Машинное обучение

Лекция 5. Метод опорных векторов

Сегодня

- Напоминание: линейная классификация
- Метод опорных векторов
- Ядра (Kernel trick) в методе опорных векторов
- Математическое дополнение: условный экстремум

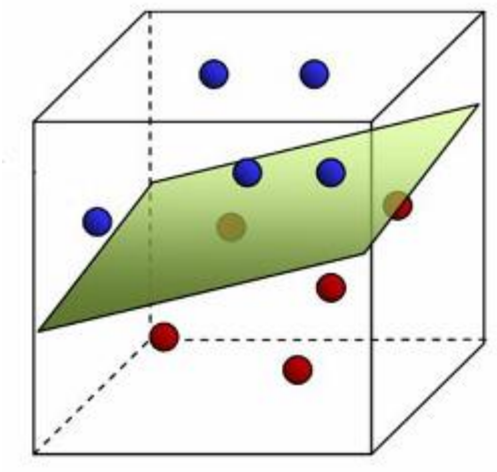
НАПОМИНАНИЕ: ЛИНЕЙНАЯ КЛАССИФИКАЦИЯ

Формализуем линейный классификатор

$$a(x) = \begin{cases} 1, \text{если } f(x) > 0 \\ -1, \text{если } f(x) \le 0 \end{cases}$$

$$f(x) = w_0 + w_1 x_1 + \cdots + w_d x_d = w_0 + \langle w, x \rangle$$

Геометрическая интерпретация: разделяем классы плоскостью



Формализуем линейный классификатор

$$a(x) = \begin{cases} 1, \text{если } f(x) > 0 \\ -1, \text{если } f(x) \leq 0 \end{cases}$$

Если добавляем $x_{(0)} = 1$, то:

$$\cancel{f(x) = w_0 + \langle w, x \rangle}$$

$$f(x) = \langle w, x \rangle = w^T x$$

Отступ (margin)

Отступом алгоритма $a(x) = sign\{f(x)\}$ на объекте x_i называется величина

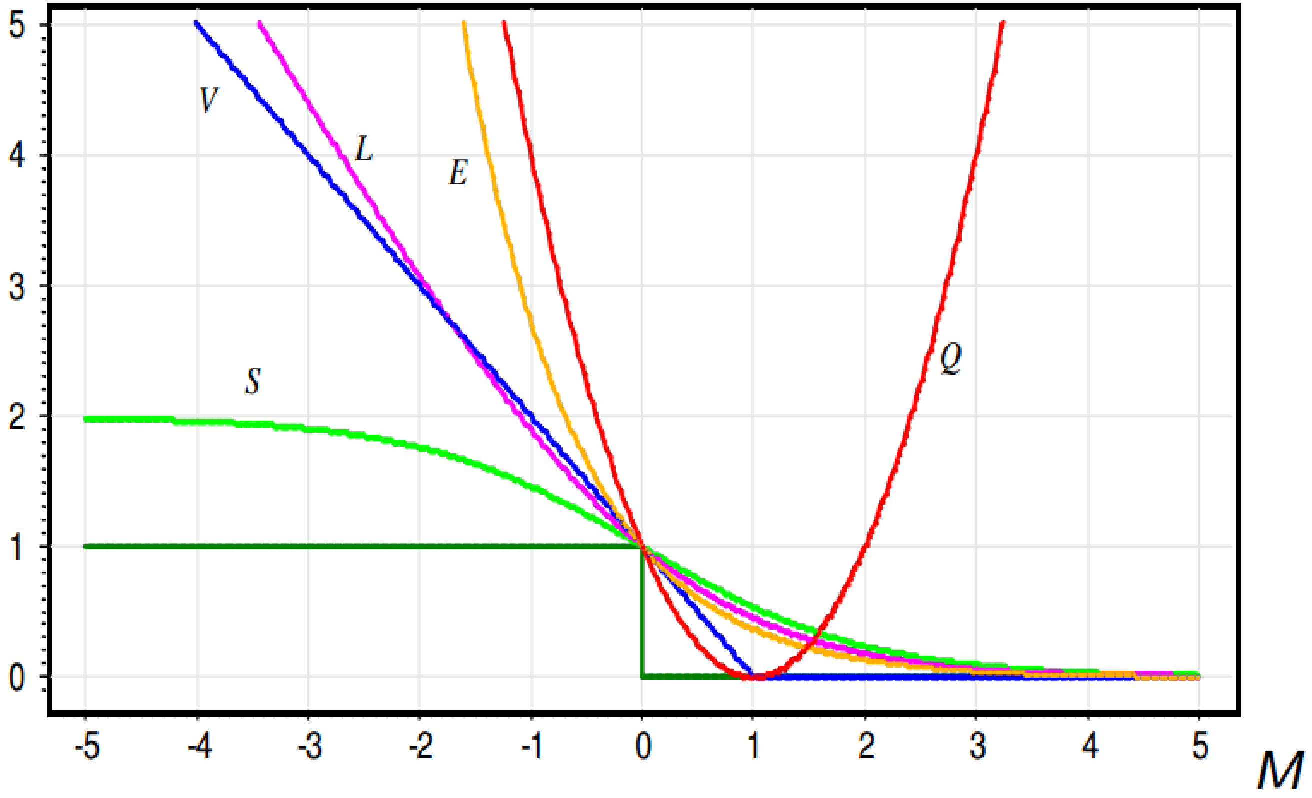
$$M_i = y_i f(x_i)$$

(y_i - класс, к которому относится x_i)

$$M_i \leq 0 \Leftrightarrow y_i \neq a(x_i)$$
$$M_i > 0 \Leftrightarrow y_i = a(x_i)$$

Функция потерь

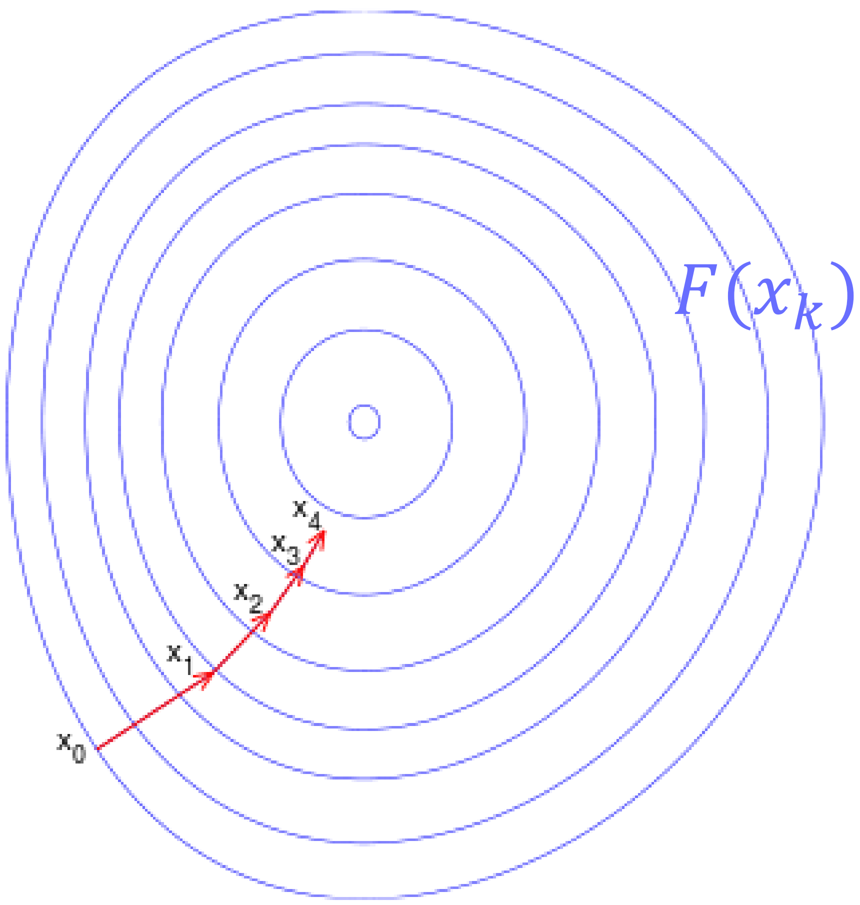
$$Q(w) = \sum_{i=1}^{\ell} [M_i(w) < 0] \leqslant \tilde{Q}(w) = \sum_{i=1}^{\ell} \mathscr{L}(M_i(w)) \to \min_w;$$



$$Q(M) = (1 - M)^2$$
$$V(M) = (1 - M)_+$$
$$S(M) = 2(1 + e^M)^{-1}$$
$$L(M) = \log_2(1 + e^{-M})$$
$$E(M) = e^{-M}$$

Градиентный спуск (GD, Gradient Decent)

$$x_{k+1} = x_k - \gamma_k \nabla F(x_k)$$



$$\nabla_w \tilde{Q} = \sum_{i=1}^{l} \nabla L(M_i) = \sum_{i=1}^{l} L'(M_i) \frac{\partial M_i}{\partial w}$$

$$M_i = y_i \langle w, x_i \rangle \Longrightarrow \frac{\partial M_i}{\partial w} = y_i x_i$$

$$\nabla \tilde{Q} = \sum_{i=1}^{l} y_i x_i L'(M_i)$$

$$\boxed{w_{k+1} = w_k - \gamma_k \sum_{i=1}^{l} y_i x_i L'(M_i)}$$

Стохастический градиент (SGD)

$$w_{k+1} = w_k - \gamma_k \cancel{\sum_{i=1}^{l}} y_i x_i L'(M_i)$$

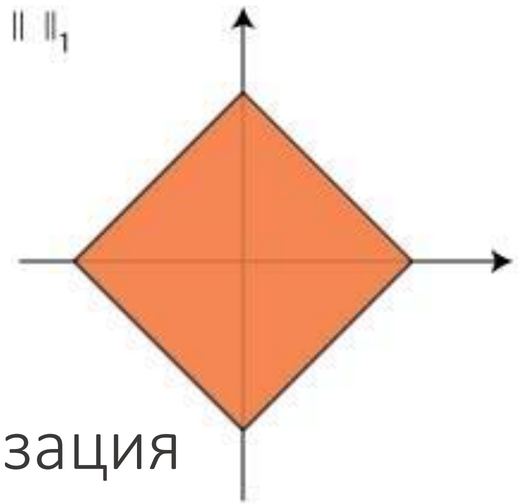
$$\boxed{w_{k+1} = w_k - \gamma_k y_i x_i L'(M_i)}$$

x_i – случайный элемент обучающей выборки

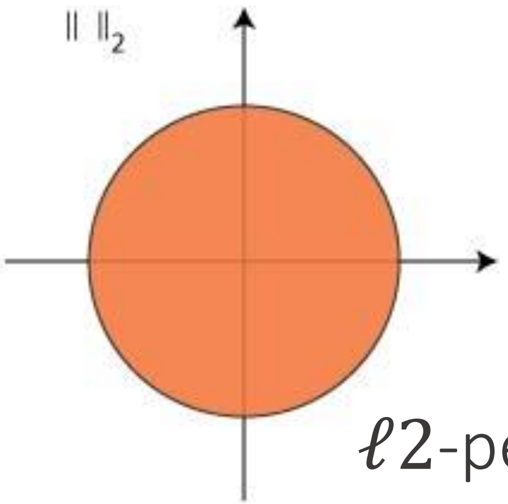
Регуляризация

$$\sum_{i=1}^{l} L(M_i) + \gamma \sum_{n=1}^{d} |w_n| \to min \qquad \sum_{i=1}^{l} L(M_i) + \gamma \sum_{n=1}^{d} {w_n}^2 \to min$$

Почему так – обсудим в математическом дополнении



$\ell 1$-регуляризация



$\ell 2$-регуляризация

Стандартные линейные классификаторы

Классификатор	Функция потерь	Регуляризатор
SVM (Support vector machine, метод опорных векторов)	$L(M) = \max\{0, 1 - M\} = $ $= (1 - M)_+$	$\sum_{k=1}^{m} {w_k}^2$
Логистическая регрессия	$L(M) = \log(1 + e^{-M})$	Обычно $\sum_{k=1}^{m} {w_k}^2$ или $\sum_{k=1}^{m} \lvert w_k \rvert$

Общий случай

$$Q = \sum_{i=1}^{\ell} L\big(y_i, f(x_i)\big) + \gamma V(w) \to \min_w$$

Функция потерь

Регуляризатор

Коэффициент
регуляризации

МЕТОД ОПОРНЫХ ВЕКТОРОВ

МЕТОД ОПОРНЫХ ВЕКТОРОВ

- Линейный классификатор:

$$a(x) = \mathrm{sign}(< w, x > -w_0)$$

- Использующий кусочно-линейную функцию потерь и $L2$-регуляризатор:

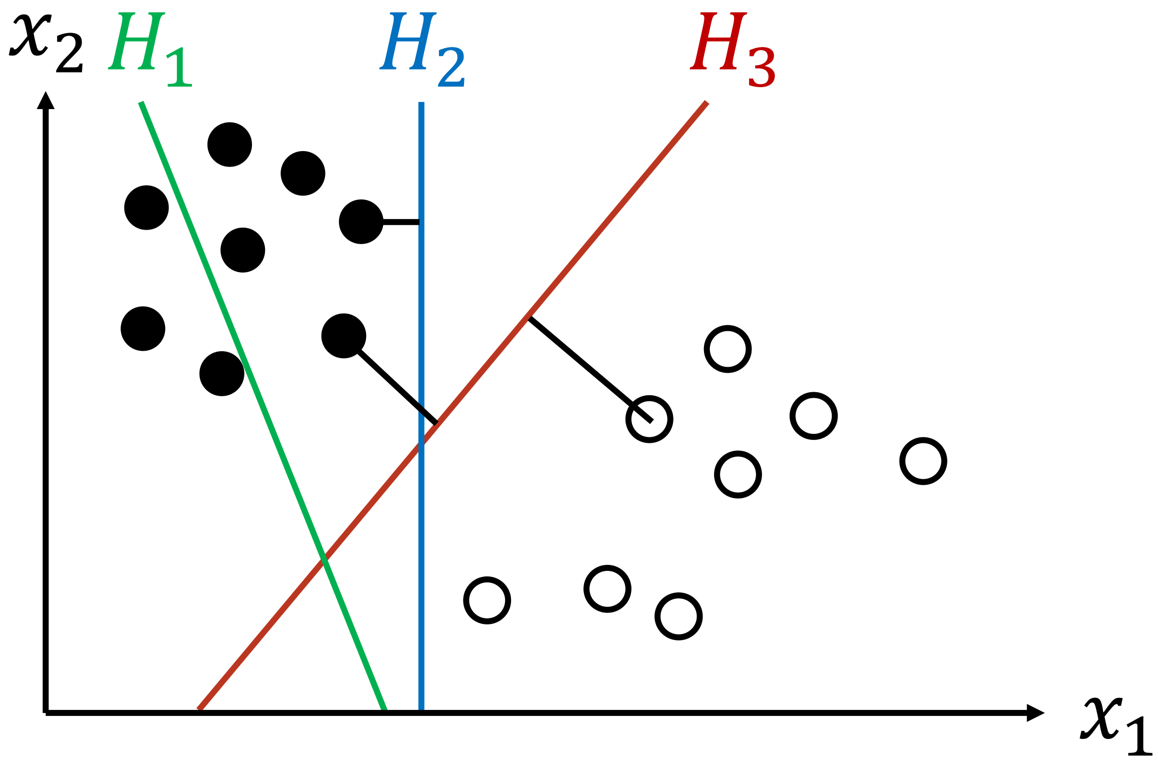
$$\sum_{i=1}^{l} \boxed{L(M_i)} + \boxed{\gamma \|w\|^2} \to \min_{w}$$

функция потерь   квадратичный регуляризатор

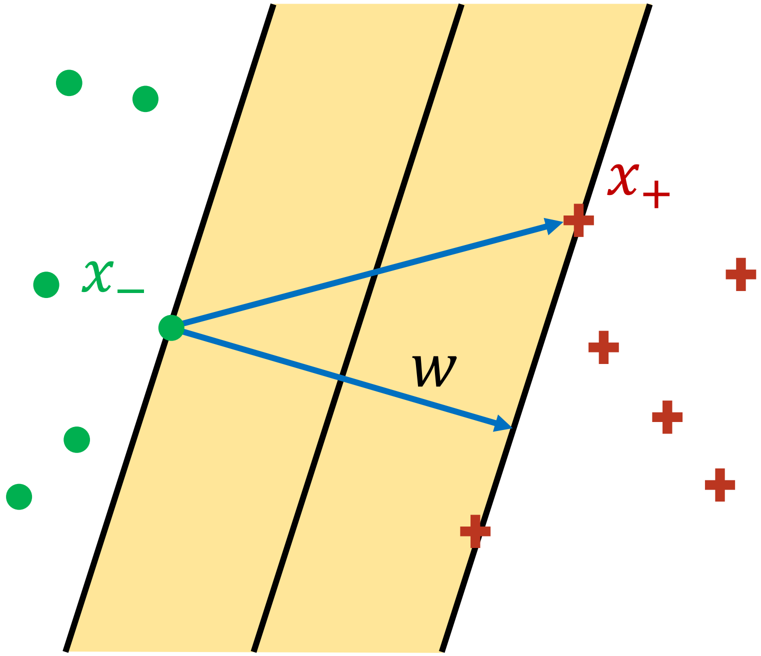
- Кусочно-линейная функция потерь (hinge loss):

$$L(M_i) = \max\{0, 1 - M_i\} = (1 - M_i)_+$$

ПОСТРОЕНИЕ РАЗДЕЛЯЮЩЕЙ ГИПЕРПЛОСКОСТИ



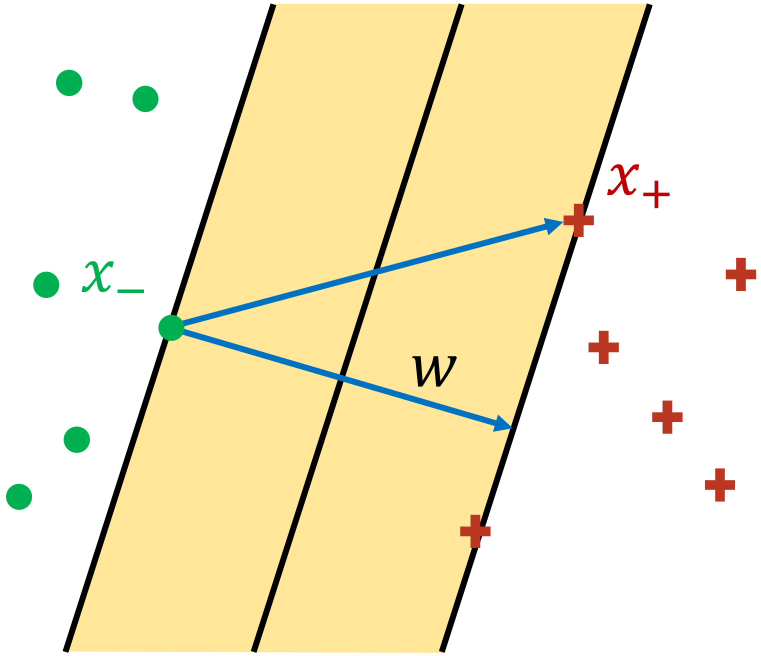
РАЗДЕЛЯЮЩАЯ ПОЛОСА



$$a(x) = \operatorname{sign}\left(\sum_{j=1}^{n} w_j x^j - w_0\right) = \operatorname{sign}(\langle w, x\rangle - w_0)$$

РАЗДЕЛЯЮЩАЯ ПОЛОСА

Обозначим $\delta = (\langle w, x_+ \rangle - w_0)$

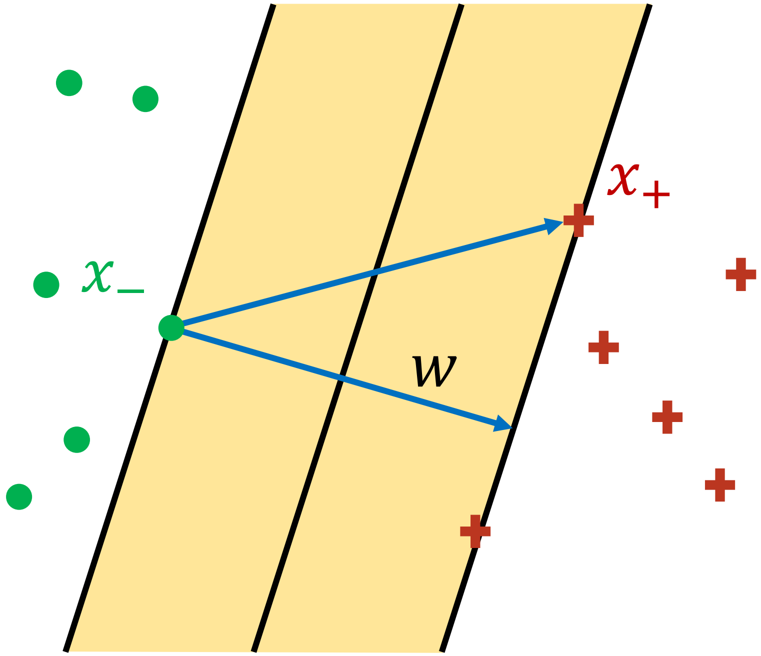


$$a(x) = \mathrm{sign}\left(\sum_{j=1}^{n} w_j x^j - w_0\right) = \mathrm{sign}(\langle w, x \rangle - w_0)$$

РАЗДЕЛЯЮЩАЯ ПОЛОСА

Обозначим $\delta = (\langle w, x_+\rangle - w_0)$

Тогда $(\langle w, x_-\rangle - w_0) = -\delta$



$$a(x) = \mathrm{sign}\left(\sum_{j=1}^{n} w_j x^j - w_0\right) = \mathrm{sign}(\langle w, x\rangle - w_0)$$

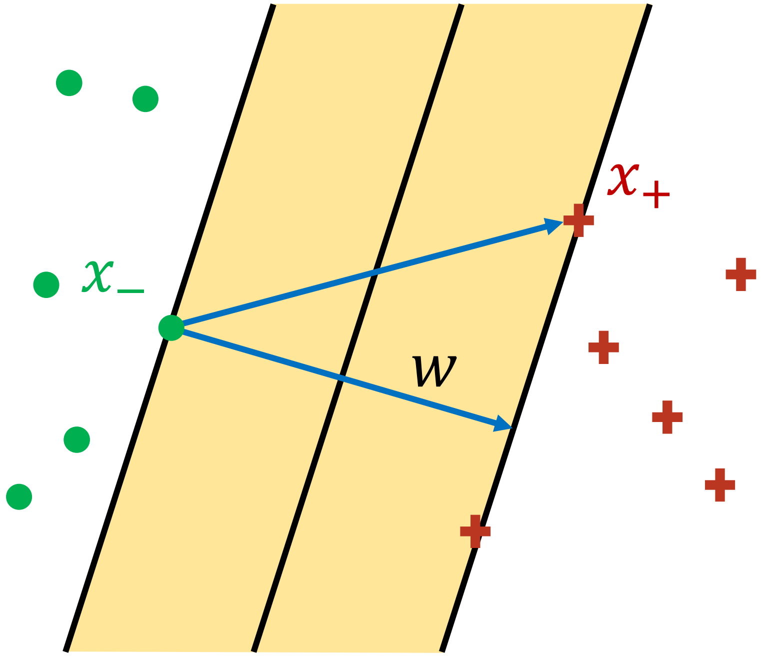
РАЗДЕЛЯЮЩАЯ ПОЛОСА

Обозначим $\delta = (\langle w, x_+\rangle - w_0)$

Тогда $(\langle w, x_-\rangle - w_0) = -\delta$

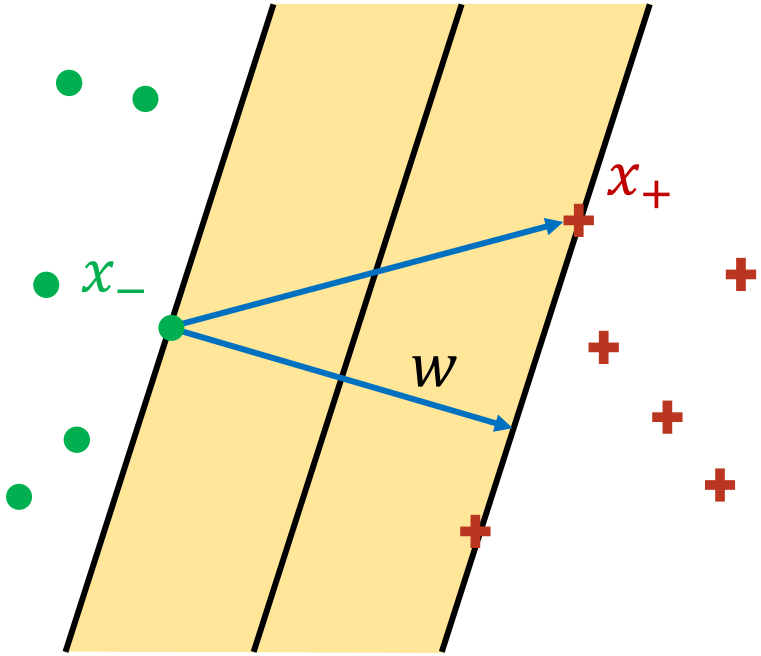
То же самое, но другими словами:

$$\min_{i=1,\dots,l} y_i(\langle w, x_i\rangle - w_0) = \delta$$



$$a(x) = \mathrm{sign}\left(\sum_{j=1}^{n} w_j x^j - w_0\right) = \mathrm{sign}(\langle w, x\rangle - w_0)$$

РАЗДЕЛЯЮЩАЯ ПОЛОСА



Обозначим $\delta = (\langle w, x_+ \rangle - w_0)$

Тогда $(\langle w, x_- \rangle - w_0) = -\delta$

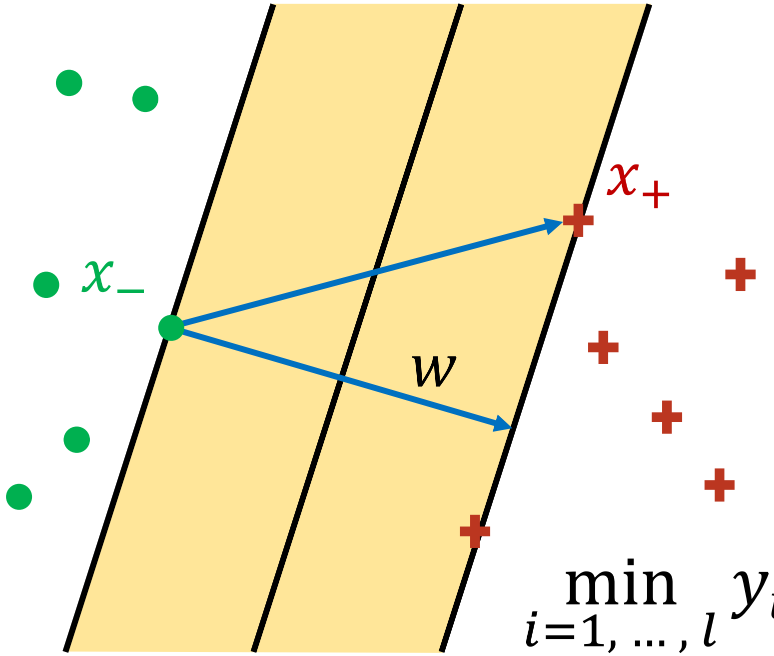
То же самое, но другими словами:

$$\min_{i=1,\dots,l} y_i(\langle w, x_i \rangle - w_0) = \delta$$

А если разделим w и w_0 на δ (разделяющая поверхность не поменяется): $\min_{i=1,\dots,l} y_i(\langle w, x_i \rangle - w_0) = 1$

$$a(x) = \operatorname{sign}\left(\sum_{j=1}^{n} w_j x^j - w_0\right) = \operatorname{sign}(\langle w, x \rangle - w_0)$$

РАЗДЕЛЯЮЩАЯ ПОЛОСА



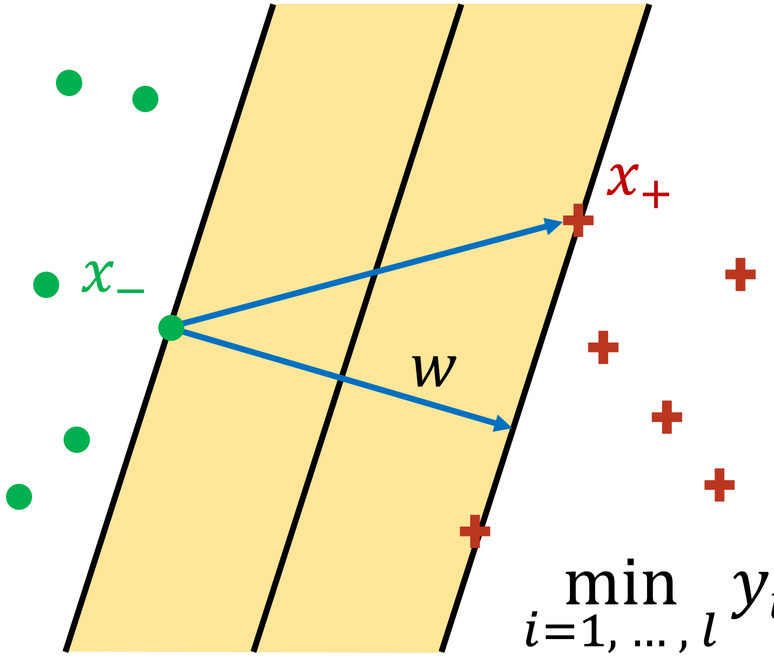
$$\langle w, x_+ \rangle - w_0 = +1$$

$$\langle w, x_- \rangle - w_0 = -1$$

$$\min_{i=1,\dots,l} y_i(\langle w, x_i \rangle - w_0) = 1$$

$$a(x) = \mathrm{sign}\left(\sum_{j=1}^{n} w_j x^j - w_0\right) = \mathrm{sign}(\langle w, x \rangle - w_0)$$

ШИРИНА РАЗДЕЛЯЮЩЕЙ ПОЛОСЫ



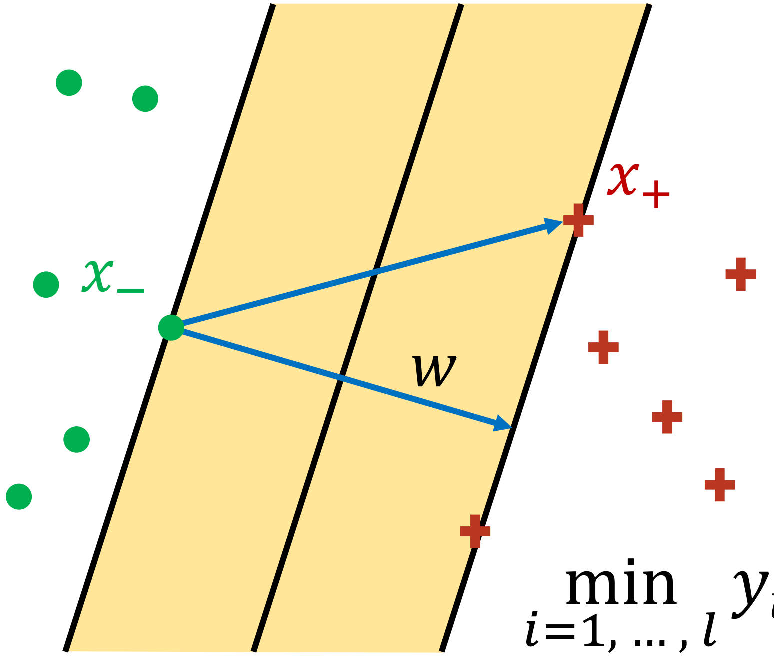
$$\langle w, x_+ \rangle - w_0 = +1$$

$$\langle w, x_- \rangle - w_0 = -1$$

$$\min_{i=1,\dots,l} y_i(\langle w, x_i \rangle - w_0) = 1$$

$$\left\langle (x_+ - x_-), \frac{w}{\|w\|} \right\rangle$$

ШИРИНА РАЗДЕЛЯЮЩЕЙ ПОЛОСЫ



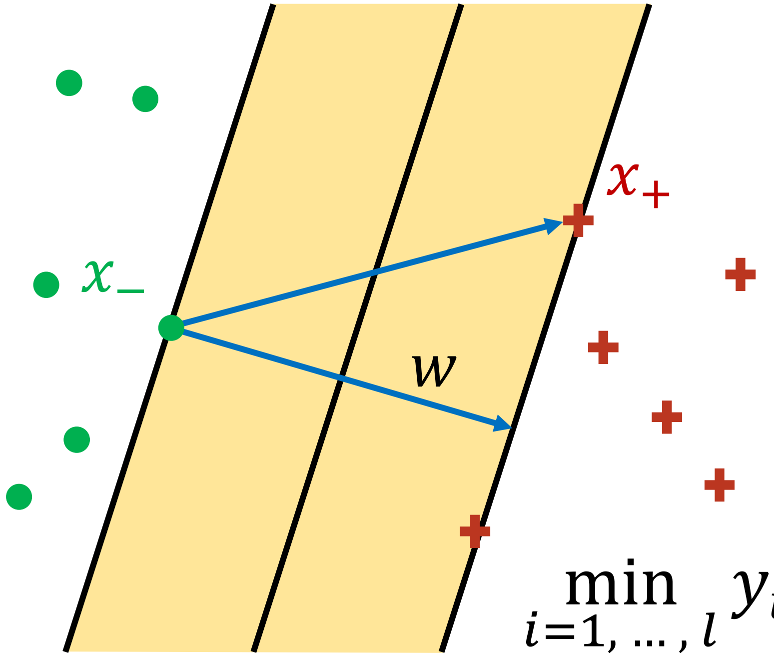
$$\langle w, x_+\rangle - w_0 = +1$$

$$\langle w, x_-\rangle - w_0 = -1$$

$$\min_{i=1,\dots,l} y_i(\langle w, x_i\rangle - w_0) = 1$$

$$\left\langle (x_+ - x_-), \frac{w}{\|w\|}\right\rangle = \frac{\langle w,x_+\rangle - \langle w,x_-\rangle}{\|w\|}$$

ШИРИНА РАЗДЕЛЯЮЩЕЙ ПОЛОСЫ

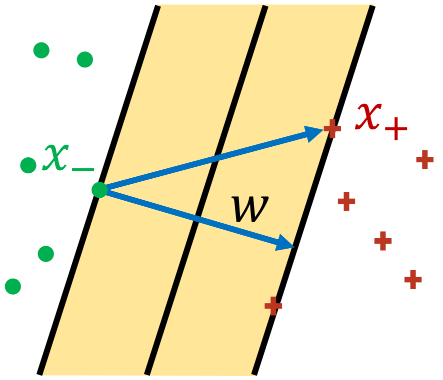


$$\langle w, x_+\rangle - w_0 = +1$$
$$\langle w, x_-\rangle - w_0 = -1$$

$$\min_{i=1,\dots,l} y_i(\langle w, x_i\rangle - w_0) = 1$$

$$\left\langle (x_+ - x_-), \frac{w}{\|w\|} \right\rangle = \frac{\langle w, x_+\rangle - \langle w, x_-\rangle}{\|w\|} = \frac{(w_0+1)-(w_0-1)}{\|w\|} = \frac{2}{\|w\|}$$

ШИРИНА РАЗДЕЛЯЮЩЕЙ ПОЛОСЫ

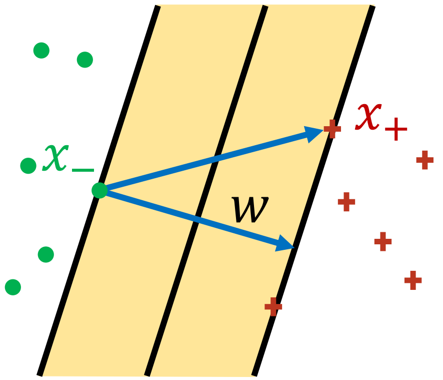


$$\min_{i=1,\dots,l} y_i(\langle w, x_i\rangle - w_0) = 1$$

$$\langle w, x_+\rangle - w_0 = +1$$
$$\langle w, x_-\rangle - w_0 = -1$$

$$\left\langle (x_+ - x_-), \frac{w}{\|w\|} \right\rangle = \frac{\langle w, x_+\rangle - \langle w, x_-\rangle}{\|w\|} = \frac{(w_0+1)-(w_0-1)}{\|w\|} = \frac{2}{\|w\|}$$

МАКСИМИЗАЦИЯ ЗАЗОРА



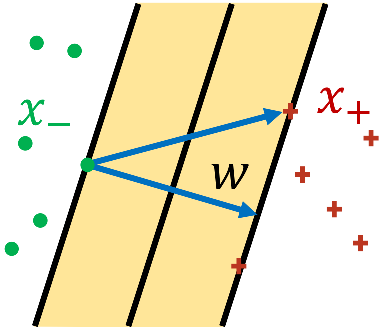
$$\min_{i=1,\ldots,l} y_i(\langle w, x_i\rangle - w_0) = 1$$

$$\langle w, x_+\rangle - w_0 = +1$$
$$\langle w, x_-\rangle - w_0 = -1$$

$$\left\langle (x_+ - x_-), \frac{w}{\|w\|} \right\rangle = \frac{\langle w,x_+\rangle - \langle w,x_-\rangle}{\|w\|} = \frac{(w_0+1)-(w_0-1)}{\|w\|} = \frac{2}{\|w\|}$$

$$\begin{cases} \langle w, w\rangle \to \min \\ y_i(\langle w, x_i\rangle - w_0) \geq 1, i = 1, \ldots, l \end{cases}$$

СЛУЧАЙ ЛИНЕЙНО НЕРАЗДЕЛИМОЙ ВЫБОРКИ



$$\begin{cases} \langle w, w \rangle \to \min \\ y_i(\langle w, x_i \rangle - w_0) \geq 1, i = 1, \dots, l \end{cases}$$

$$\begin{cases} \frac{1}{2}\langle w, w \rangle + C \sum_{i=1}^{l} \xi_i \to \min\limits_{w, w_0, \xi} \\ y_i(\langle w, x_i \rangle - w_0) \geq 1 - \xi_i, i = 1, \dots, l \\ \xi_\mathrm{i} \geq 0, i = 1, \dots, l \end{cases}$$

ОПТИМИЗИЦИОННАЯ ЗАДАЧА В SVM

$$\begin{cases} \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{l} \xi_i \to \min\limits_{w, w_0, \xi} \\ y_i(\langle w, x_i\rangle - w_0) \geq 1 - \xi_i, i = 1, \dots, l \\ \xi_i \geq 0, i = 1, \dots, l \end{cases}$$

Причем здесь линейный классификатор в привычном нам виде?

БЕЗУСЛОВНАЯ ОПТИМИЗИЦИОННАЯ ЗАДАЧА В SVM

$$\begin{cases} \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{l} \xi_i \to \min_{w, w_0, \xi} \\ y_i(\langle w, x_i\rangle - w_0) \geq 1 - \xi_i, i = 1, \dots, l \\ \xi_i \geq 0, i = 1, \dots, l \end{cases}$$

Напоминание:

$M_i = y_i(\langle w, x_i\rangle - w_0)$

отступ на i-том объекте

БЕЗУСЛОВНАЯ ОПТИМИЗИЦИОННАЯ ЗАДАЧА В SVM

$$\begin{cases} \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{l} \xi_i \to \min\limits_{w, w_0, \xi} \\ y_i(\langle w, x_i\rangle - w_0) \geq 1 - \xi_i, i = 1, \dots, l \\ \xi_i \geq 0, i = 1, \dots, l \end{cases}$$

$$\xi_i \geq 0$$

$$\xi_i \geq 1 - M_i$$

$$\sum_{i=1}^{l} \xi_i \to \min$$

Напоминание:

$M_i = y_i(\langle w, x_i\rangle - w_0)$

отступ на i-том объекте

БЕЗУСЛОВНАЯ ОПТИМИЗИЦИОННАЯ ЗАДАЧА В SVM

$$\begin{cases} \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{l}\xi_i \to \min\limits_{w, w_0, \xi} \\ y_i(\langle w, x_i\rangle - w_0) \geq 1 - \xi_i, i = 1, \dots, l \\ \xi_i \geq 0, i = 1, \dots, l \end{cases}$$

Напоминание:

$M_i = y_i(\langle w, x_i\rangle - w_0)$

отступ на i-том объекте

$$\xi_i \geq 0$$

$$\xi_i \geq 1 - M_i$$

$$\sum_{i=1}^{l}\xi_i \to \min$$

$$\Rightarrow \xi_i = \max\{0{,}1 - M_i\} = (1 - M_i)_+$$

БЕЗУСЛОВНАЯ ОПТИМИЗИЦИОННАЯ ЗАДАЧА В SVM

$$\begin{cases} \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{l}\xi_i \to \min\limits_{\mathrm{w,w_0},\xi} \\ y_i(\langle w, x_i\rangle - w_0) \geq 1 - \xi_i, i = 1, \dots, l \\ \xi_i \geq 0, i = 1, \dots, l \end{cases}$$

Напоминание:

$M_i = y_i(\langle w, x_i\rangle - w_0)$

отступ на i-том объекте

$$\xi_i \geq 0$$

$$\xi_i \geq 1 - M_i$$

$$\Rightarrow \xi_i = \max\{0,1 - M_i\} = (1 - M_i)_+$$

$$\sum_{i=1}^{l}\xi_i \to \min$$

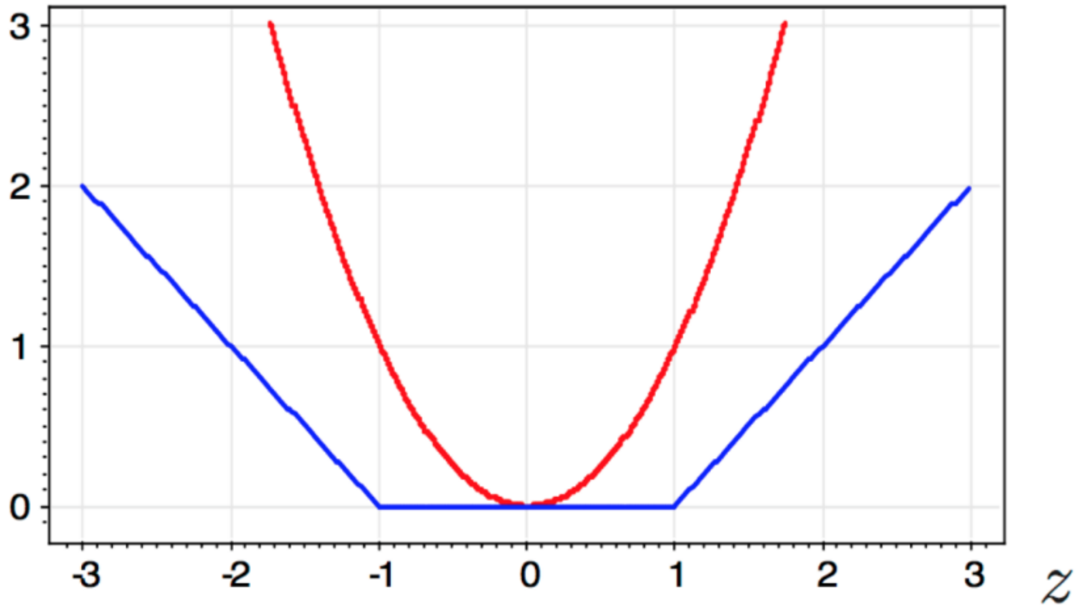
$$Q(w, w_0) = \sum_{i=1}^{l}(1 - M_i(w, w_0))_+ + \frac{1}{2C}\|w\|^2 \to \min_{w,w_0}$$

БЕЗУСЛОВНАЯ ОПТИМИЗИЦИОННАЯ ЗАДАЧА В SVM

$$Q(w, w_0) = \sum_{i=1}^{l} (1 - M_i(w, w_0))_+ + \frac{1}{2C} \|w\|^2 \to \min_{w, w_0}$$

БЕЗУСЛОВНАЯ ОПТИМИЗИЦИОННАЯ ЗАДАЧА В SVM: РЕГРЕССИЯ

$$Q_\epsilon(a, X^l) = \sum_{i=1}^{l} |< w, x_i > -w_0 - y_i|_\epsilon + \tau < w, w > \to min_{w,w_0}$$



РЕЗЮМЕ

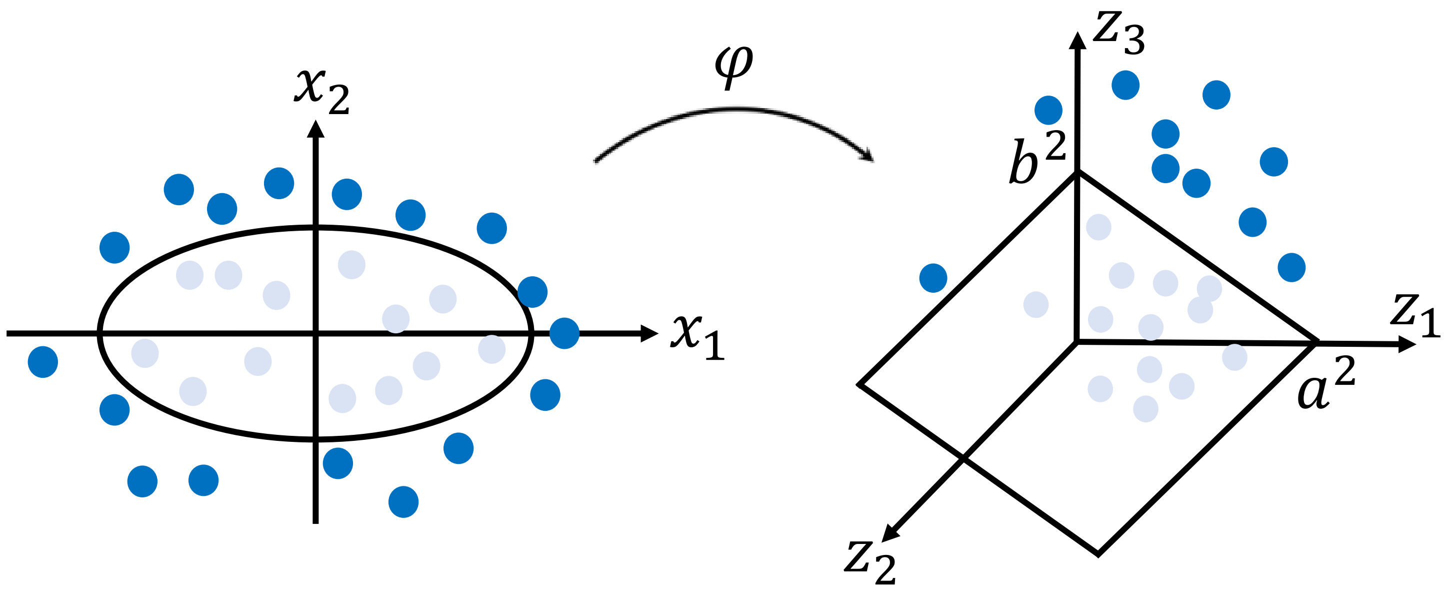
- Метод опорных векторов – линейный классификатор с кусочно-линейной функцией потерь (hinge loss) и $L2$-регуляризатором
- Придуман метод был из соображений максимизации зазора между классами

РЕЗЮМЕ

- В случае линейно разделимой выборки это означает просто максимизацию ширины разделяющей полосы
- А в случае линейно неразделимой выборки просто добавляется возможность попадания объектов в полосу и штрафы за это

ЯДРА В МЕТОДЕ ОПОРНЫХ ВЕКТОРОВ

ДОБАВЛЕНИЕ НОВЫХ ПРИЗНАКОВ



Спрямляющее пространство

$$\varphi: (x_1, x_2) \to (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

$$\left(\frac{x_1}{a}\right)^2 + \left(\frac{x_2}{b}\right)^2 = 1 \to \frac{z_1}{a^2} + \frac{z_3}{b^2} = 1$$

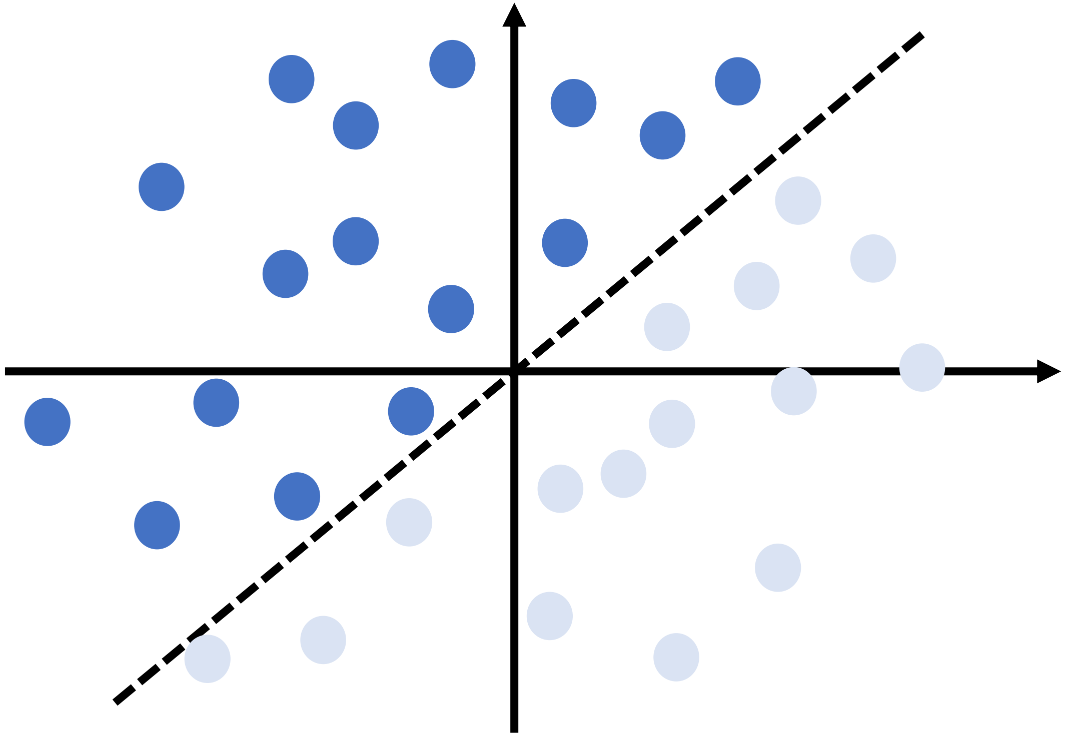
KERNEL TRICK

$$\left.\begin{array}{r} x \mapsto \varphi(x) \\ w \mapsto \varphi(\mathrm{w}) \end{array}\right. \Rightarrow \langle w, x\rangle \mapsto \langle \varphi(w), \varphi(x)\rangle$$

Можно не делать преобразование признаков явно, а вместо скалярного произведения $\langle \mathrm{w}, x\rangle$ использовать функцию $K(w, x)$, представимую в виде:

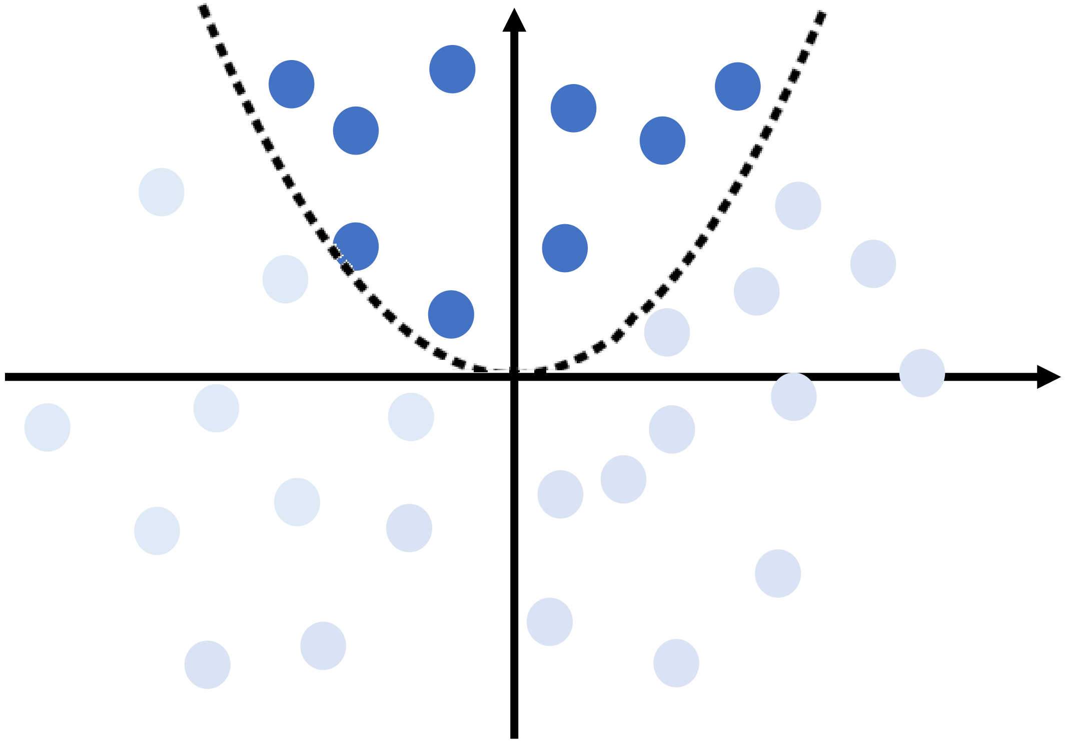
$$K(w, x) = \langle \varphi(w), \varphi(x)\rangle$$

ЛИНЕЙНОЕ ЯДРО



$$K(w, x) = \langle w, x \rangle$$

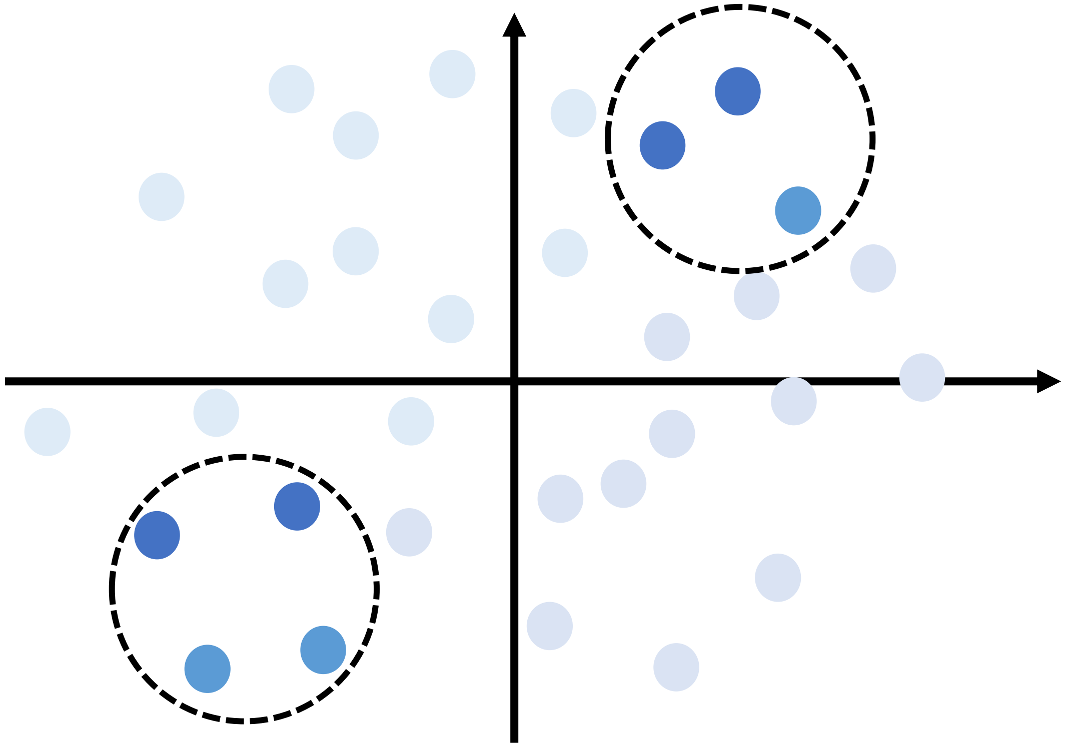
ПОЛИНОМИАЛЬНОЕ ЯДРО



$$K(w, x) = (\gamma \langle w, x \rangle + r)^d$$

Упражнение: какое спрямляющее пространство соответствует этому ядру? Какая у него размерность?

РАДИАЛЬНОЕ ЯДРО



$$K(w, x) = e^{-\gamma \|w - x\|^2}$$

Упражнение: какое спрямляющее пространство соответствует этому ядру? Какая у него размерность?

ЯДРА И БИБЛИОТЕКИ

- LibSVM – можно выбирать ядра
- LibLinear – только линейное ядро
- В scikit-learn: обёртка над LibSVM и LibLinear
- Vowpal Wabbit – только линейное ядро

РЕЗЮМЕ

- Kernel trick
- Линейное, полиномиальное
  и радиальное ядра
- Библиотеки

Математическое дополнение: условный экстремум

Метод множителей Лагранжа: пример

$$f(X) = x_1^2 + x_2^2 \to extr \qquad \varphi_1(X) = x_1 + x_2 = 2$$

Метод множителей Лагранжа: пример

$$f(X) = x_1^2 + x_2^2 \to extr \qquad \varphi_1(X) = x_1 + x_2 = 2$$

1. Запишем функцию Лагранжа: $L(X,\lambda) = x_1^2 + x_2^2 + \lambda_1(x_1 + x_2 - 2)$
2. Запишем необходимые условия экстремума.

$$\left.\begin{aligned} \frac{\partial L}{\partial x_1} &= 2x_1 + \lambda_1 = 0 \\ \frac{\partial L}{\partial x_2} &= 2x_2 + \lambda_1 = 0 \\ x_1 + x_2 - 2 &= 0 \end{aligned}\right\}$$

3. Найдем координаты условно-стационарных точек.

$$\begin{cases} 2x_1 + \lambda_1 = 0 \\ 2x_2 + \lambda_1 = 0 \\ x_1 + x_2 - 2 = 0 \end{cases} \Rightarrow \begin{cases} 2x_1 + \lambda_1 = 0 \\ x_1 - x_2 = 0 \\ x_1 + x_2 - 2 = 0 \end{cases} \Rightarrow \begin{cases} x_1^* = 1 \\ x_2^* = 1 \\ \lambda_1^* = -2 \end{cases}$$

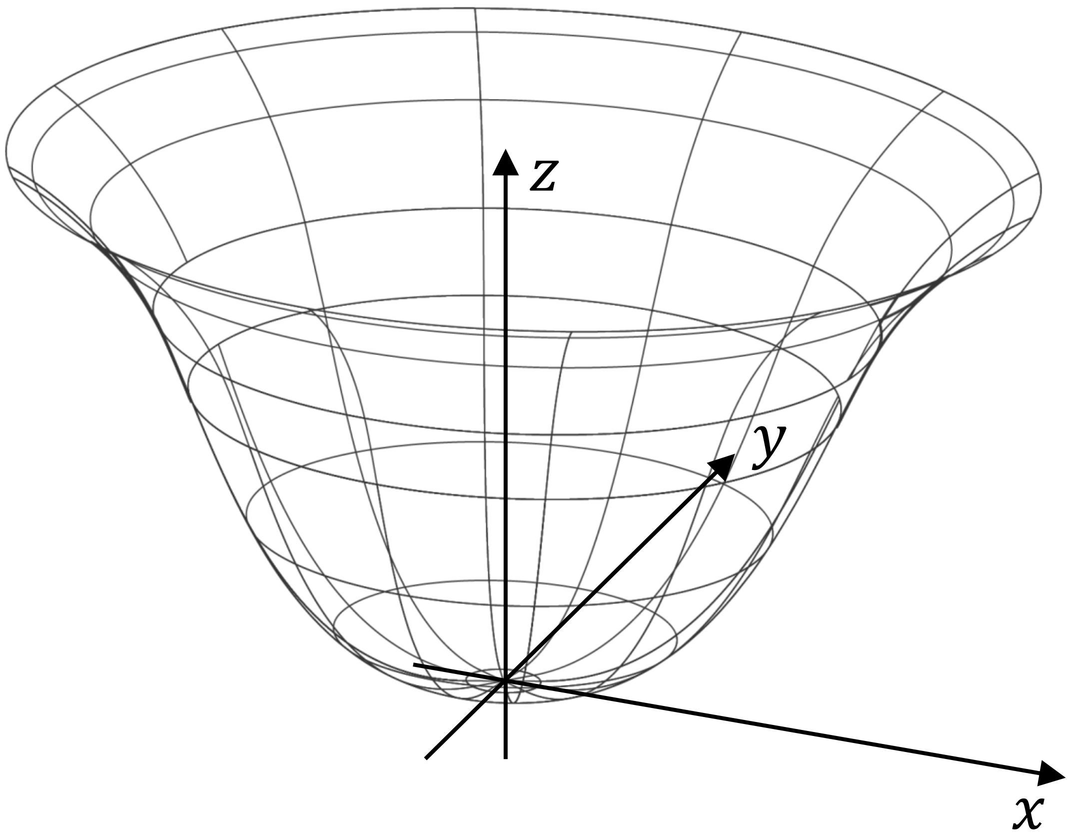
ЧАСТНЫЕ ПРОИЗВОДНЫЕ

- $\frac{\Delta f}{\Delta x} \xrightarrow{\Delta x \to 0} f'_x,\ \ y$- фиксирован
- $\frac{\Delta f}{\Delta y} \xrightarrow{\Delta y \to 0} f'_y,\ \ x$ - фиксирован

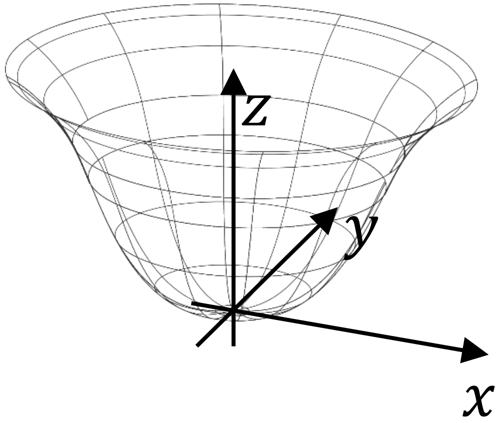
ГРАДИЕНТ

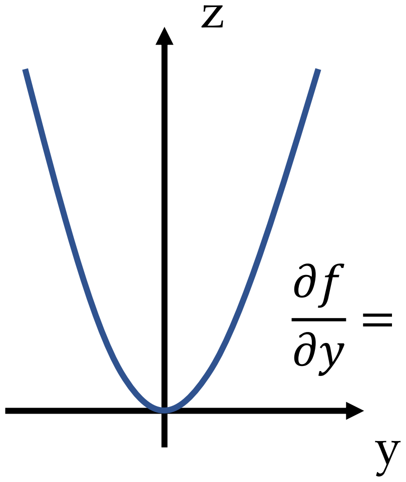
$$\nabla f(x_0, y_0) = \begin{pmatrix} f_x' & (x_0, y_0) \\ f_y' & (x_0, y_0) \end{pmatrix}$$

НЕОБХОДИМОЕ УСЛОВИЕ БЕЗУСЛОВНОГО ЭКСТРЕМУМА

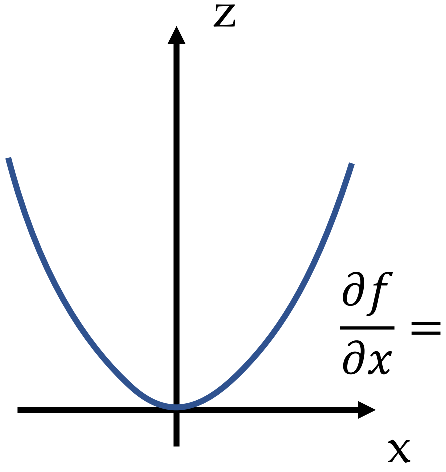


НЕОБХОДИМОЕ УСЛОВИЕ БЕЗУСЛОВНОГО ЭКСТРЕМУМА





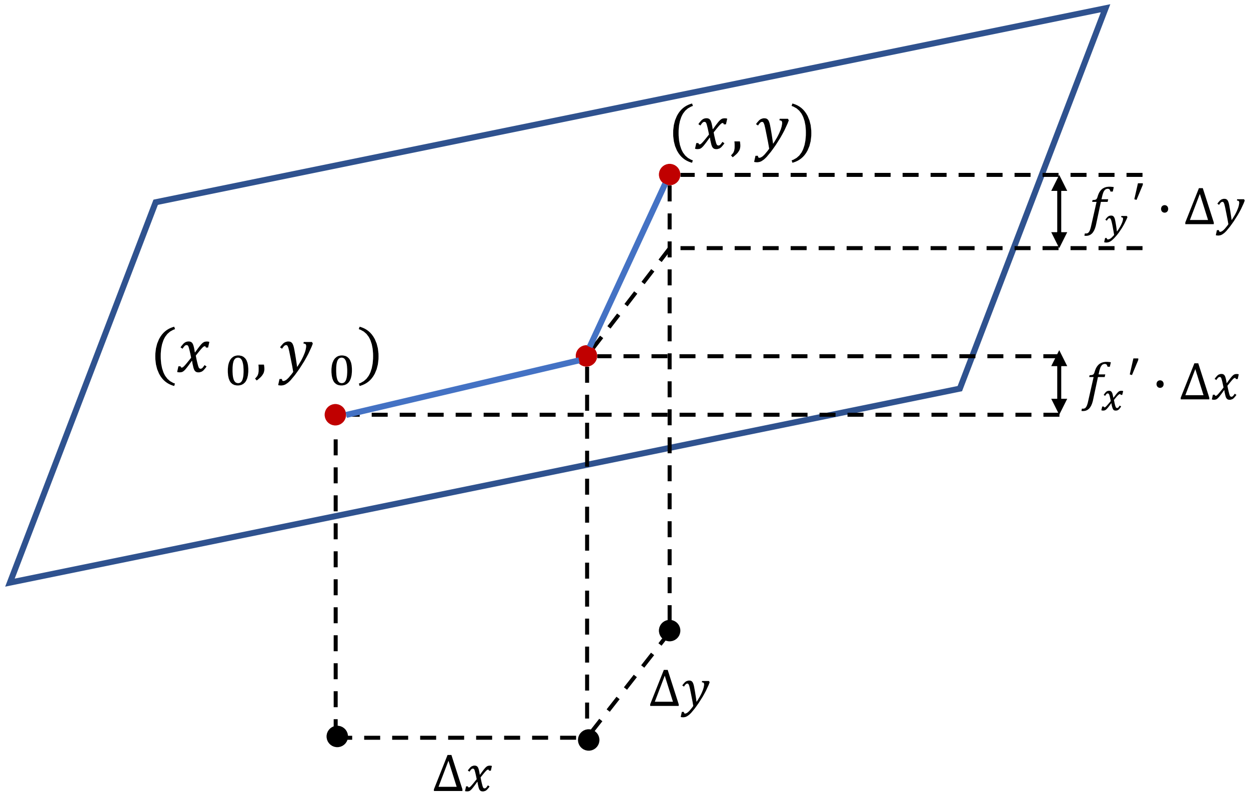
$$\frac{\partial f}{\partial y} = 0$$



$$\frac{\partial f}{\partial x} = 0$$

$$\nabla f = 0$$

ГЕОМЕТРИЧЕСКИЙ СМЫСЛ ЧАСТНЫХ ПРОИЗВОДНЫХ



$$\Delta f = f_x' \cdot \Delta x + f_y' \cdot \Delta y$$

ЛИНЕЙНАЯ ЧАСТЬ ПРИРАЩЕНИЯ

$$\nabla f(x_0, y_0) = \begin{pmatrix} f'_x & (x_0, y_0) \\ f'_y & (x_0, y_0) \end{pmatrix}$$

$\Delta f = f'_x \cdot \Delta x + f'_y \cdot \Delta y$ – линейная часть приращения

$$f(x) \approx f(x_0) + f'_x \cdot \Delta x + f'_y \cdot \Delta y = f(x_0) + \left\langle \nabla f(x_0, y_0), \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \right\rangle$$

ГРАДИЕНТ – НАПРАВЛЕНИЕ НАИСКОРЕЙШЕГО РОСТА

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x_0}) + \langle \nabla f(\boldsymbol{x_0}), \boldsymbol{x} - \boldsymbol{x_0} \rangle$$

Пусть $\boldsymbol{x} = \boldsymbol{x_0} + \eta \boldsymbol{g}$, где $\boldsymbol{g}$ это единичный вектор, сонаправленный $\boldsymbol{x} - \boldsymbol{x_0}$:

$$f(x) - f(x_0) \approx \langle \nabla f(x_0), \eta \boldsymbol{g} \rangle$$

ГРАДИЕНТ – НАПРАВЛЕНИЕ НАИСКОРЕЙШЕГО РОСТА

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x_0}) + \langle \nabla f(\boldsymbol{x_0}), \boldsymbol{x} - \boldsymbol{x_0} \rangle$$

Пусть $\boldsymbol{x} = \boldsymbol{x_0} + \eta \boldsymbol{g}$, где $\boldsymbol{g}$ это единичный вектор, сонаправленный $\boldsymbol{x} - \boldsymbol{x_0}$:

$$f(x) - f(x_0) \approx \langle \nabla f(x_0), \eta \boldsymbol{g} \rangle$$

Если $|\boldsymbol{x} - \boldsymbol{x_0}| = \eta$ зафиксировано, какой вектор $\boldsymbol{g}$ максимизирует $f(x) - f(x_0)$?

ГРАДИЕНТ – НАПРАВЛЕНИЕ НАИСКОРЕЙШЕГО РОСТА

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x_0}) + \langle \nabla f(\boldsymbol{x_0}), \boldsymbol{x} - \boldsymbol{x_0} \rangle$$

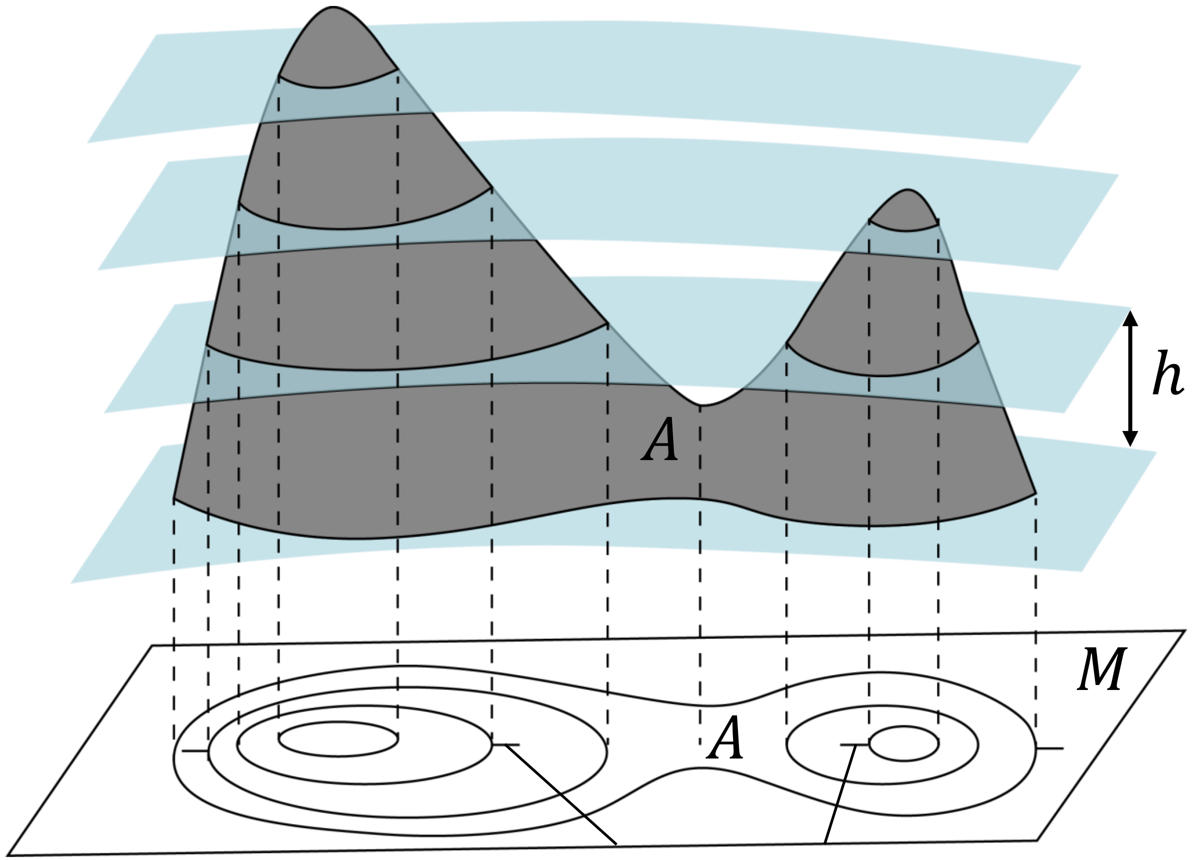
Пусть $\boldsymbol{x} = \boldsymbol{x_0} + \eta \boldsymbol{g}$, где $\boldsymbol{g}$ это единичный вектор, сонаправленный $\boldsymbol{x} - \boldsymbol{x_0}$:

$$f(x) - f(x_0) \approx \langle \nabla f(x_0), \eta \boldsymbol{g} \rangle$$

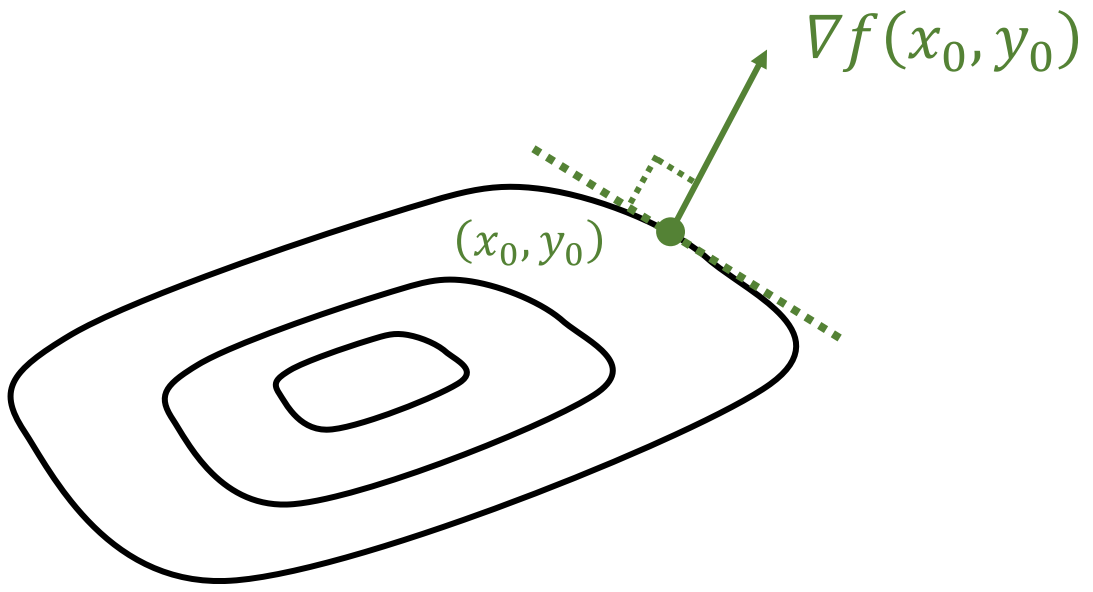
Если $|\boldsymbol{x} - \boldsymbol{x_0}| = \eta$ зафиксировано, какой вектор $\boldsymbol{g}$ максимизирует $f(x) - f(x_0)$?

$\boldsymbol{g}$ сонаправленный $\nabla f(x_0)$

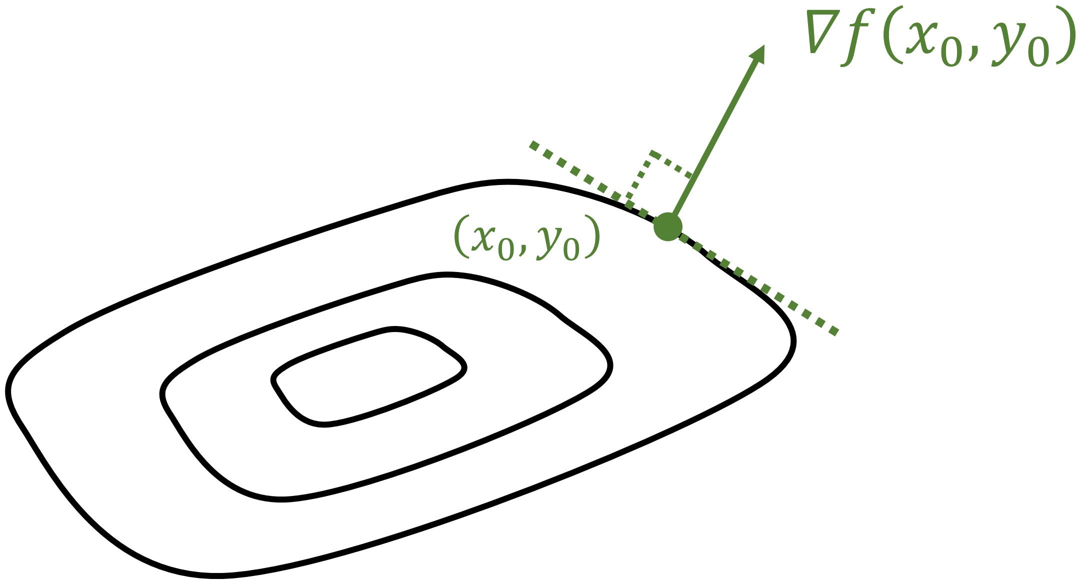
ЛИНИИ УРОВНЯ



ЛИНИИ УРОВНЯ И ГРАДИЕНТ

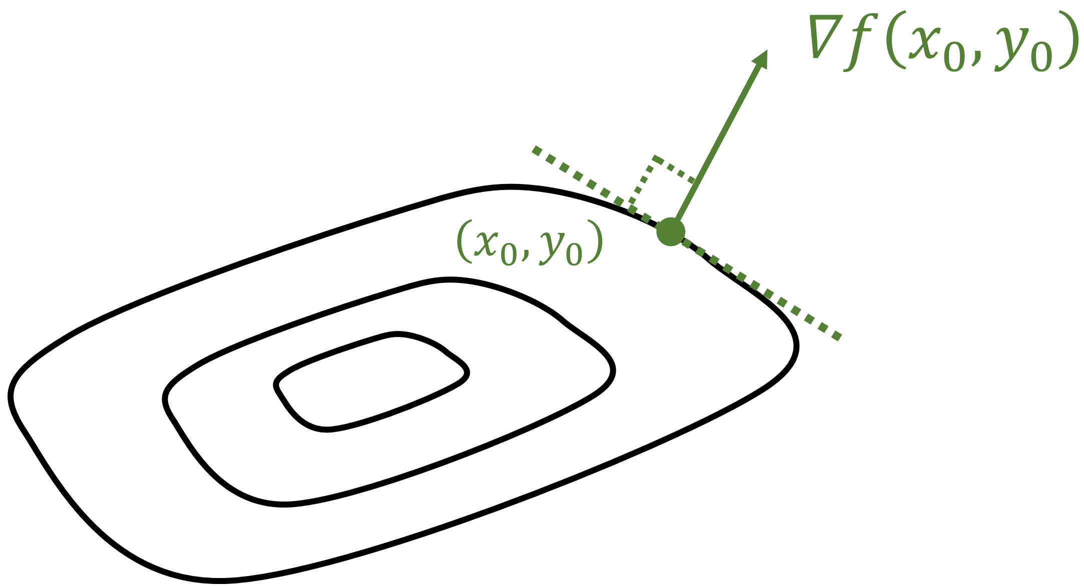


ЛИНИИ УРОВНЯ И ГРАДИЕНТ



$$f(x, y) \approx f(x_0, y_0) + \left\langle \nabla f(x_0, y_0), \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \right\rangle$$

ЛИНИИ УРОВНЯ И ГРАДИЕНТ



$$f(x, y) \approx f(x_0, y_0) + \left\langle \nabla f(x_0, y_0), \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \right\rangle$$

Чтобы при сдвиге вдоль линии уровня второе слагаемое было нулевым, градиент должен быть ортогонален линии уровня

Условный экстремум

$$\begin{cases} f(x,y) \to min \\ g(x,y) = 0 \end{cases}$$

Метод множителей Лагранжа

$$\begin{cases} f(x,y) \to min \\ g(x,y) = 0 \end{cases} \quad \Longrightarrow \quad \begin{aligned} & L(x,y) = f(x,y) + \lambda g(x,y) \\ & \nabla L(x,y) = 0 \\ & g(x,y) = 0 \end{aligned}$$

Метод множителей Лагранжа: пример

$$f(X) = x_1^2 + x_2^2 \to \text{extr} \qquad \varphi_1(X) = x_1 + x_2 = 2$$

Метод множителей Лагранжа: пример

$$f(X) = x_1^2 + x_2^2 \to extr \qquad \varphi_1(X) = x_1 + x_2 = 2$$

1. Запишем функцию Лагранжа: $L(X,\lambda) = x_1^2 + x_2^2 + \lambda_1(x_1 + x_2 - 2)$

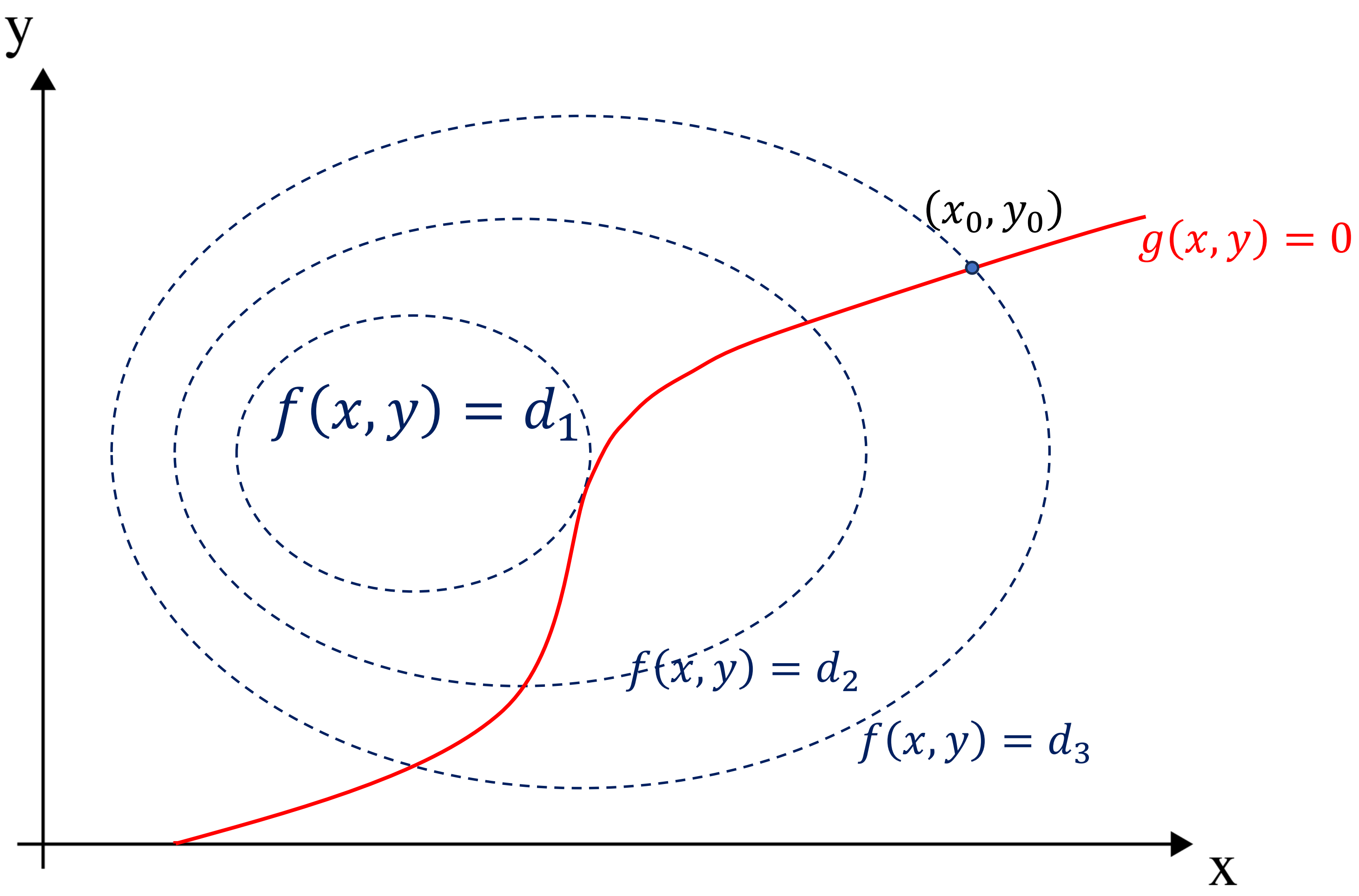
2. Запишем необходимые условия экстремума.

$$\left.\begin{aligned} \frac{\partial L}{\partial x_1} &= 2x_1 + \lambda_1 = 0 \\ \frac{\partial L}{\partial x_2} &= 2x_2 + \lambda_1 = 0 \\ x_1 + x_2 - 2 &= 0 \end{aligned}\right\}$$

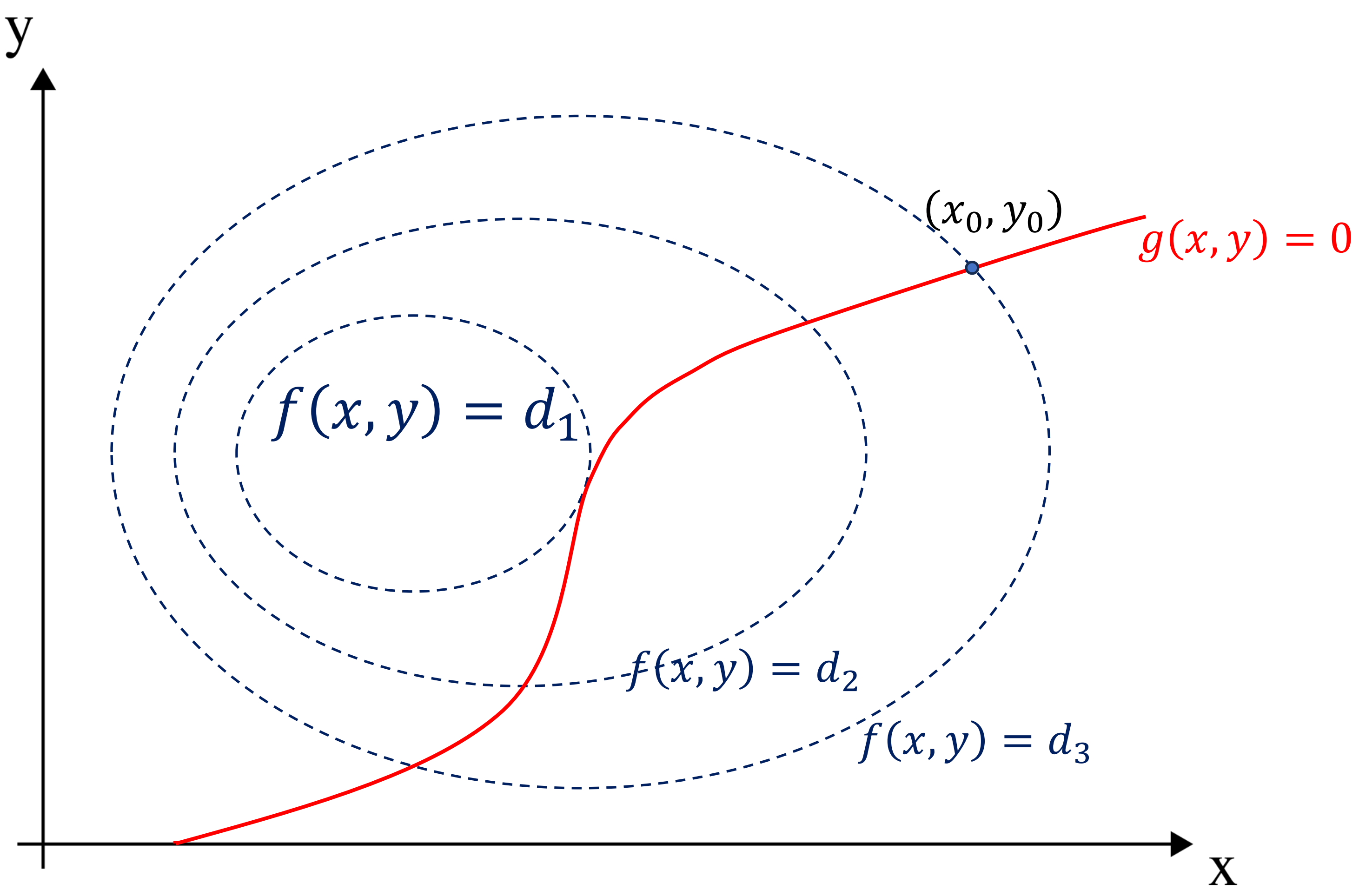
3. Найдем координаты условно-стационарных точек.

$$\begin{cases} 2x_1 + \lambda_1 = 0 \\ 2x_2 + \lambda_1 = 0 \\ x_1 + x_2 - 2 = 0 \end{cases} \Rightarrow \begin{cases} 2x_1 + \lambda_1 = 0 \\ x_1 - x_2 = 0 \\ x_1 + x_2 - 2 = 0 \end{cases} \Rightarrow \begin{cases} x_1^* = 1 \\ x_2^* = 1 \\ \lambda_1^* = -2 \end{cases}$$

Метод множителей Лагранжа

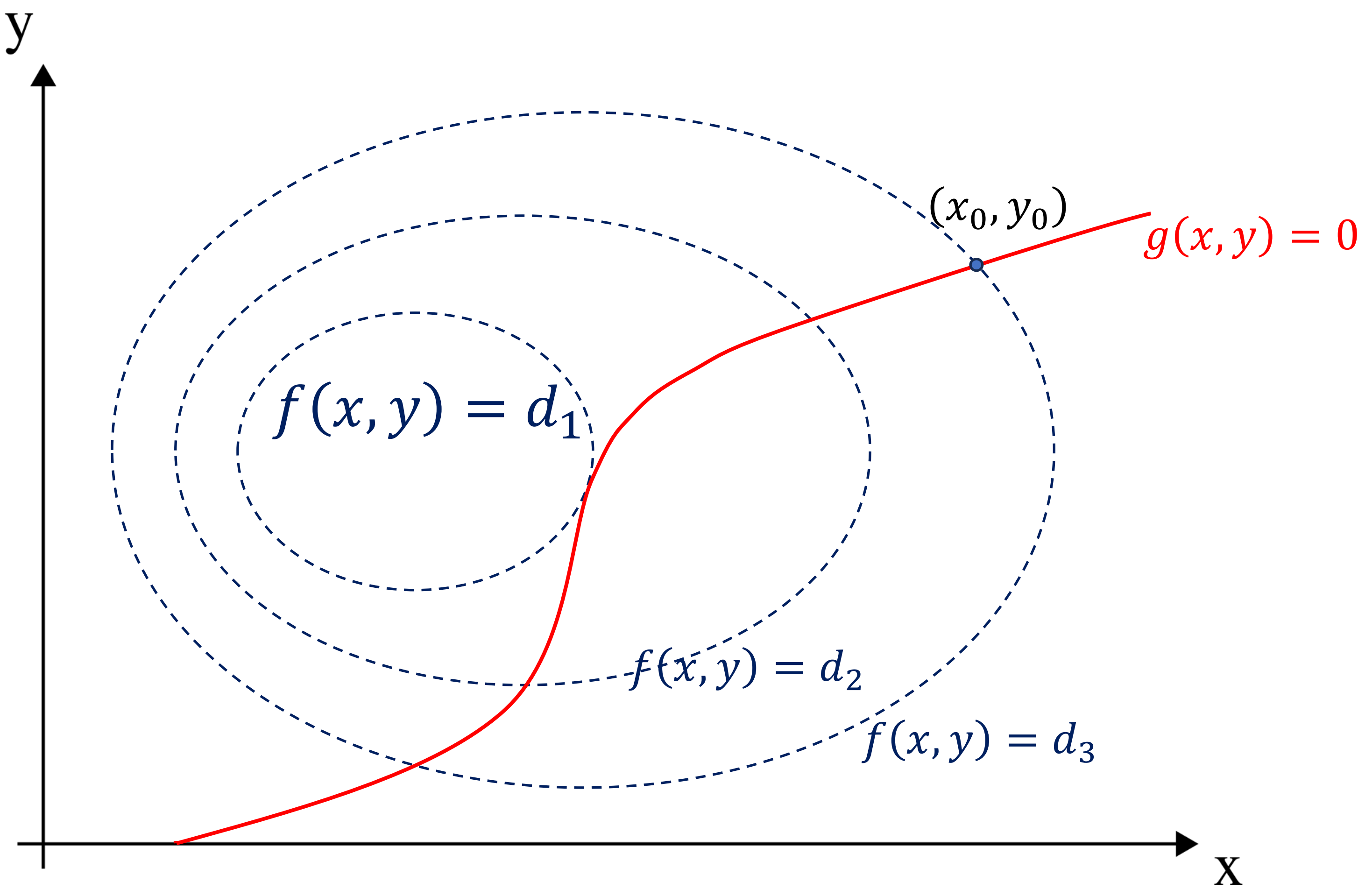


Метод множителей Лагранжа



Если в точке (x_0, y_0) кривая $g(x, y) = 0$ пересекает линию уровня $f(x, y)$ под ненулевым углом:

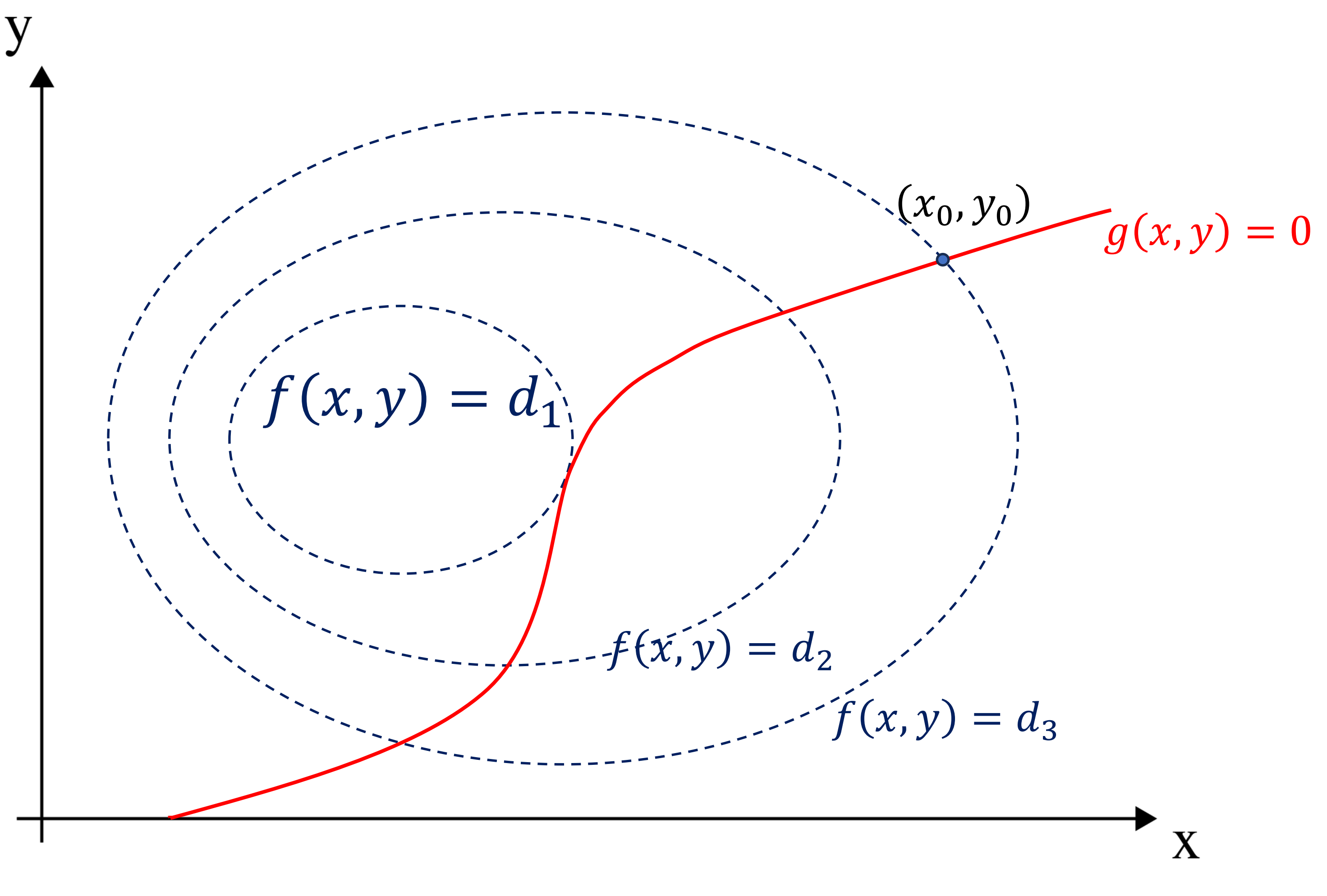
Метод множителей Лагранжа



Если в точке (x_0, y_0) кривая $g(x,y) = 0$ пересекает линию уровня $f(x,y)$ под ненулевым углом:

- в одну сторону вдоль $g(x,y) = 0$ от точки (x_0, y_0) функция $f(x,y)$ возрастает
- в другую сторону вдоль $g(x,y) = 0$ от точки (x_0, y_0) функция $f(x,y)$ убывает

Метод множителей Лагранжа



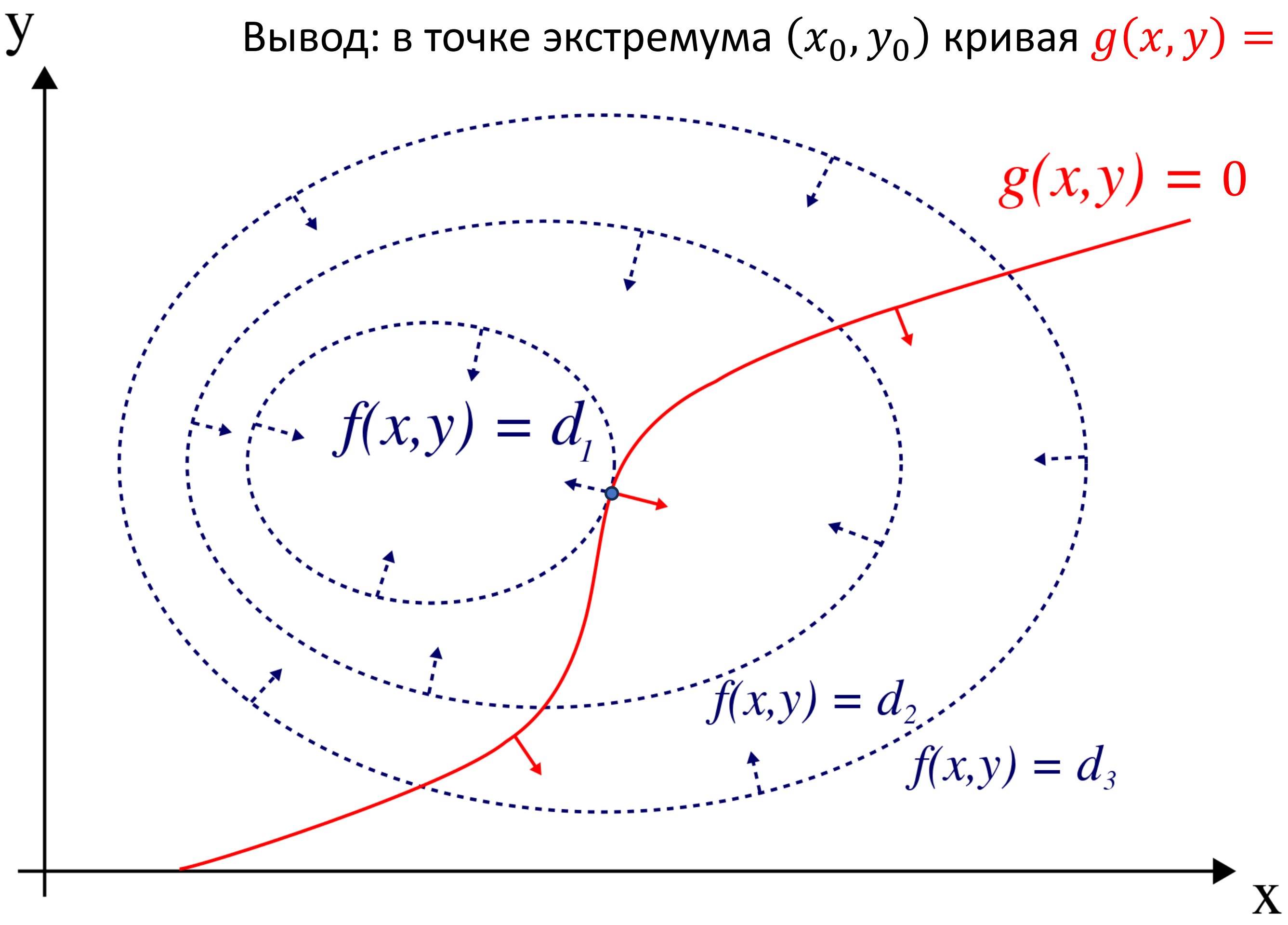
Если в точке (x_0, y_0) кривая $g(x,y) = 0$ пересекает линию уровня $f(x,y)$ под ненулевым углом:

- в одну сторону вдоль $g(x,y) = 0$ от точки (x_0, y_0) функция $f(x,y)$ возрастает
- в другую сторону вдоль $g(x,y) = 0$ от точки (x_0, y_0) функция $f(x,y)$ убывает

Значит (x_0, y_0) - не точка экстремума

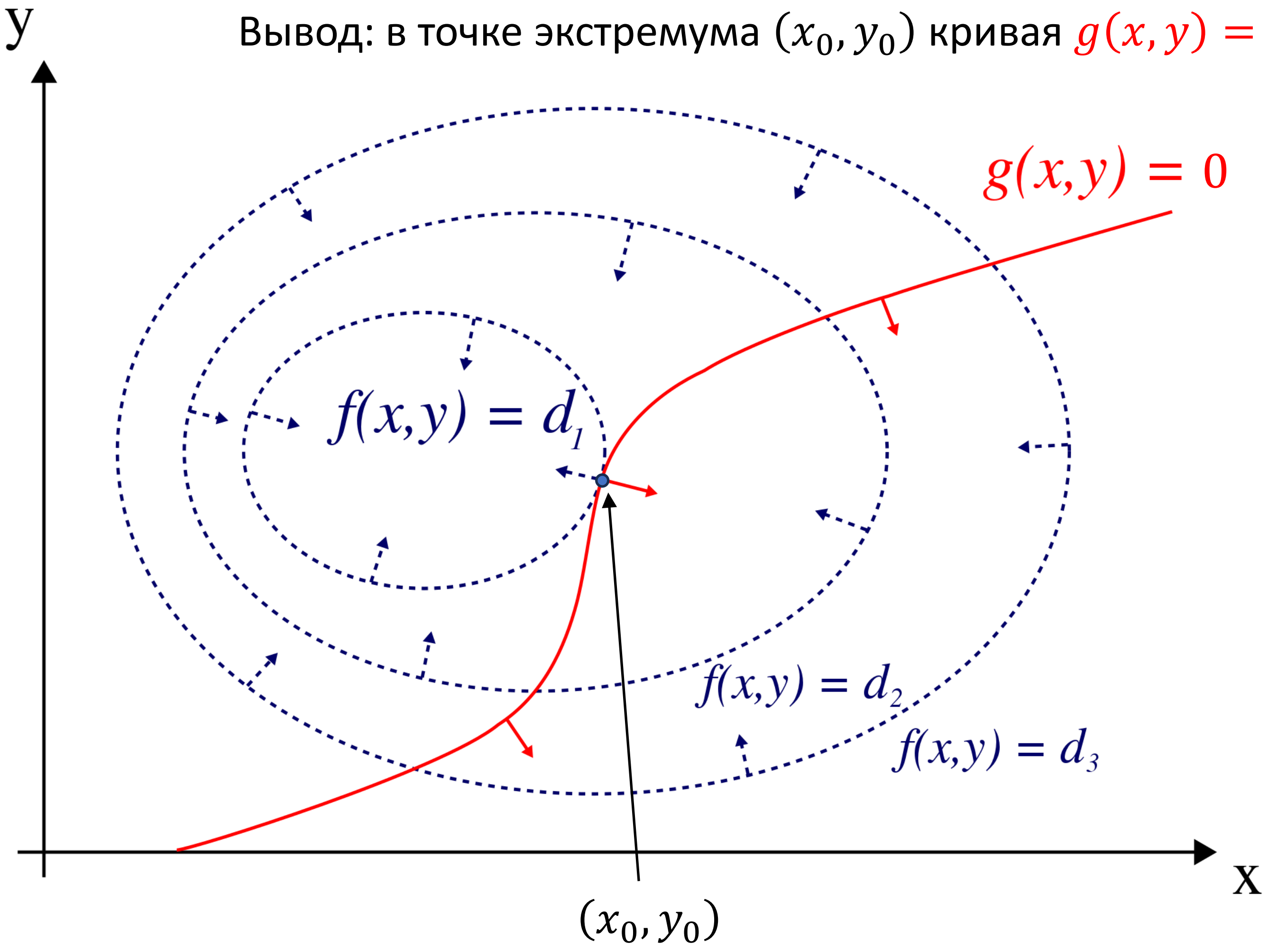
Метод множителей Лагранжа

Вывод: в точке экстремума (x_0, y_0) кривая $g(x, y) = 0$ касается линии уровня $f(x, y)$

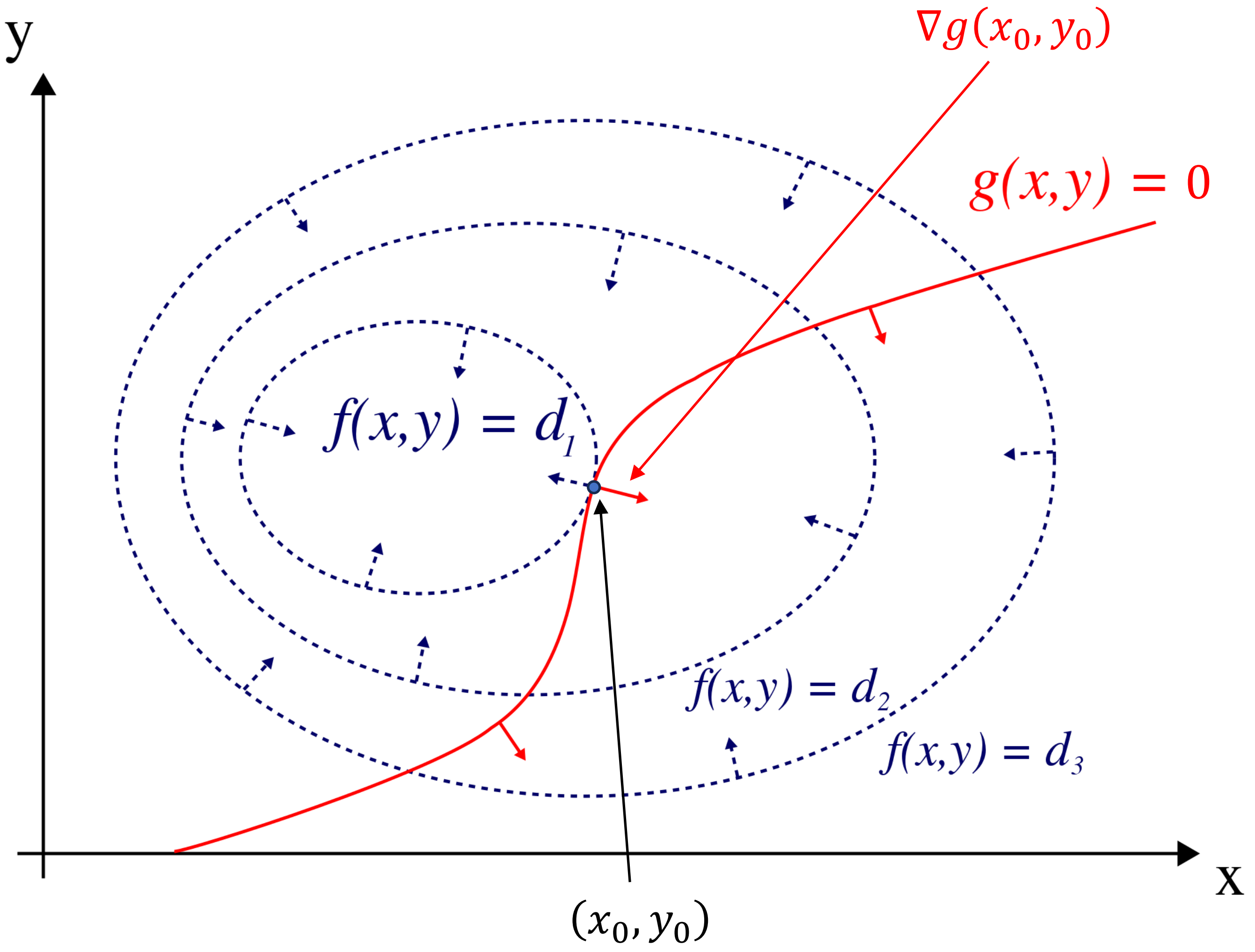


Метод множителей Лагранжа

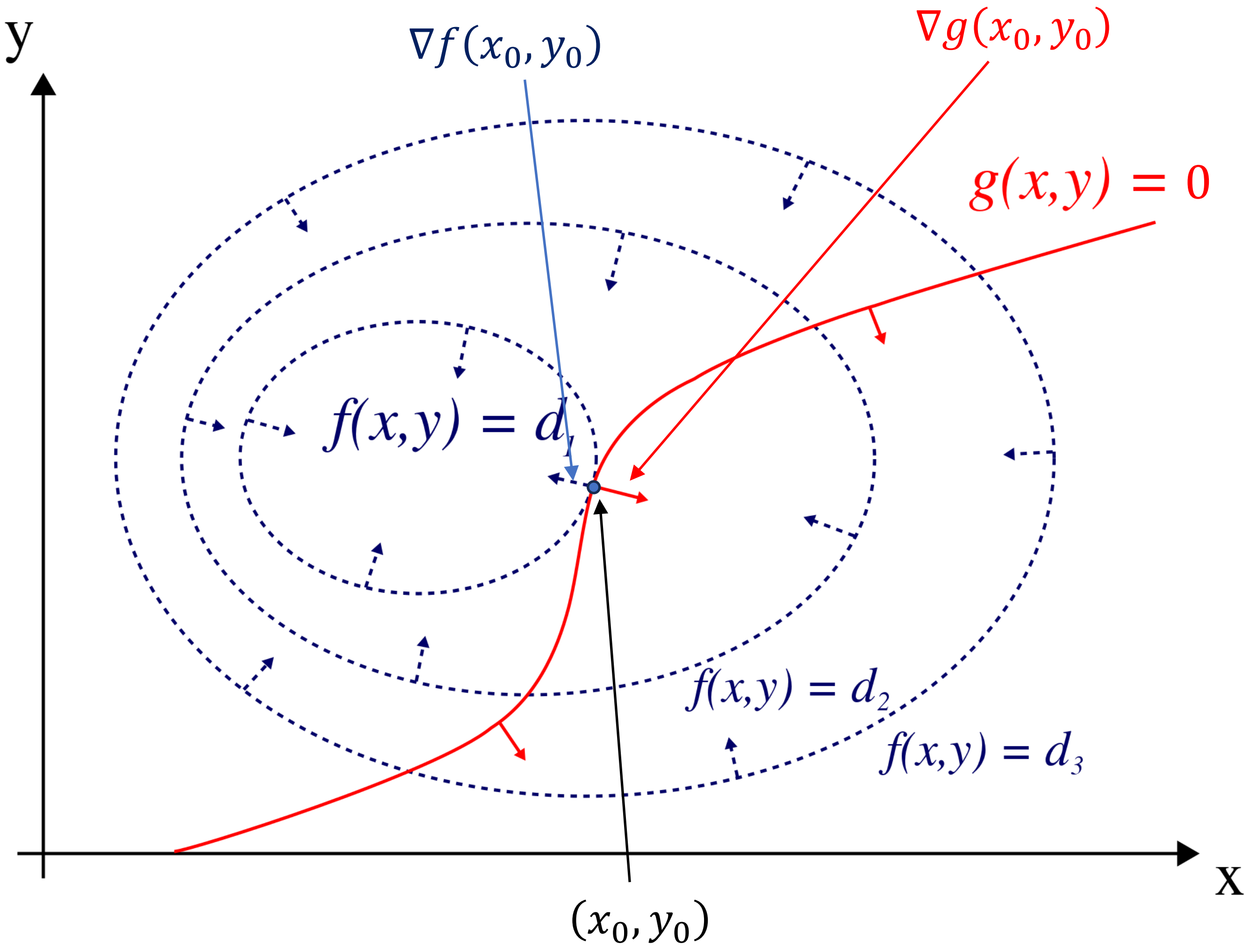
Вывод: в точке экстремума (x_0, y_0) кривая $g(x,y) = 0$ касается линии уровня $f(x,y)$



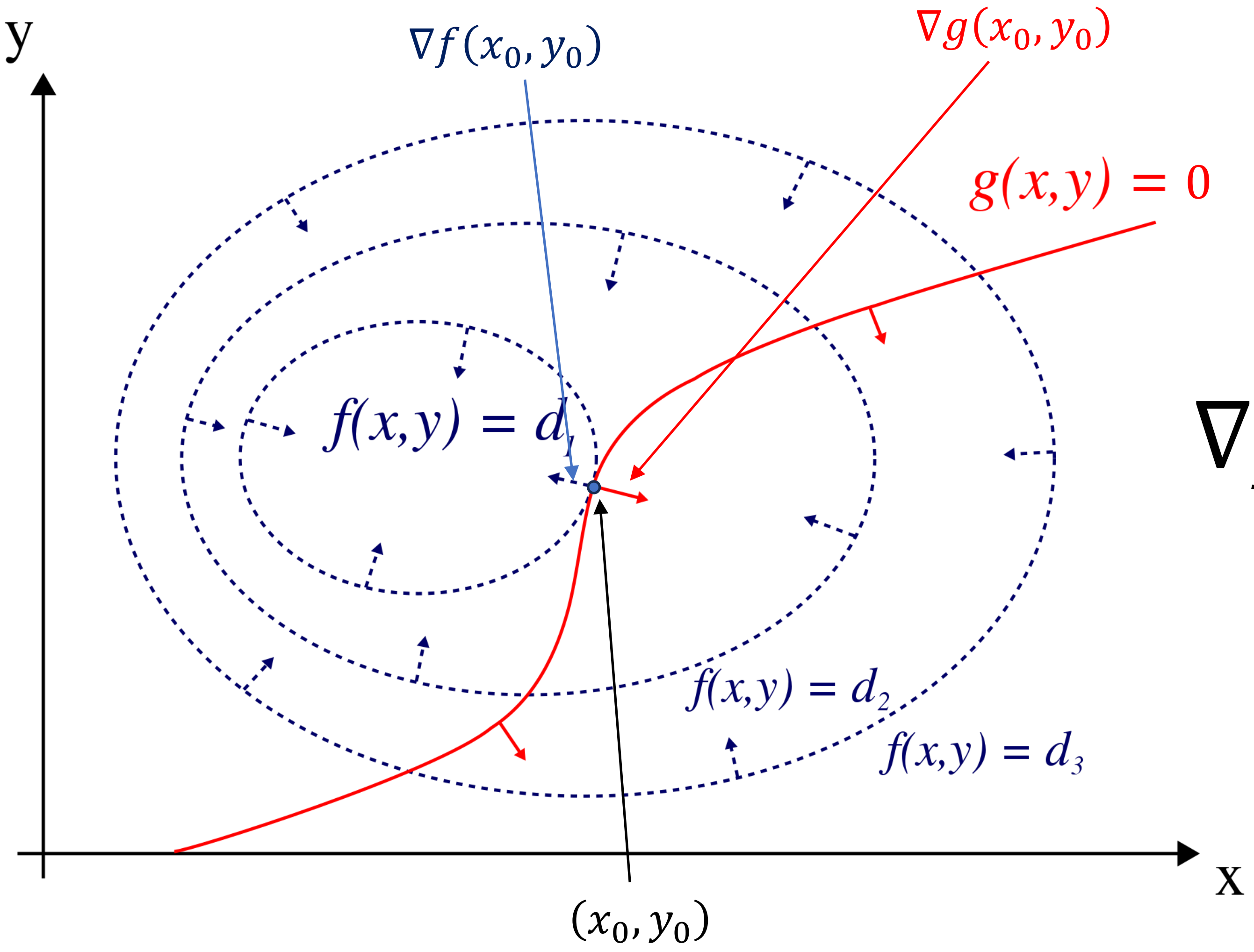
Метод множителей Лагранжа



Метод множителей Лагранжа

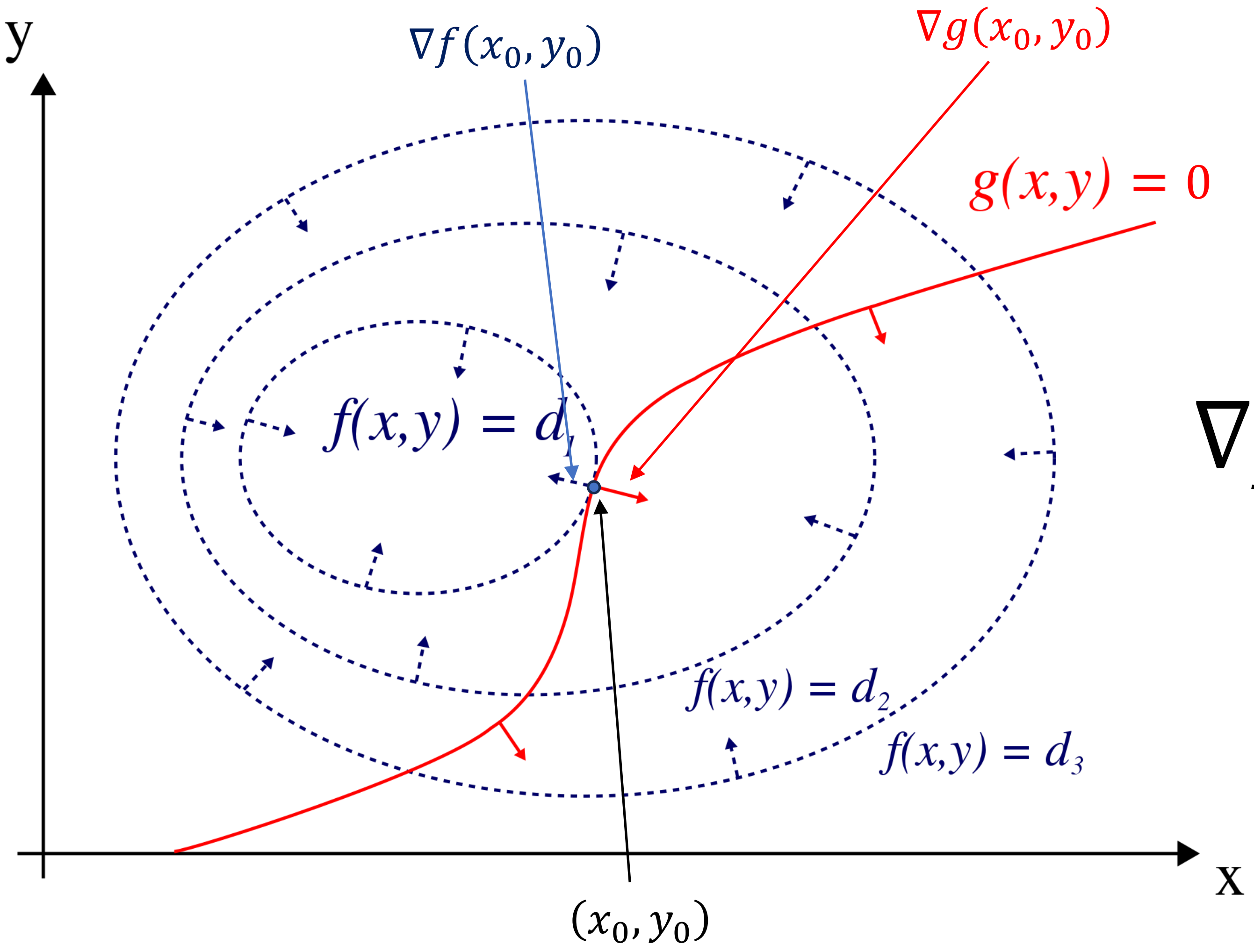


Метод множителей Лагранжа



$$\nabla f(x,y) + \lambda \nabla g(x,y) = 0$$

Метод множителей Лагранжа



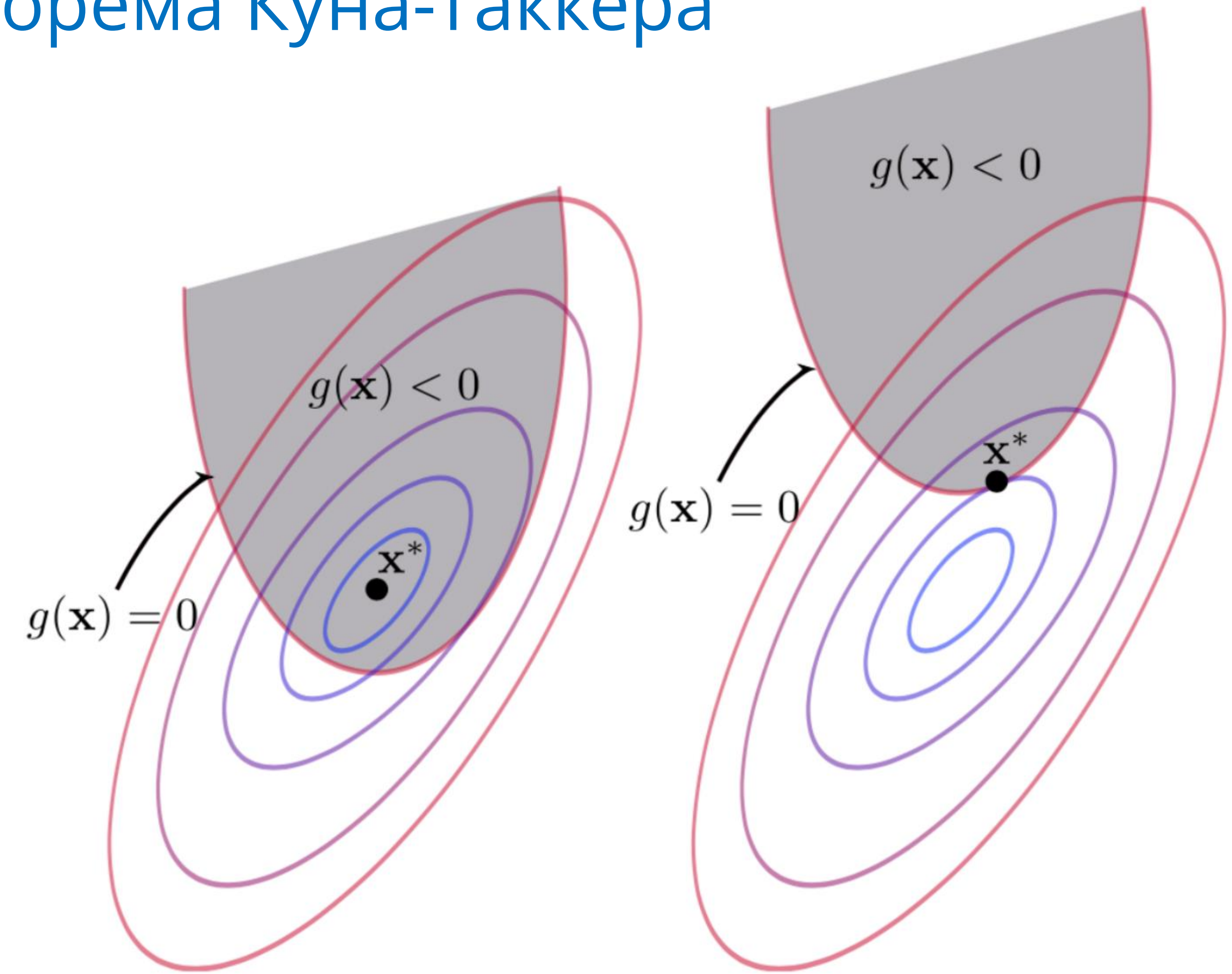
$$\nabla f(x,y) + \lambda \nabla g(x,y) = 0$$

$$\boxed{\begin{aligned} &\nabla L(x,y) = 0 \\ &L(x,y) = f(x,y) + \lambda g(x,y) \end{aligned}}$$

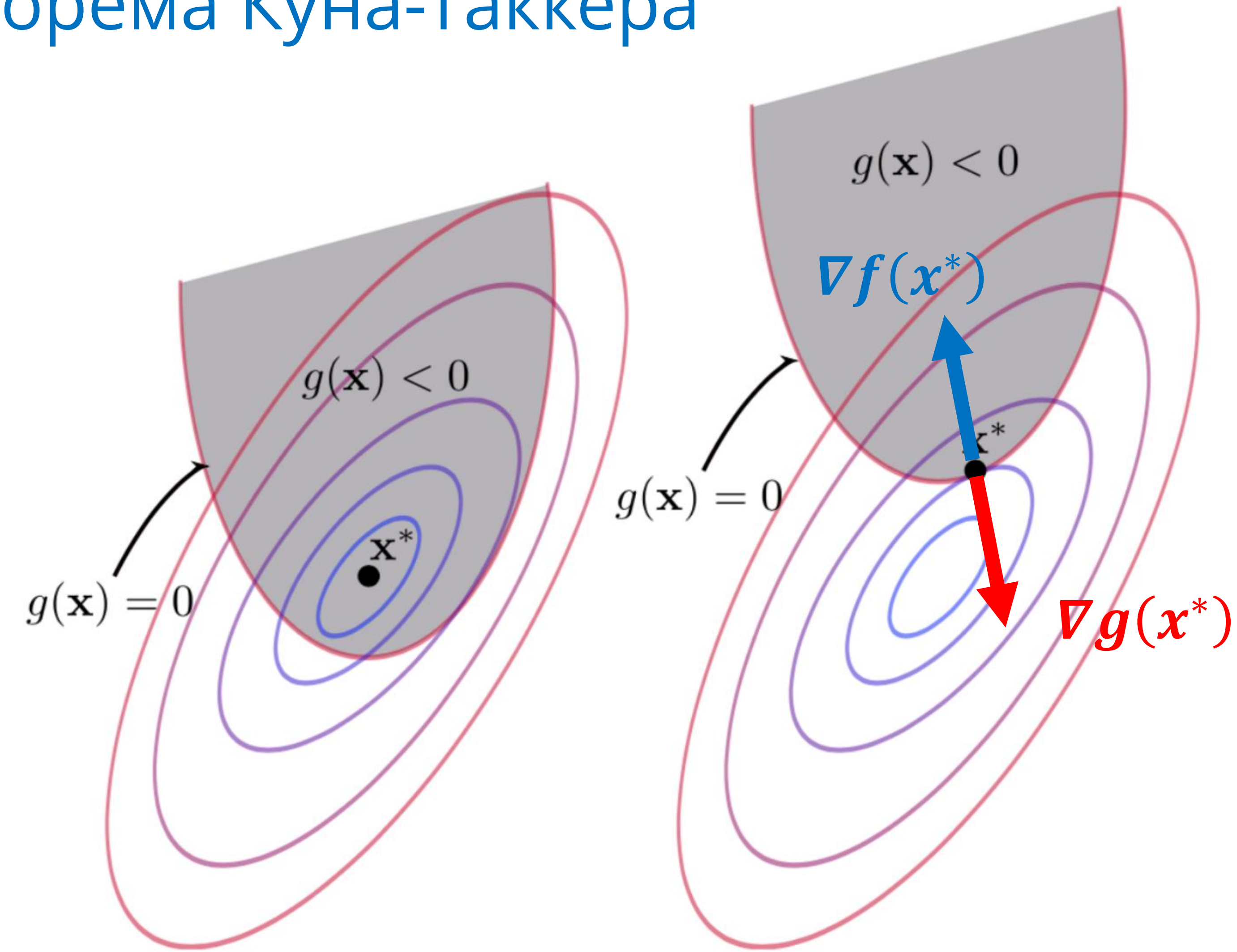
Теорема Куна-Таккера

$$\begin{cases} f(x) \to \min\limits_{x}; \\ g(x) \leq 0 \end{cases} \implies \begin{cases} \nabla_x L(x, \mu) = \nabla_x (f(x) + \mu g(x)) = 0 \\ \mu g(x) = 0 \\ \mu \geq 0 \end{cases}$$

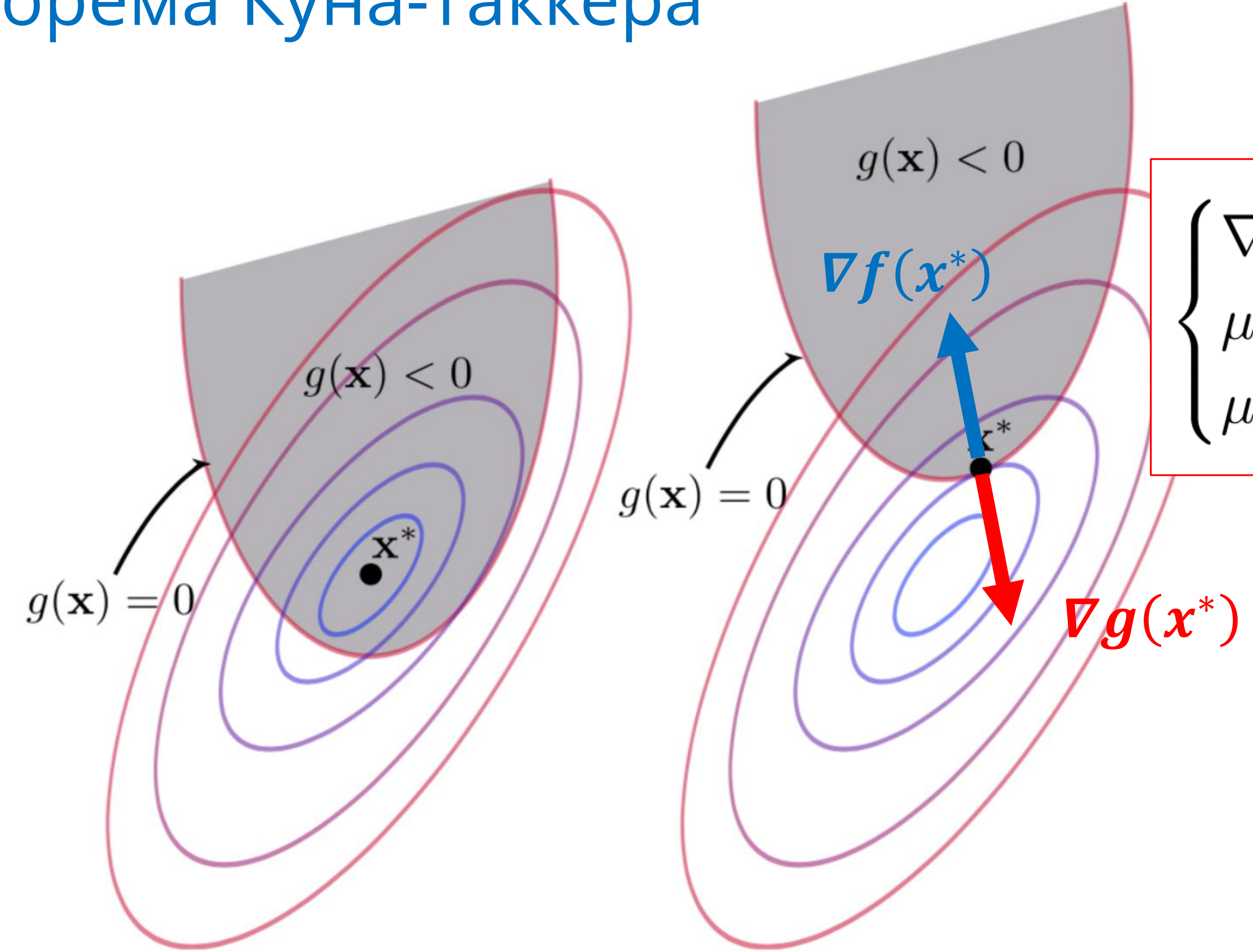
Теорема Куна-Таккера



Теорема Куна-Таккера



Теорема Куна-Таккера



$$\begin{cases} \nabla_x L(x,\mu) = \nabla_x(f(x) + \mu g(x)) = 0 \\ \mu g(x) = 0 \\ \mu \geq 0 \end{cases}$$

Теорема Куна-Таккера: пример 1 - регуляризация

$$\begin{cases} \tilde{Q} = \sum_{i=1}^{l} L(M_i) \to min \\ \sum_{n=1}^{d} {w_n}^2 \leq \tau \end{cases} \Longrightarrow \sum_{i=1}^{l} L(M_i) + \gamma \sum_{n=1}^{d} {w_n}^2 \to min$$

Теорема Куна-Таккера: пример 2 - SVM

$$\begin{cases} \frac{1}{2} < w, w > +C \sum\limits_{i=1}^{l} \xi_i \to min_{w,w_0,\xi}; \\ y_i(< w, x_i > -w_0) \geq 1 - \xi_i, & i = 1, ..., l; \\ \xi_i \geq 0, & i = 1, ..., l \end{cases}$$

Не забываем, что:

$$M_i = y_i(< w, x_i - w_0 >)$$

Выпишем лагранжиан и подставим в него выражение для отступа:

$$\mathcal{L}(w, w_0, \xi; \lambda, \eta) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\lambda_i(M_i(w, w_0) - 1 + \xi_i) - \sum_{i=1}^{l}\xi_i\eta_i =$$

$$= \frac{1}{2}||w||^2 - \sum_{i=1}^{l}\lambda_i(M_i(w, w_0) - 1) - \sum_{i=1}^{l}\xi_i(\lambda_i + \eta_i - C)$$

Теорема Куна-Таккера: пример 2 - SVM

$$\mathcal{L}(w, w_0, \xi; \lambda, \eta) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\lambda_i(M_i(w, w_0) - 1 + \xi_i) - \sum_{i=1}^{l}\xi_i\eta_i =$$

$$= \frac{1}{2}||w||^2 - \sum_{i=1}^{l}\lambda_i(M_i(w, w_0) - 1) - \sum_{i=1}^{l}\xi_i(\lambda_i + \eta_i - C)$$

$$\begin{cases} \nabla\mathcal{L}(w, w_0, \xi; \lambda, \eta) = 0; \\ \xi_i, \lambda, \eta_i \geq 0, i = 1, ..., l; \\ \lambda_i = 0, \text{либо } M_i(w, w_0) = 1 - \xi_i, i = 1, ..., l; \\ \eta_i = 0, \text{либо } \xi_i = 0, i = 1, ..., l \end{cases}$$

Теорема Куна-Таккера: пример 2 - SVM

$$\boxed{\begin{aligned}\mathcal{L}(w, w_0, \xi; \lambda, \eta) &= \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\lambda_i(M_i(w, w_0) - 1 + \xi_i) - \sum_{i=1}^{l}\xi_i\eta_i = \\ &= \frac{1}{2}||w||^2 - \sum_{i=1}^{l}\lambda_i(M_i(w, w_0) - 1) - \sum_{i=1}^{l}\xi_i(\lambda_i + \eta_i - C)\end{aligned}}$$

$$\begin{cases}\nabla\mathcal{L}(w, w_0, \xi; \lambda, \eta) = 0; \\ \xi_i, \lambda, \eta_i \geq 0, i = 1, ..., l; \\ \lambda_i = 0, \text{либо } M_i(w, w_0) = 1 - \xi_i, i = 1, ..., l; \\ \eta_i = 0, \text{либо } \xi_i = 0, i = 1, ..., l\end{cases}$$

$$\frac{\partial\mathcal{L}}{\partial w} = w - \sum_{i=1}^{l}\lambda_i y_i x_i = 0 \Longrightarrow w = \sum_{i=1}^{l}\lambda_i y_i x_i$$

Теорема Куна-Таккера: пример 2 - SVM

$$\mathcal{L}(w, w_0, \xi; \lambda, \eta) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\lambda_i(M_i(w, w_0) - 1 + \xi_i) - \sum_{i=1}^{l}\xi_i\eta_i =$$

$$= \frac{1}{2}||w||^2 - \sum_{i=1}^{l}\lambda_i(M_i(w, w_0) - 1) - \sum_{i=1}^{l}\xi_i(\lambda_i + \eta_i - C)$$

$$\begin{cases} \nabla\mathcal{L}(w, w_0, \xi; \lambda, \eta) = 0; \\ \xi_i, \lambda, \eta_i \geq 0, i = 1, ..., l; \\ \lambda_i = 0, \text{либо } M_i(w, w_0) = 1 - \xi_i, i = 1, ..., l; \\ \eta_i = 0, \text{либо } \xi_i = 0, i = 1, ..., l \end{cases}$$

$$\frac{\partial\mathcal{L}}{\partial w} = w - \sum_{i=1}^{l}\lambda_i y_i x_i = 0 \Longrightarrow w = \sum_{i=1}^{l}\lambda_i y_i x_i$$

$$\frac{\partial\mathcal{L}}{\partial\xi_i} = -\lambda_i - \eta_i + C = 0 \Longrightarrow \eta_i + \lambda_i, i = 1, ..., l$$

Теорема Куна-Таккера: пример 2 - SVM

$$\mathcal{L}(w, w_0, \xi; \lambda, \eta) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\lambda_i(M_i(w, w_0) - 1 + \xi_i) - \sum_{i=1}^{l}\xi_i\eta_i =$$

$$= \frac{1}{2}||w||^2 - \sum_{i=1}^{l}\lambda_i(M_i(w, w_0) - 1) - \sum_{i=1}^{l}\xi_i(\lambda_i + \eta_i - C)$$

$$\begin{cases} \nabla\mathcal{L}(w, w_0, \xi; \lambda, \eta) = 0; \\ \xi_i, \lambda, \eta_i \geq 0, i = 1, ..., l; \\ \lambda_i = 0, \text{либо } M_i(w, w_0) = 1 - \xi_i, i = 1, ..., l; \\ \eta_i = 0, \text{либо } \xi_i = 0, i = 1, ..., l \end{cases}$$

$$\frac{\partial\mathcal{L}}{\partial w} = w - \sum_{i=1}^{l}\lambda_i y_i x_i = 0 \Longrightarrow w = \sum_{i=1}^{l}\lambda_i y_i x_i$$

$$\frac{\partial\mathcal{L}}{\partial\xi_i} = -\lambda_i - \eta_i + C = 0 \Longrightarrow \eta_i + \lambda_i, i = 1, ..., l$$

$$\Longrightarrow \begin{cases} \mathcal{L}(\lambda) = -\sum\limits_{i=1}^{l}\lambda_i + \frac{1}{2}\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{l}\lambda_i\lambda_j y_i y_j < x_i x_j > \to min_\lambda; \\ 0 \leq \lambda_i \leq C, i = 1, ..., l; \\ \sum\limits_{i=1} l\lambda_i y_i = 0. \end{cases}$$

Теорема Куна-Таккера: пример 2 - SVM

$$\mathcal{L}(w, w_0, \xi; \lambda, \eta) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\lambda_i(M_i(w, w_0) - 1 + \xi_i) - \sum_{i=1}^{l}\xi_i\eta_i =$$

$$= \frac{1}{2}||w||^2 - \sum_{i=1}^{l}\lambda_i(M_i(w, w_0) - 1) - \sum_{i=1}^{l}\xi_i(\lambda_i + \eta_i - C)$$

$$\begin{cases} \nabla\mathcal{L}(w, w_0, \xi; \lambda, \eta) = 0; \\ \xi_i, \lambda, \eta_i \geq 0, i = 1, ..., l; \\ \lambda_i = 0, \text{либо } M_i(w, w_0) = 1 - \xi_i, i = 1, ..., l; \\ \eta_i = 0, \text{либо } \xi_i = 0, i = 1, ..., l \end{cases}$$

$$\frac{\partial\mathcal{L}}{\partial w} = w - \sum_{i=1}^{l}\lambda_i y_i x_i = 0 \Longrightarrow w = \sum_{i=1}^{l}\lambda_i y_i x_i$$

$$\frac{\partial\mathcal{L}}{\partial \xi_i} = -\lambda_i - \eta_i + C = 0 \Longrightarrow \eta_i + \lambda_i, i = 1, ..., l$$

$$\Longrightarrow \begin{cases} \mathcal{L}(\lambda) = -\sum_{i=1}^{l}\lambda_i + \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\lambda_i\lambda_j y_i y_j < x_i x_j > \to min_\lambda; \\ 0 \leq \lambda_i \leq C, i = 1, ..., l; \\ \sum_{i=1} l\lambda_i y_i = 0. \end{cases}$$

Т.е. для настройки w достаточно знать скалярные произведения объектов из выборки (или значения ядра на объектах из выборки)

Теорема Куна-Таккера: пример 2 - SVM

$$\frac{\partial \mathcal{L}}{\partial w} = w - \sum_{i=1}^{l} \lambda_i y_i x_i = 0 \Longrightarrow w = \sum_{i=1}^{l} \lambda_i y_i x_i$$

$$\begin{cases} \mathcal{L}(\lambda) = -\sum\limits_{i=1}^{l} \lambda_i + \frac{1}{2} \sum\limits_{i=1}^{l} \sum\limits_{j=1}^{l} \lambda_i \lambda_j y_i y_j < x_i x_j > \to min_{\lambda}; \\ 0 \le \lambda_i \le C, i = 1, ..., l; \\ \sum\limits_{i=1} l \lambda_i y_i = 0. \end{cases}$$

Т.е. для настройки w достаточно знать скалярные произведения объектов из выборки (или значения ядра на объектах из выборки)

Формула для прогнозирования на новых объектах:

$$a(x) = sign\left(\sum_{i=1}^{l} \lambda_i y_i < x_i, x > -w_0\right)$$

с линейным ядром

$$a(x) = sign\left(\sum_{i=1}^{h} \lambda_i y_i K(x_i, x) - w_0\right)$$

с произвольным ядром

Теорема Куна-Таккера: пример 2 - SVM

$$\frac{\partial \mathcal{L}}{\partial w} = w - \sum_{i=1}^{l} \lambda_i y_i x_i = 0 \Longrightarrow w = \sum_{i=1}^{l} \lambda_i y_i x_i$$

$$\begin{cases} \mathcal{L}(\lambda) = -\sum\limits_{i=1}^{l} \lambda_i + \frac{1}{2} \sum\limits_{i=1}^{l} \sum\limits_{j=1}^{l} \lambda_i \lambda_j y_i y_j < x_i x_j > \to min_\lambda; \\ 0 \leq \lambda_i \leq C, i = 1, ..., l; \\ \sum\limits_{i=1} l \lambda_i y_i = 0. \end{cases}$$

Т.е. для настройки w достаточно знать скалярные произведения объектов из выборки (или значения ядра на объектах из выборки)

Формула для прогнозирования на новых объектах:

$$a(x) = sign\left(\sum_{i=1}^{l} \lambda_i y_i < x_i, x > -w_0\right)$$

с линейным ядром

$$a(x) = sign\left(\sum_{i=1}^{h} \lambda_i y_i K(x_i, x) - w_0\right)$$

с произвольным ядром

Векторы x_i, коэффициент λ_i перед которыми **не равен нулю**, называются **опорными векторами**, т.к. оптимальные веса линейного классификатора зависят **только от этих объектов выборки**

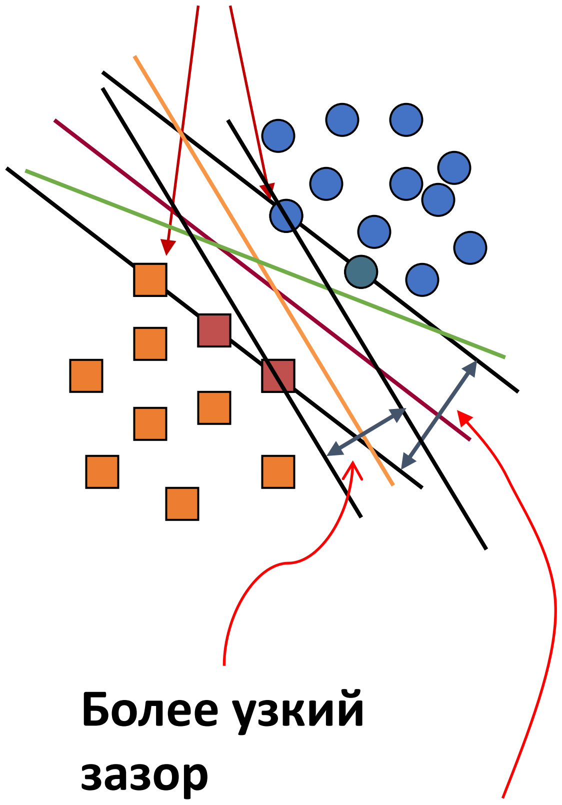
Метод опорных векторов (SVM): ключевые особенности

1 **SVM максимизирует** отступ от разделяющей гиперплоскости

2 **Дискриминантная функция** полностью **задается** подмножеством объектов, называемым **опорными векторами** (следует из двойственной задачи)

3 **Обучать SVM** можно, решая **задачу квадратичного программирования** (двойственная задача) либо «в лоб» решая задачу безусловной оптимизации

4 В SVM можно эффективно заменять скалярное произведение **нелинейными ядрами** и строить **нелинейные разделяющие поверхности**, пользуясь тем, что для обучения достаточно знать скалярные произведения векторов признаков **объектов из выборки** (следует из двойственной задачи)



Сегодня

- Метод опорных векторов
- Ядра (Kernel trick) в методе опорных векторов
- Математическое дополнение: условный экстремум