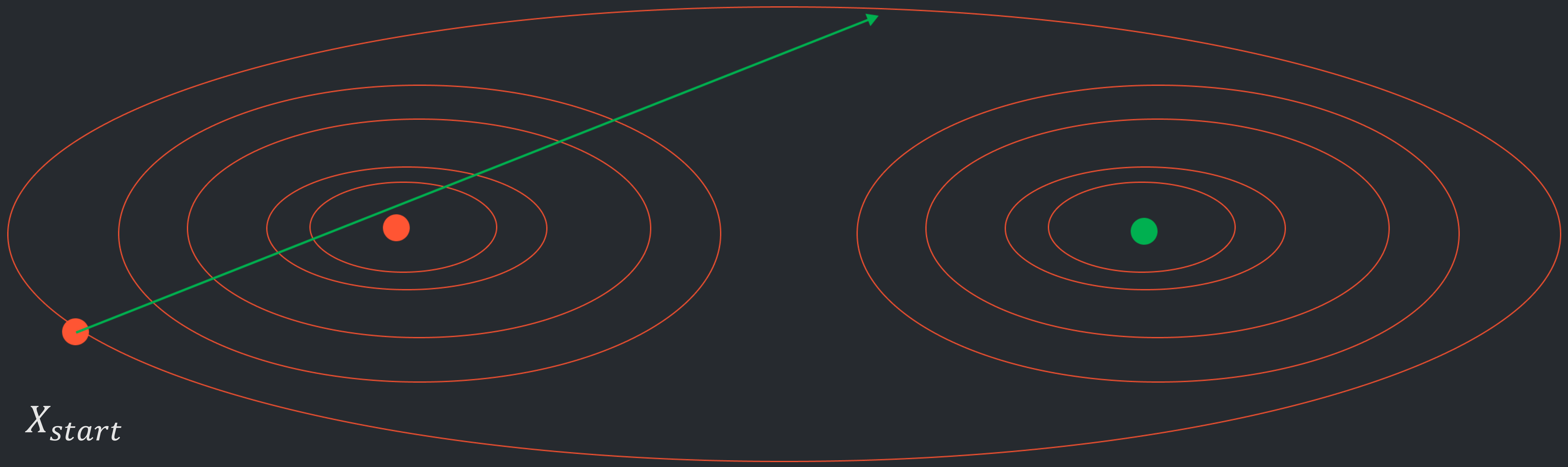
ГРАДИЕНТНЫЙ СПУСК: БОЛЬШАЯ ДЛИНА ШАГА

Функция нескольких переменных $z(x_1, x_2, \ldots, x_n)$



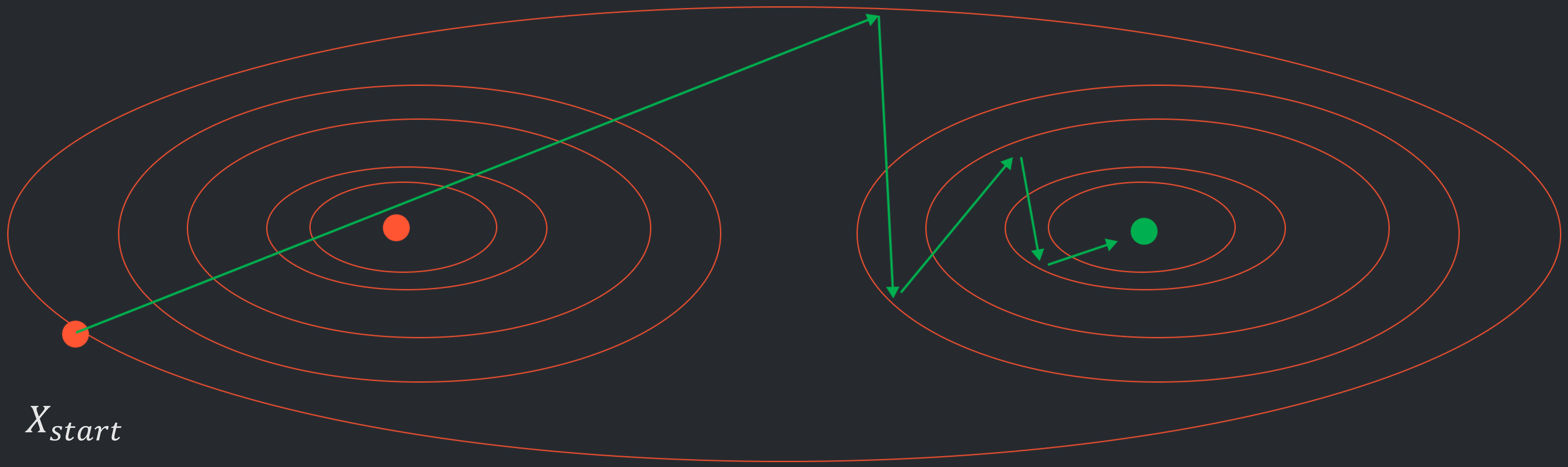
ГРАДИЕНТНЫЙ СПУСК: БОЛЬШАЯ ДЛИНА ШАГА

Функция нескольких переменных $z(x_1, x_2, \dots, x_n)$



ГРАДИЕНТНЫЙ СПУСК: БОЛЬШАЯ ДЛИНА ШАГА

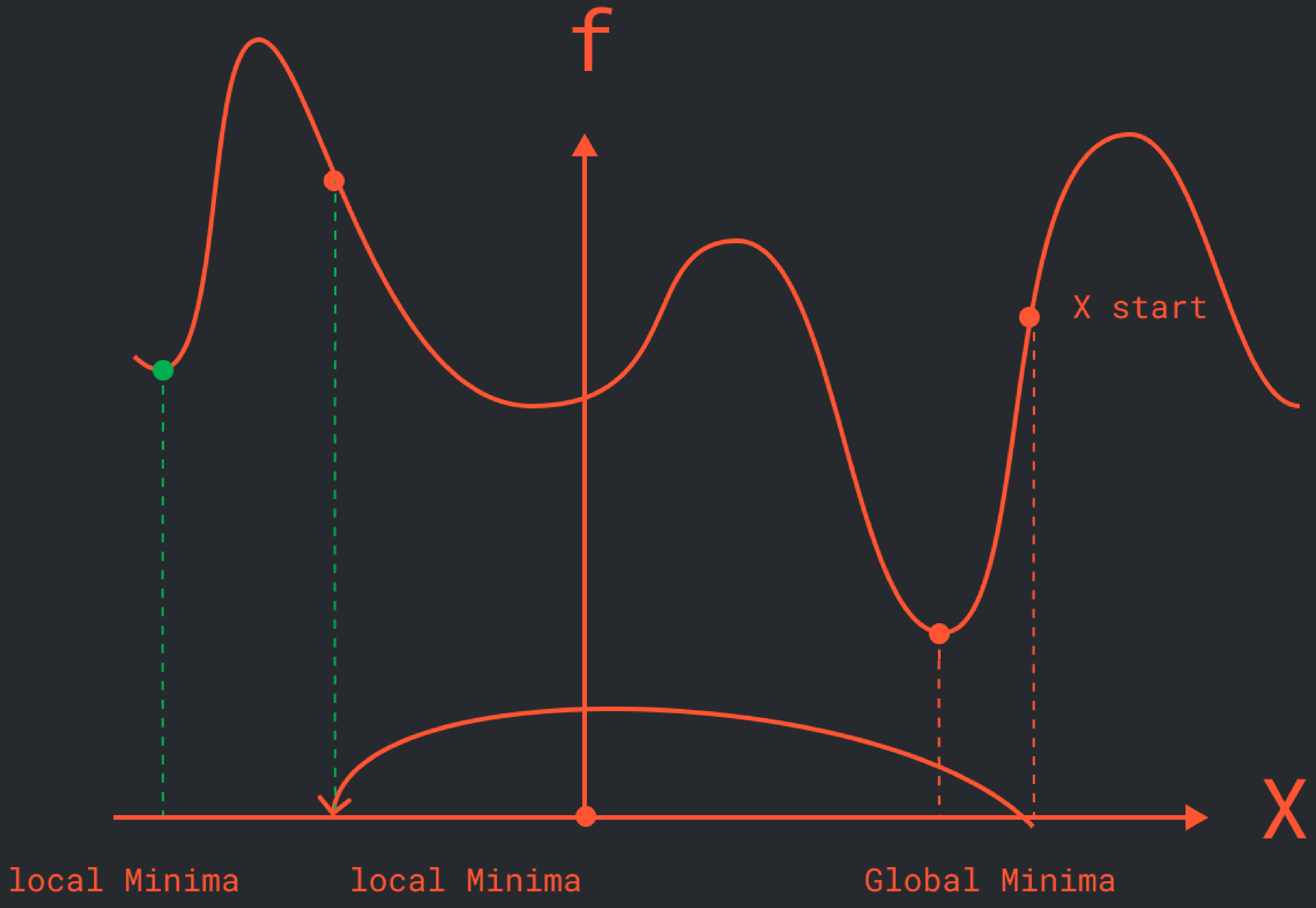
Функция нескольких переменных $z(x_1, x_2, \dots, x_n)$



ГРАДИЕНТНЫЙ СПУСК: ВЗРЫВ ГРАДИЕНТА

Функция одной переменной f(x)

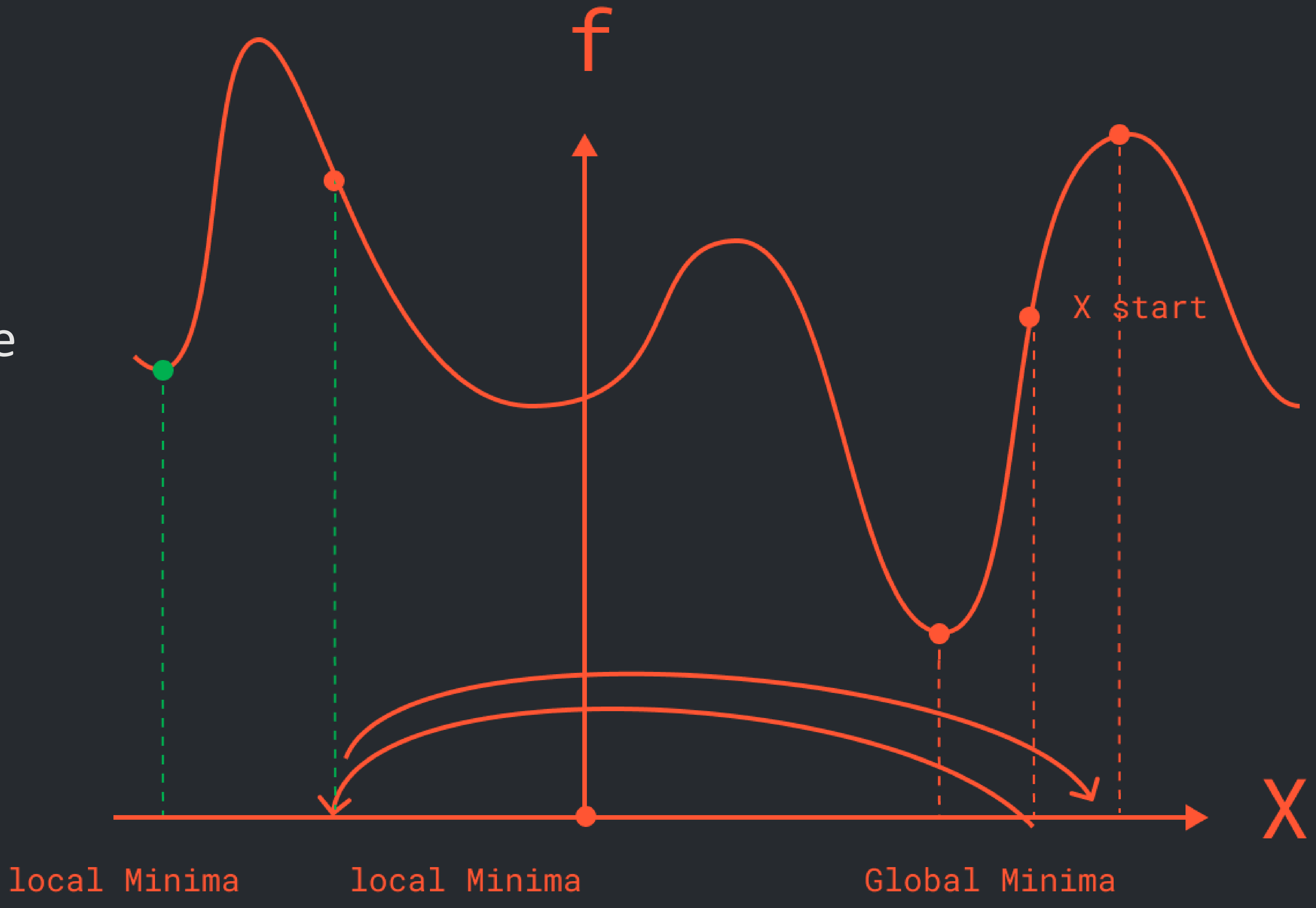
- Градиент может взорваться
- То есть начать перепрыгивать минимумы



ГРАДИЕНТНЫЙ СПУСК: ВЗРЫВ ГРАДИЕНТА

Функция одной переменной f(x)

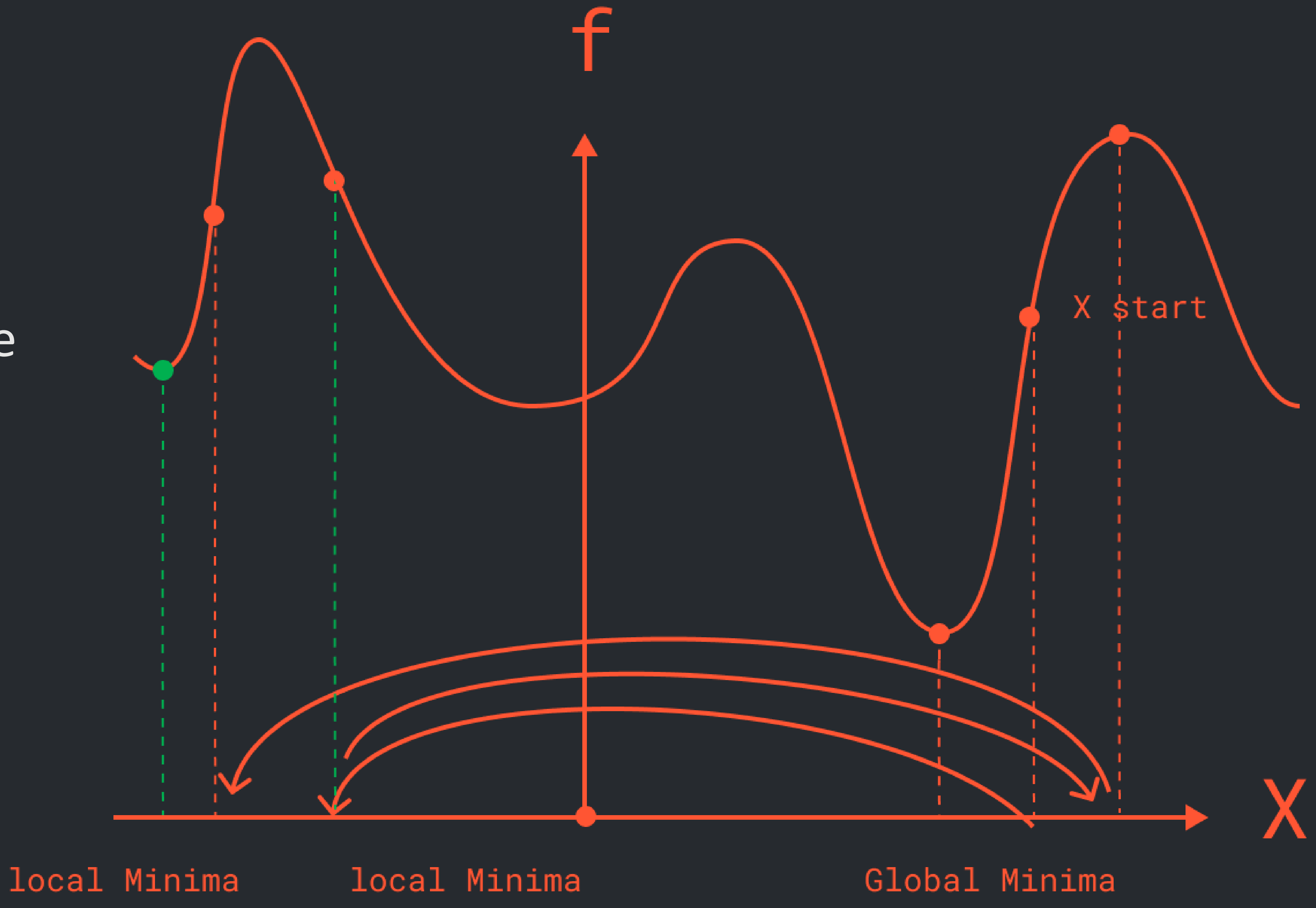
- Градиент может взорваться
- То есть начать перепрыгивать минимумы
- С каждой итерацией все дальше и дальше
- Будем бесконечно блуждать!



ГРАДИЕНТНЫЙ СПУСК: ВЗРЫВ ГРАДИЕНТА

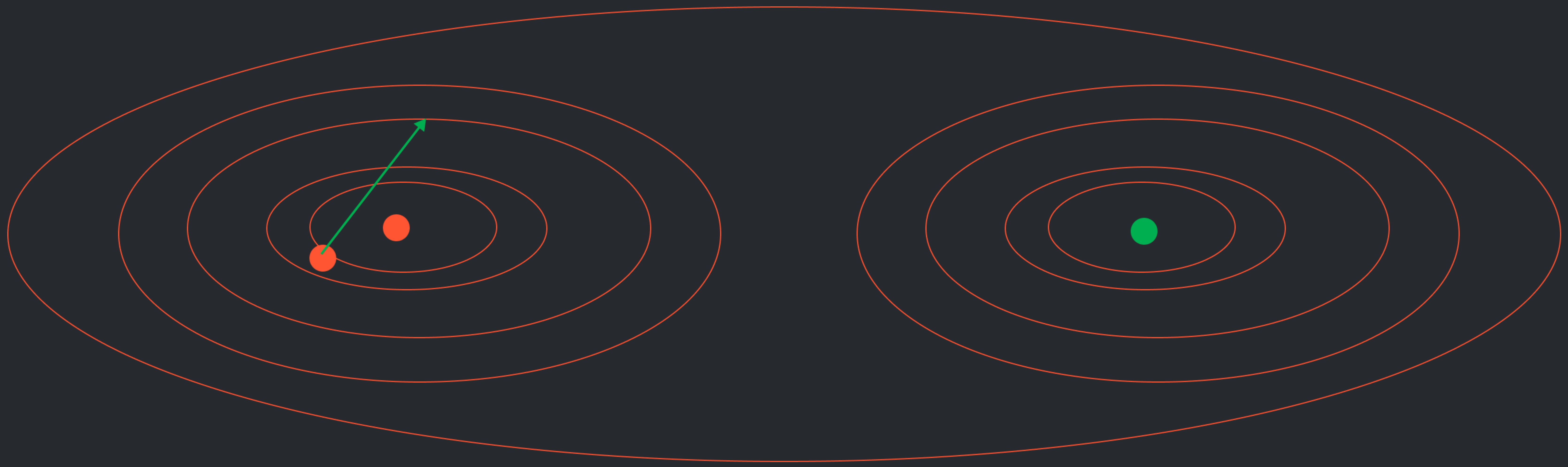
Функция одной переменной f(x)

- Градиент может взорваться
- То есть начать перепрыгивать минимумы
- С каждой итерацией все дальше и дальше
- Будем бесконечно блуждать!



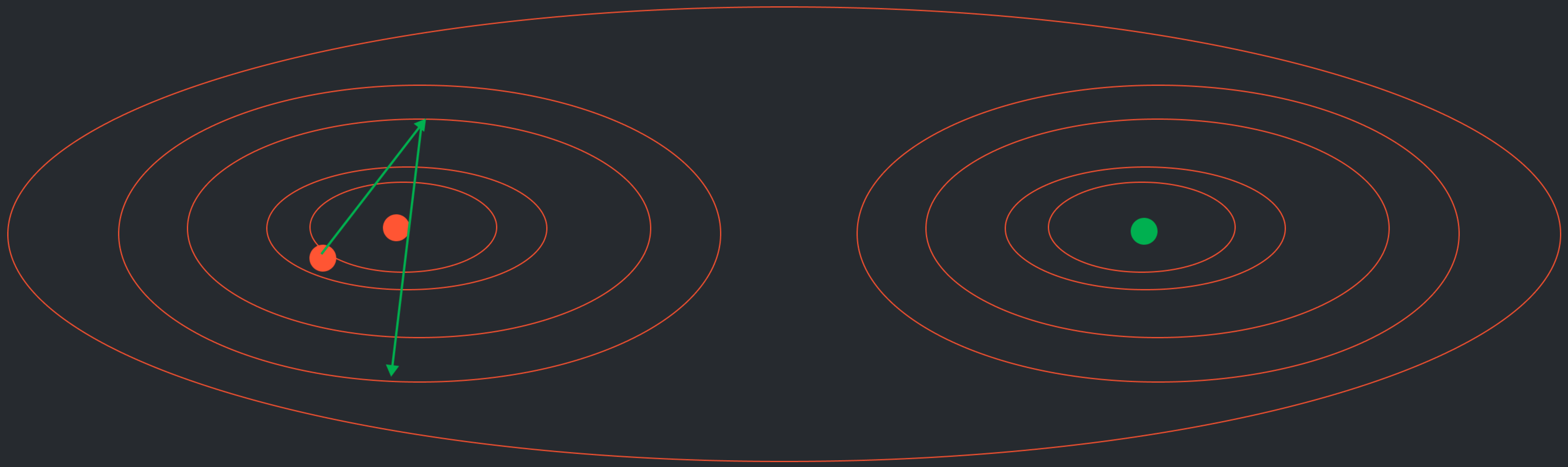
ГРАДИЕНТНЫЙ СПУСК: ВЗРЫВ ГРАДИЕНТА

Функция нескольких переменных $z(x_1, x_2, \dots, x_n)$



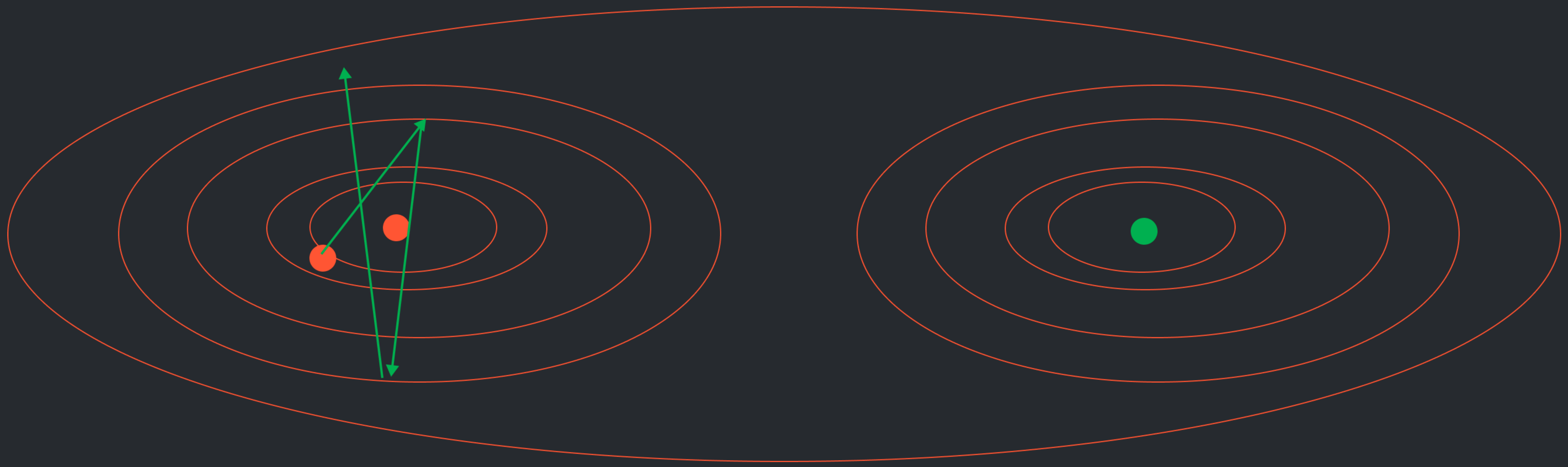
ГРАДИЕНТНЫЙ СПУСК: ВЗРЫВ ГРАДИЕНТА

Функция нескольких переменных $z(x_1, x_2, \dots, x_n)$



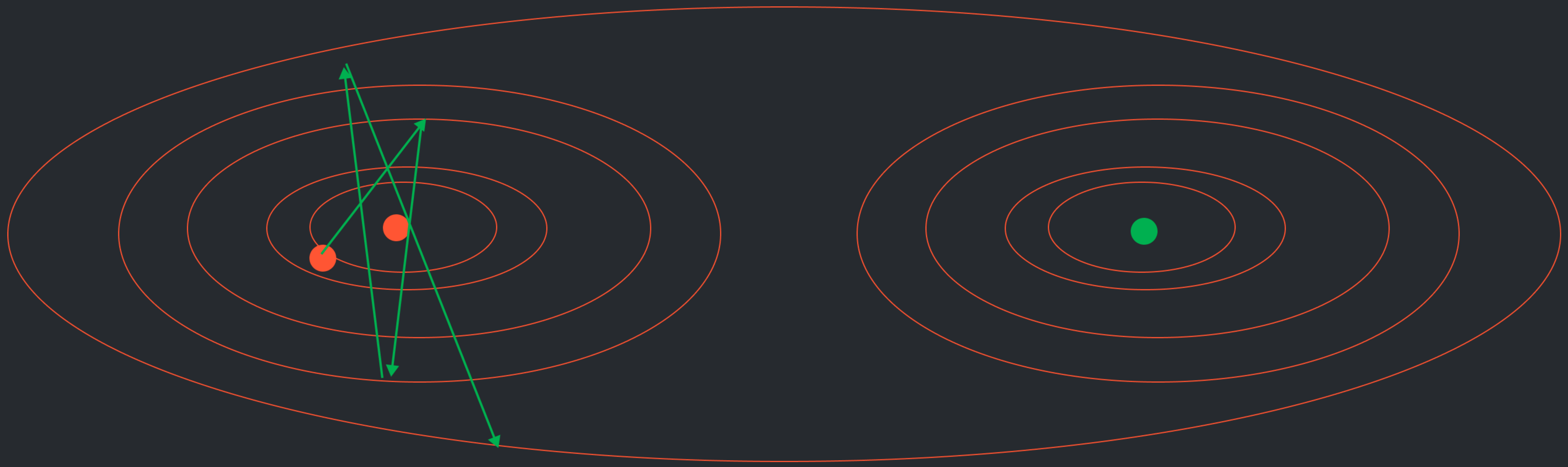
ГРАДИЕНТНЫЙ СПУСК: ВЗРЫВ ГРАДИЕНТА

Функция нескольких переменных $z(x_1, x_2, \dots, x_n)$



ГРАДИЕНТНЫЙ СПУСК: ВЗРЫВ ГРАДИЕНТА

Функция нескольких переменных $z(x_1, x_2, \dots, x_n)$



ГРАДИЕНТНЫЙ СПУСК: МАЛЕНЬКАЯ ДЛИНА ШАГА

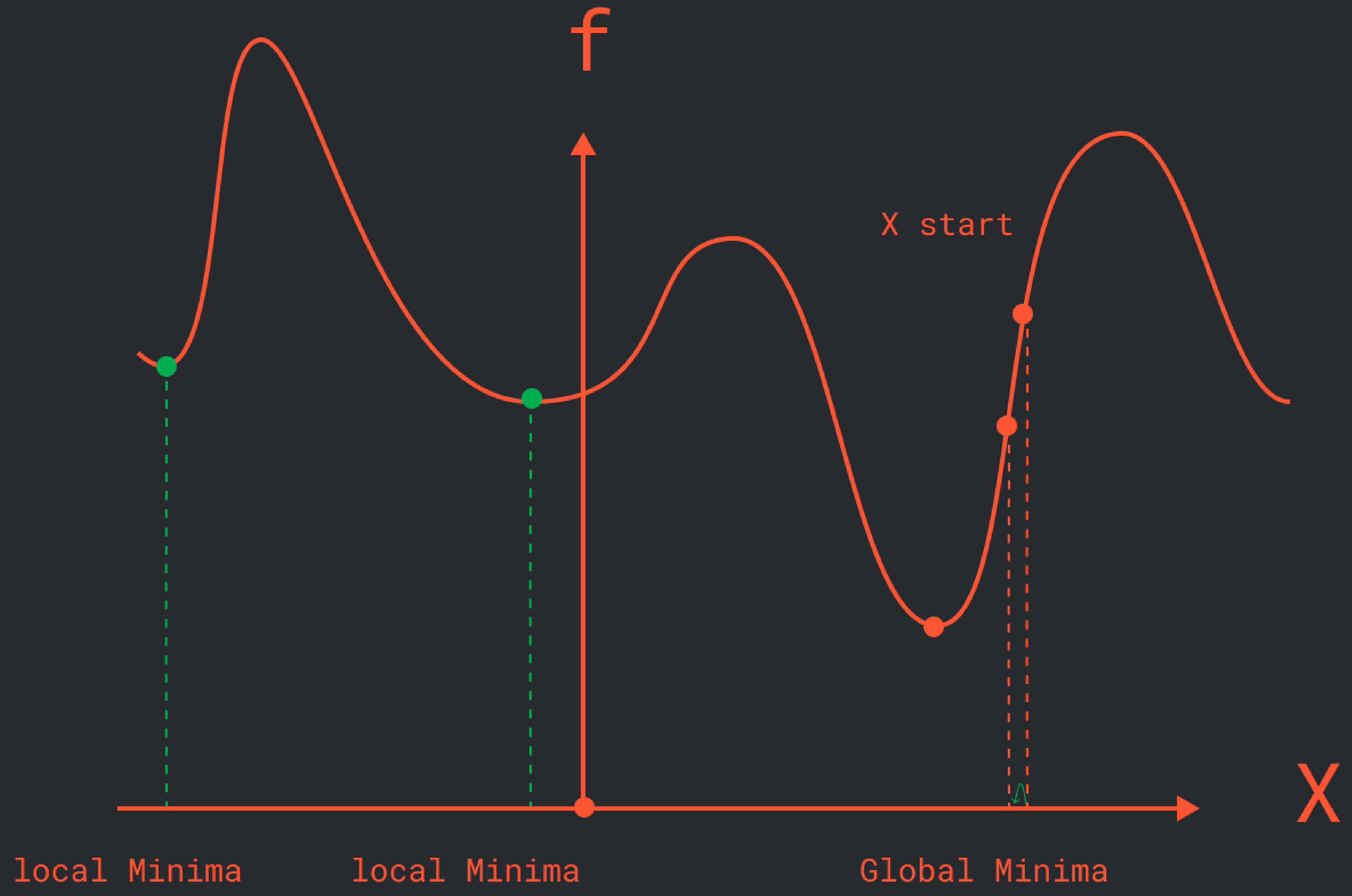
Функция одной переменной f(x)

— Можем не добежать до минимума

— Рано сработает критерия останова

— Затухание градиента

— Или добежать, но
за очень больше количество операций



ГРАДИЕНТНЫЙ СПУСК: МАЛЕНЬКАЯ ДЛИНА ШАГА

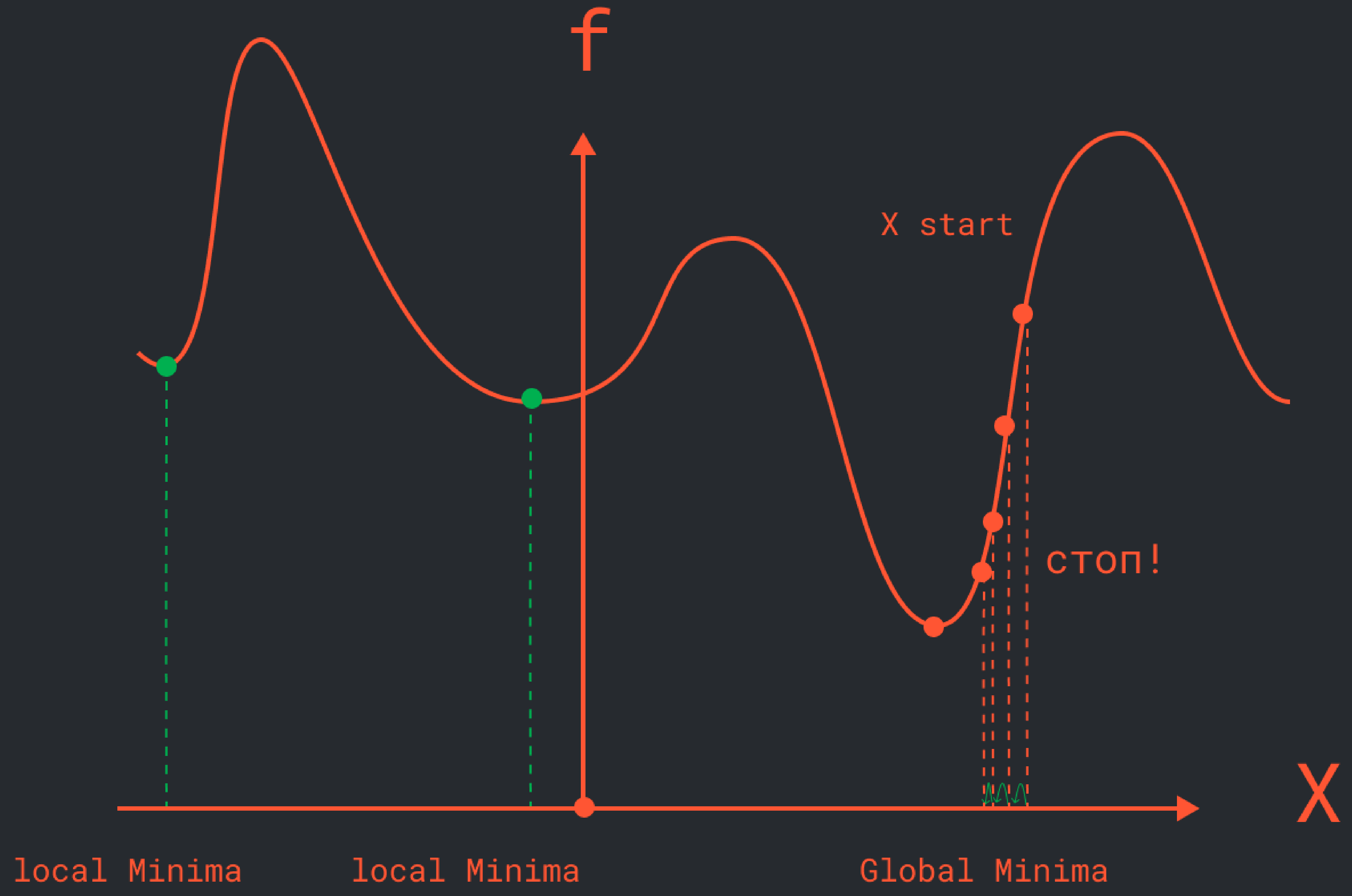
Функция одной переменной f(x)

— Можем не добежать до минимума

— Рано сработает критерия останова

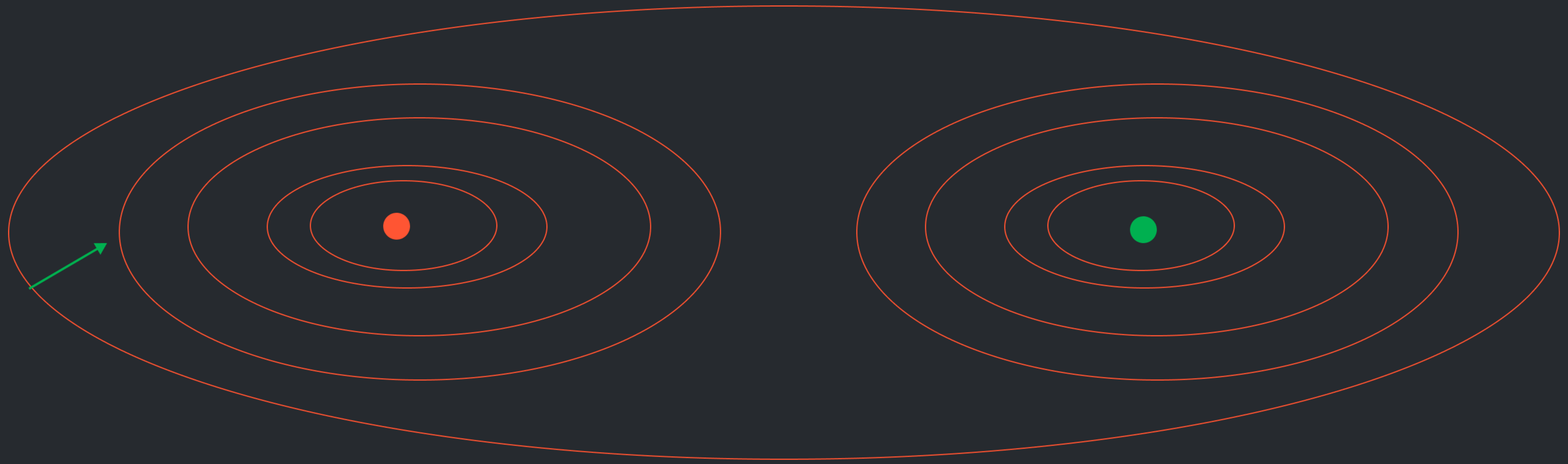
— Затухание градиента

— Или добежать, но
за очень больше количество операций



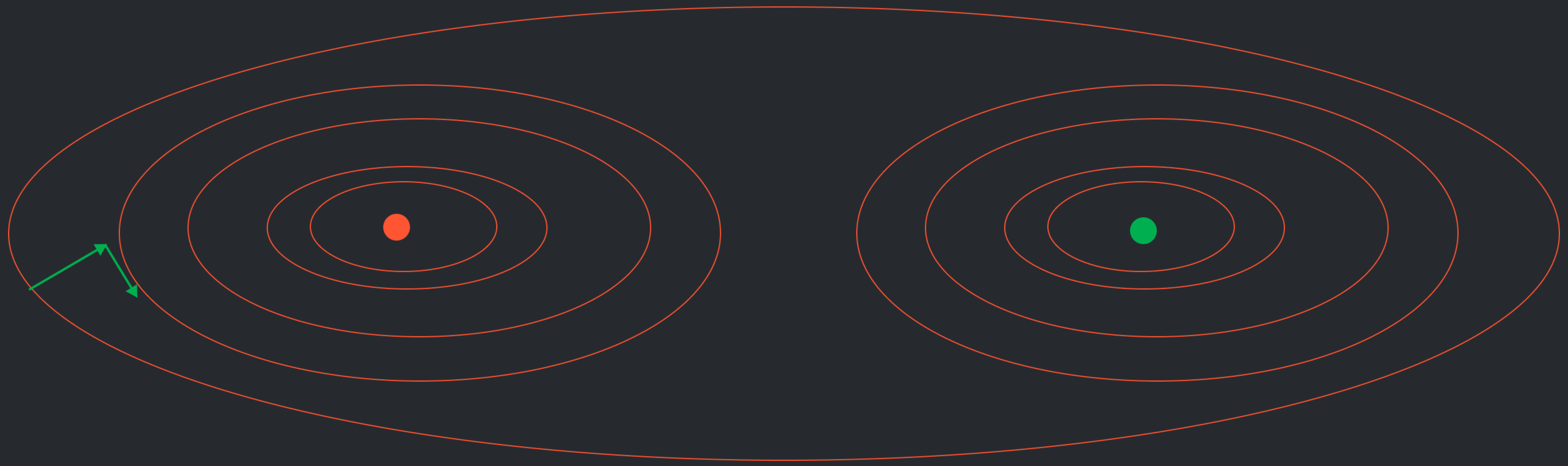
ГРАДИЕНТНЫЙ СПУСК: МАЛЕНЬКАЯ ДЛИНА ШАГА

Функция нескольких переменных $z(x_1, x_2, \dots, x_n)$



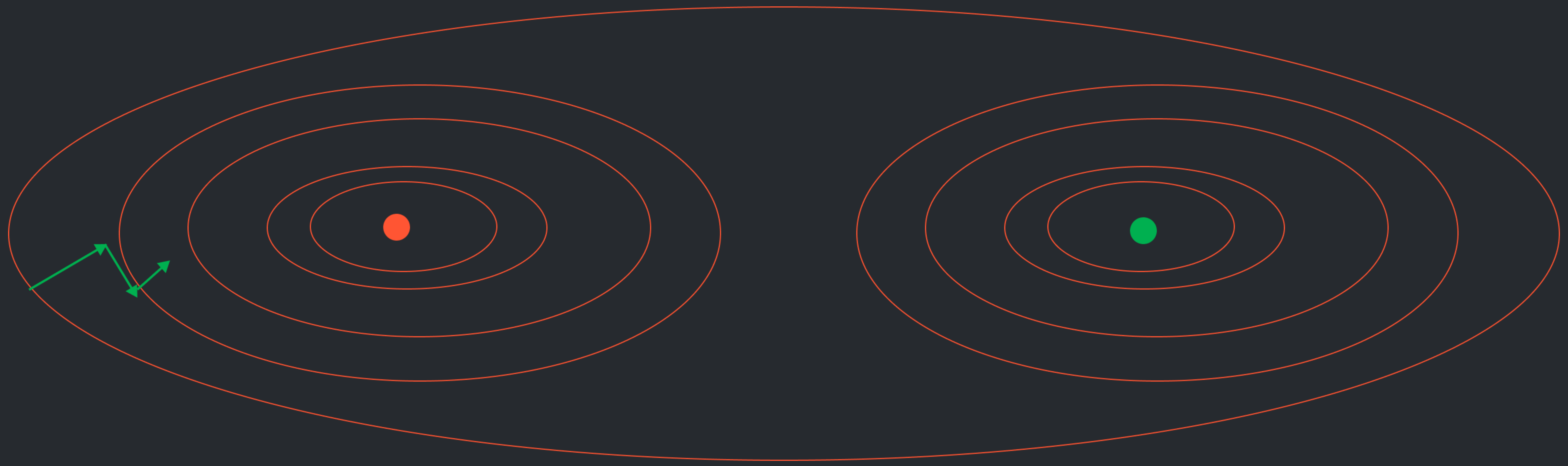
ГРАДИЕНТНЫЙ СПУСК: МАЛЕНЬКАЯ ДЛИНА ШАГА

Функция нескольких переменных $z(x_1, x_2, \dots, x_n)$



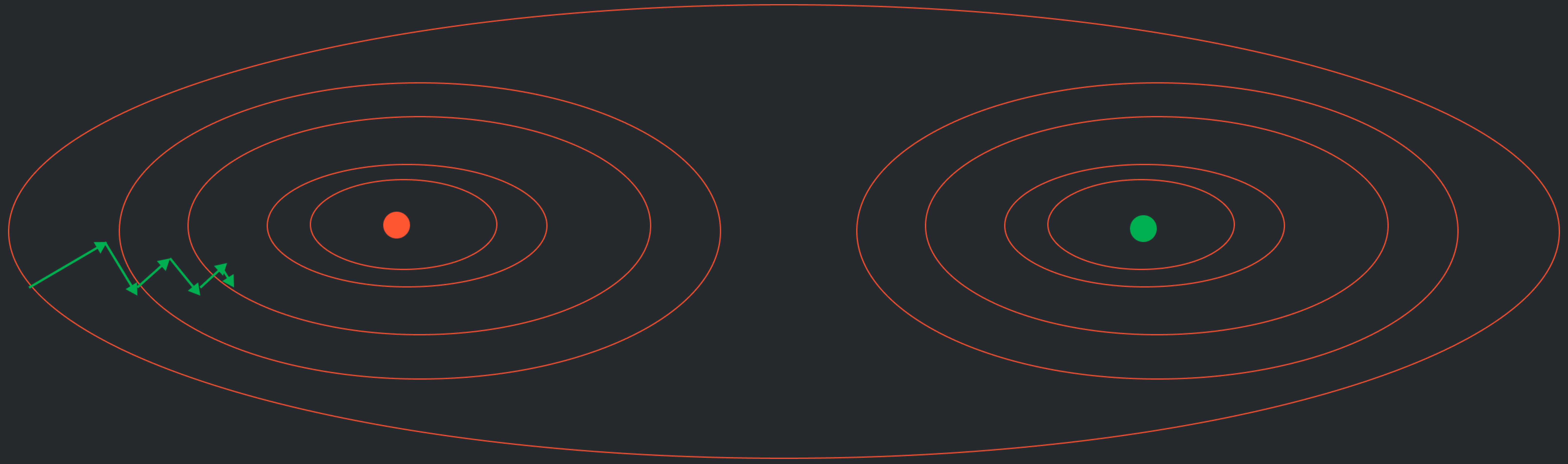
ГРАДИЕНТНЫЙ СПУСК: МАЛЕНЬКАЯ ДЛИНА ШАГА

Функция нескольких переменных $z(x_1, x_2, \ldots, x_n)$



ГРАДИЕНТНЫЙ СПУСК: МАЛЕНЬКАЯ ДЛИНА ШАГА

Функция нескольких переменных $z(x_1, x_2, \dots, x_n)$



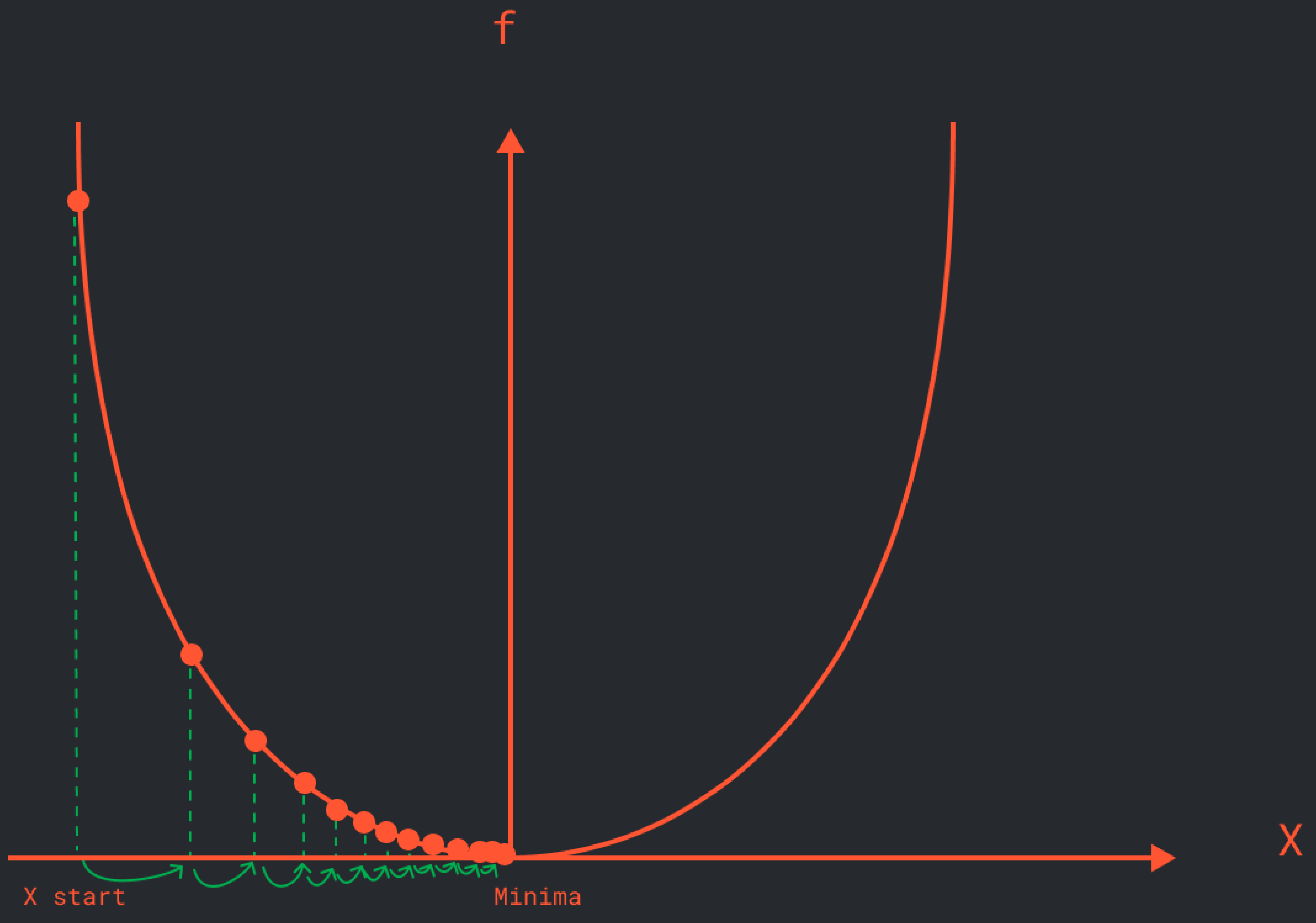
РЕЗЮМЕ

- Узнали, какие проблемы могут быть при плохом выборе обучающего шага
- Это затухание и взрыв градиента
- К тому же, даже если спуститься в адекватное значение удастся,
- Плохо выбранный шаг = долгое обучение
- Оптимальный параметр η заранее неизвестен
- Поэтому, если у функции очевидно не 1 минимум или много, то
- Нужны эксперименты!

ГРАДИЕНТНЫЙ СПУСК

НА ЧТО ВЛИЯЕТ ВЫБОР THRESHOLD?

— Если он достаточно консервативный, то итераций может понадобиться много, зато мы будем максимально близки к минимуму!



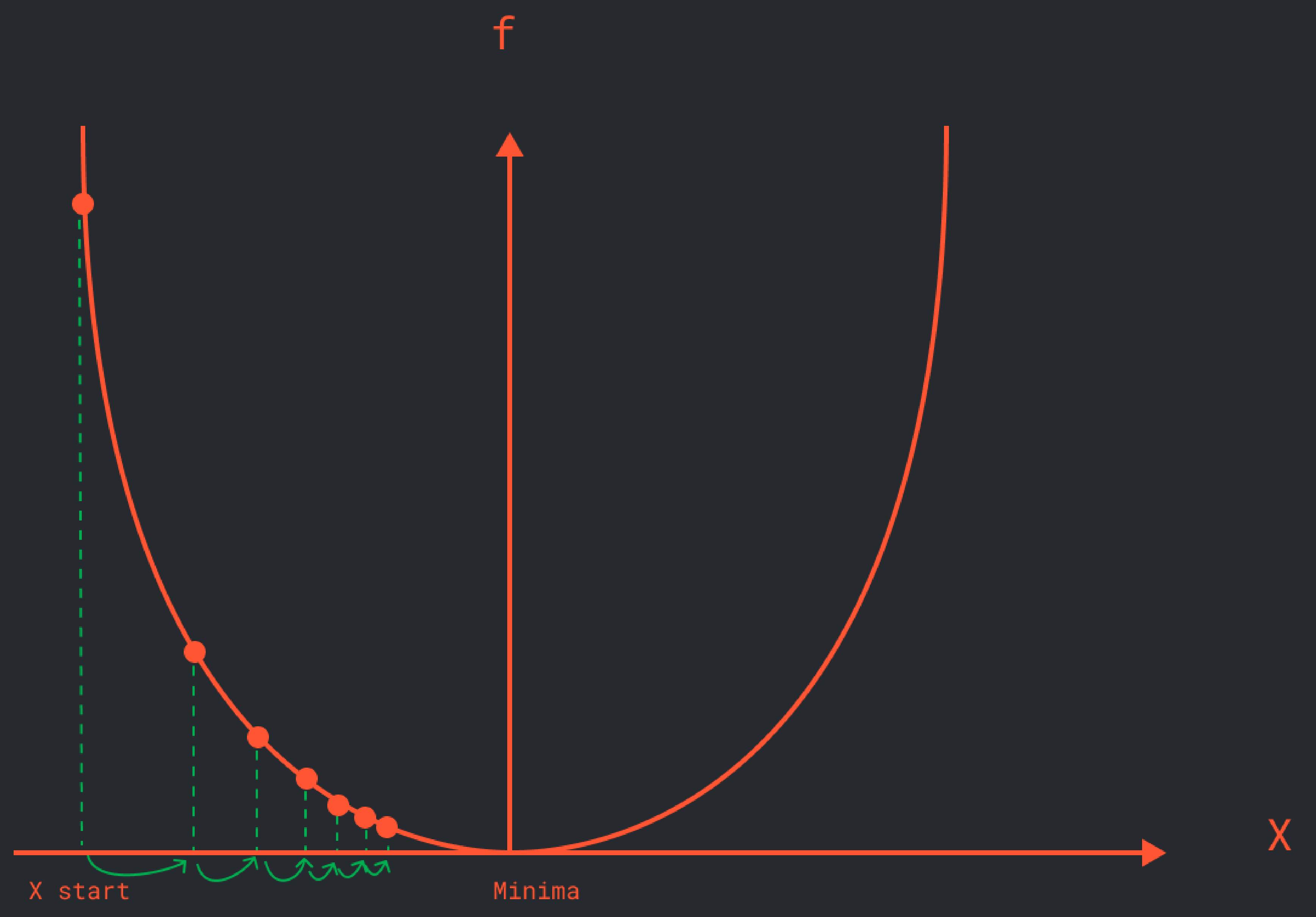
Совсем маленькие изменения, стоп нескоро

ГРАДИЕНТНЫЙ СПУСК

НА ЧТО ВЛИЯЕТ ВЫБОР THRESHOLD?

— Если он достаточно консервативный, то итераций может понадобиться много, зато мы будем максимально близки к минимуму!

— Если он достаточно большой, то итераций понадобится мало, но можем получить далеко не лучший результат

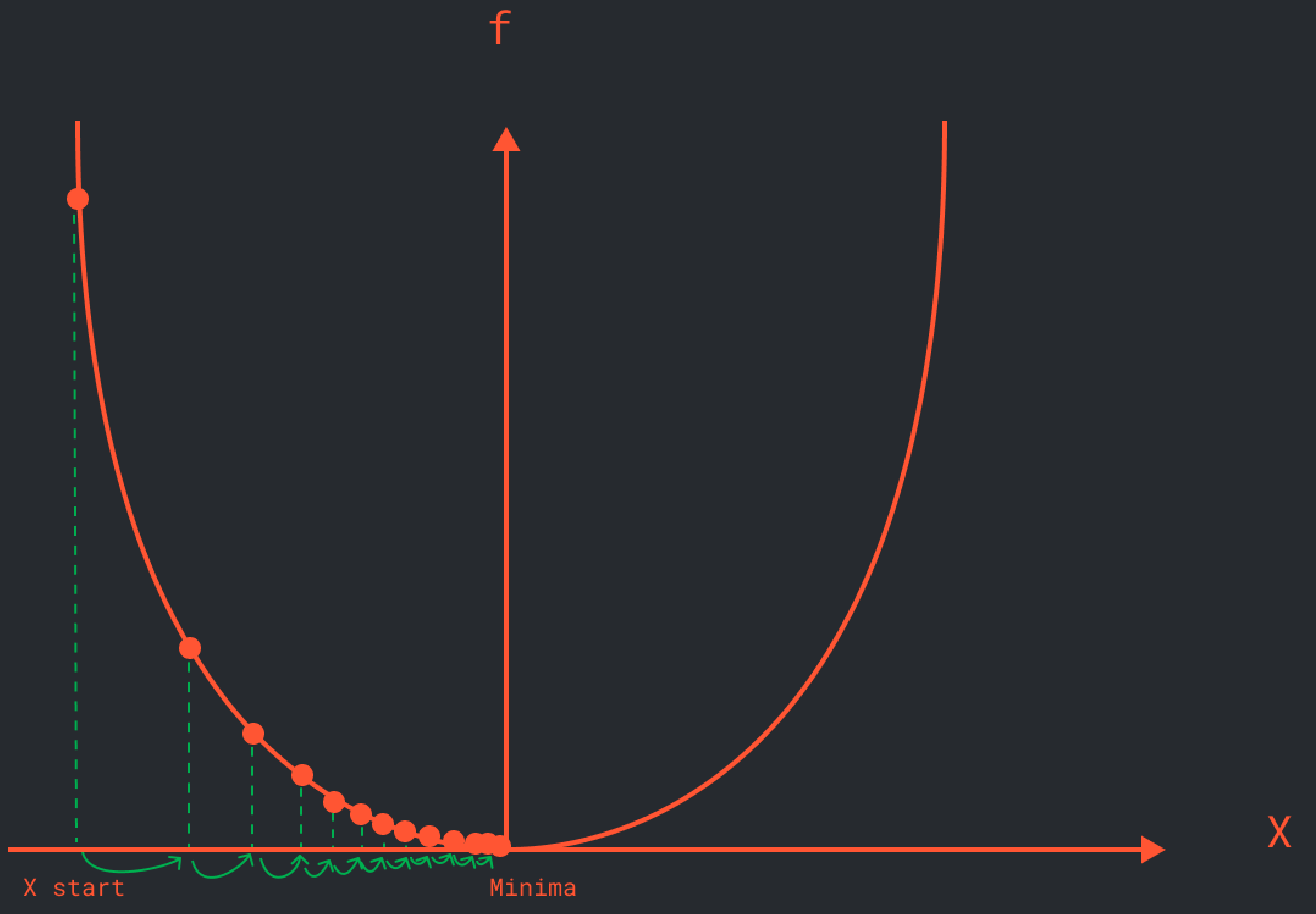


Изменения не маленькие, но рано стоп

ГРАДИЕНТНЫЙ СПУСК

НА ЧТО ВЛИЯЕТ ВЫБОР THRESHOLD?

- Если он достаточно консервативный, то итераций может понадобиться много, зато мы будем максимально близки к минимуму!
- Если он достаточно большой, то итераций понадобится мало, но можем получить далеко не лучший результат
- Нужно опытным путем находить золотую середину или использовать более продвинутые методы спуска!



РЕЗЮМЕ

- Узнали, что критерий (порог) останова тоже стоит аккуратно выбирать!
- Сильно большой порог ведет к раннему стопу и мы можем сильно “недоспуститься”
- Сильно маленький порог почти гарантирует большое количество операций
- Тоже экспериментируем с выбором ξ!
- Обычно начинаем с большого и спускаемся, смотря на то, какие значения получаются

РЕЗЮМЕ

- Неужели все?
- Кажется, что в методе градиентного спуска есть еще один параметр!
- А именно – точка старта!

ГРАДИЕНТНЫЙ СПУСК

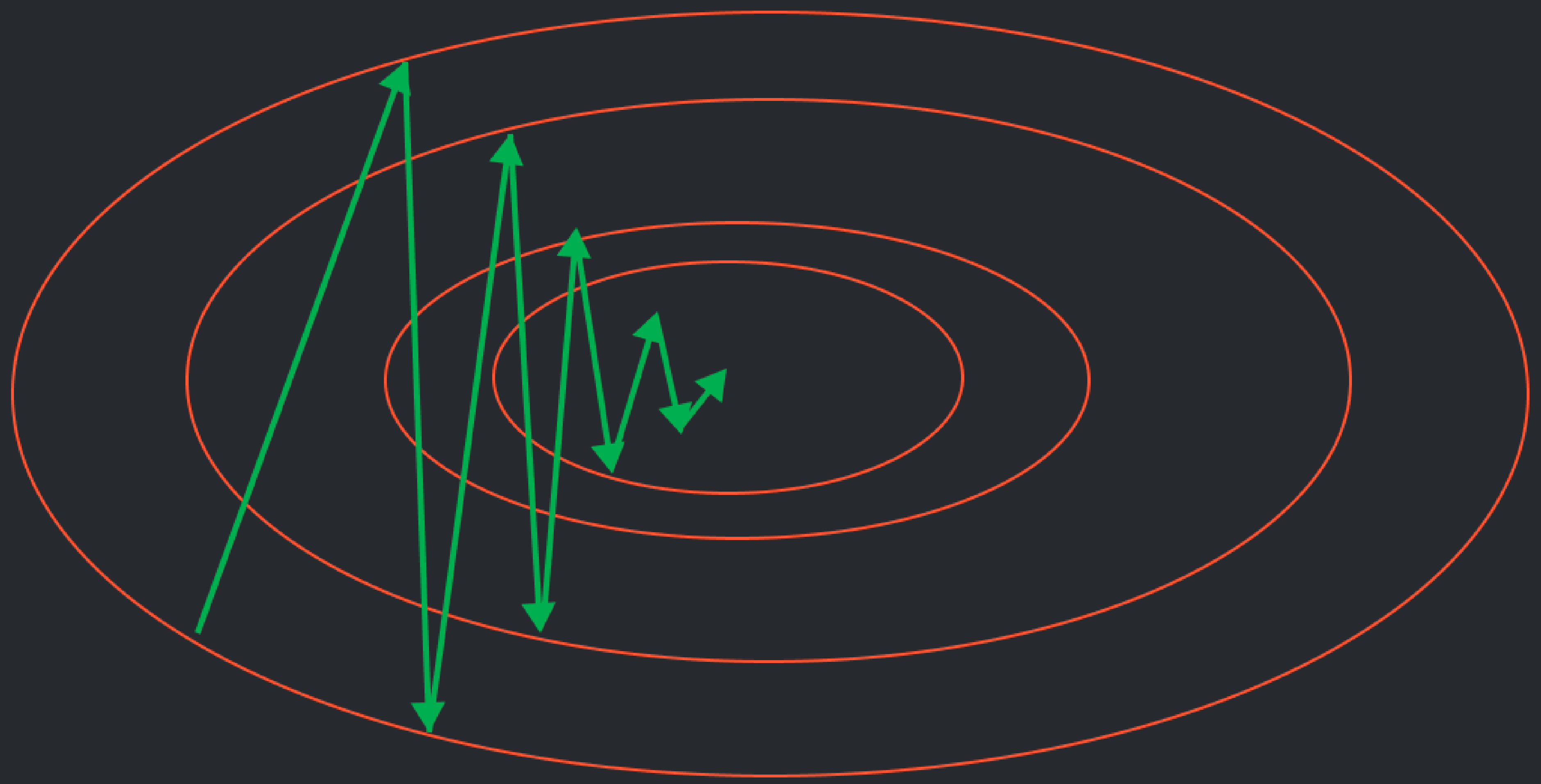
— Инициализируемся в случайной точке X_{start}

— До сходимости:

$$\text{step} = \nabla z'(X_{start})$$
$$X_{next} = X_{start} - \eta_i \cdot step$$
$$X_{start} = X_{next}$$

— Три варианта порога (threshold):

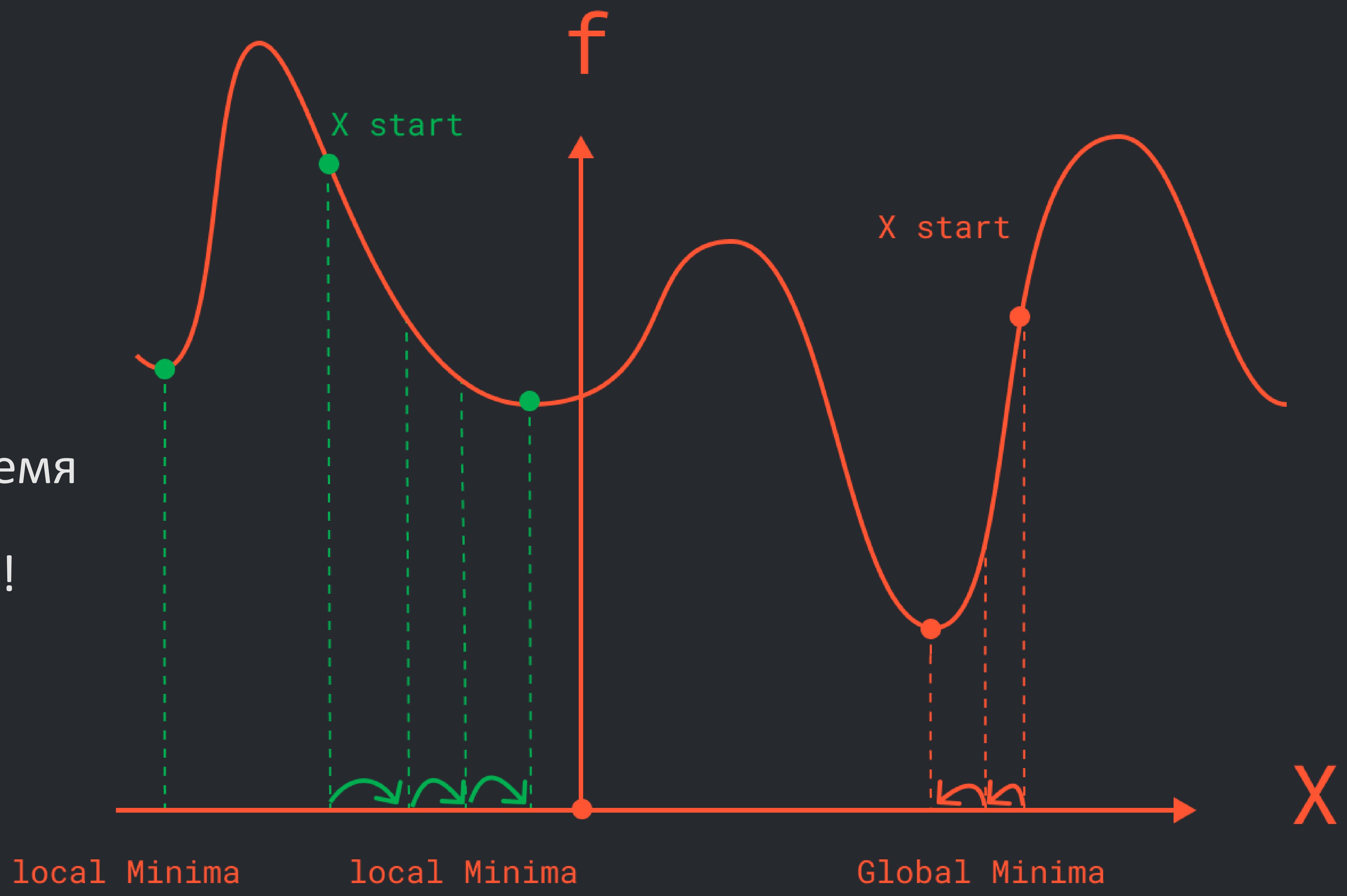
$$||\nabla z(\mathrm{X_{start}})|| \leq \xi$$
$$|f(X_{next}) - f(X_{start})| \leq \xi$$
$$|X_{start} - X_{next}| \leq \xi$$



ГРАДИЕНТНЫЙ СПУСК: ИНИЦИАЛИЗАЦИЯ

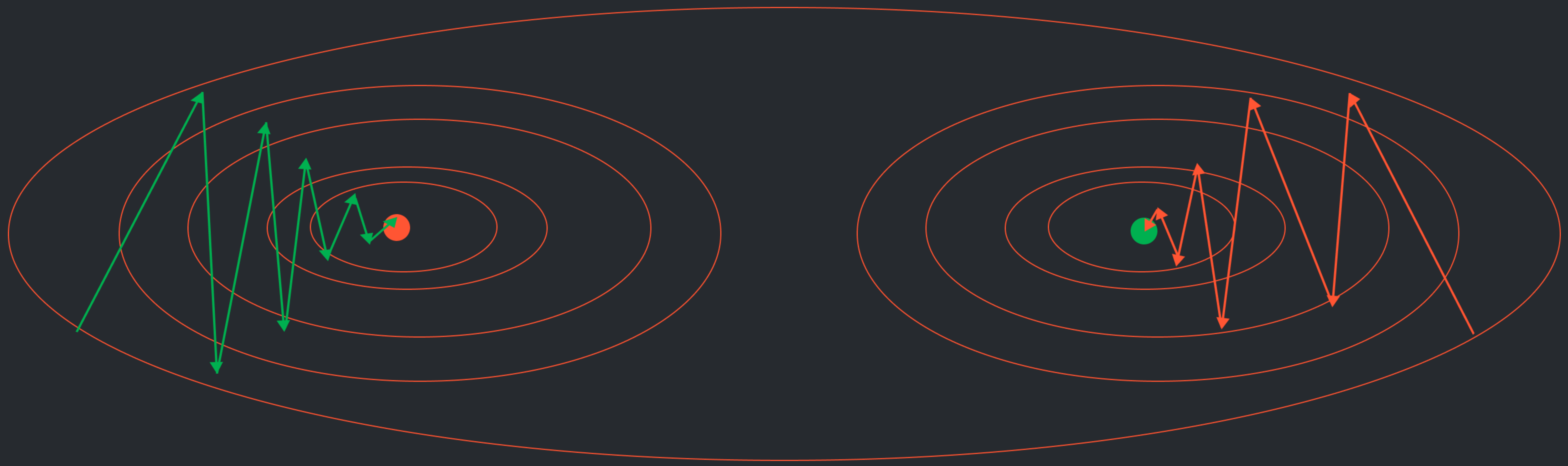
Функция одной переменной f(x)

— Пусть мы угадали с идеальными
— Мы не перескакиваем минимумы
— Градиент не взрывается
— Градиент не затухает
— Критерий останова работает вовремя
— А вот инициализировались не там!



ГРАДИЕНТНЫЙ СПУСК: ИНИЦИАЛИЗАЦИЯ

Функция нескольких переменных $z(x_1, x_2, \dots, x_n)$



РЕЗЮМЕ: ГРАДИЕНТНЫЙ СПУСК

ПРЕИМУЩЕСТВА

— Универсальный метод

— Можно вычислять параллельно

— Легко переделать для задачи поиска максимума функции

НЕДОСТАТКИ

— Нужна аккуратная настройка

— Непонятно, когда остановиться

СПАСИБО

ТАБАКАЕВ НИКИТА