

Лекція 4

Лінійні моделі

§16 Лінійні моделі в задачах регресії

Позначення

- X — простір об'єктів
- Y — простір відповідей
- $x = (x^1, \dots, x^d)$ — ознаковий опис
- $X = (x_i, y_i)_{i=1}^l$ — навчальна вибірка

Задача регресії

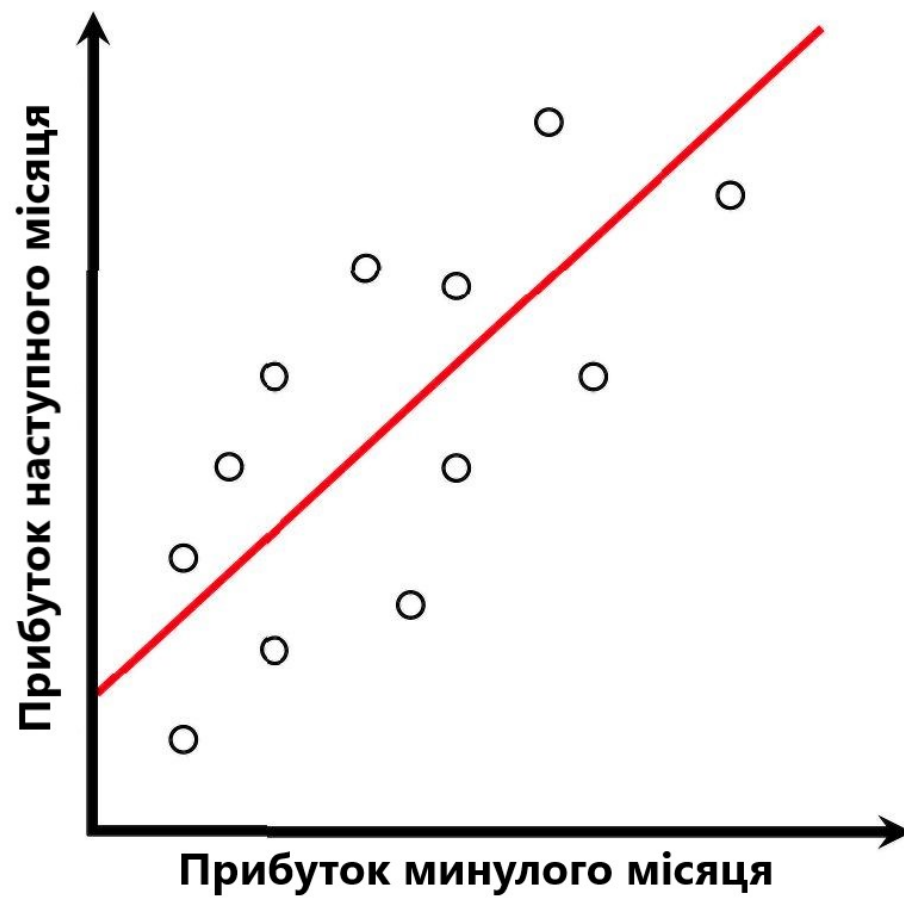
В задачі регресії простір відповідей $Y = \mathbb{R}$.

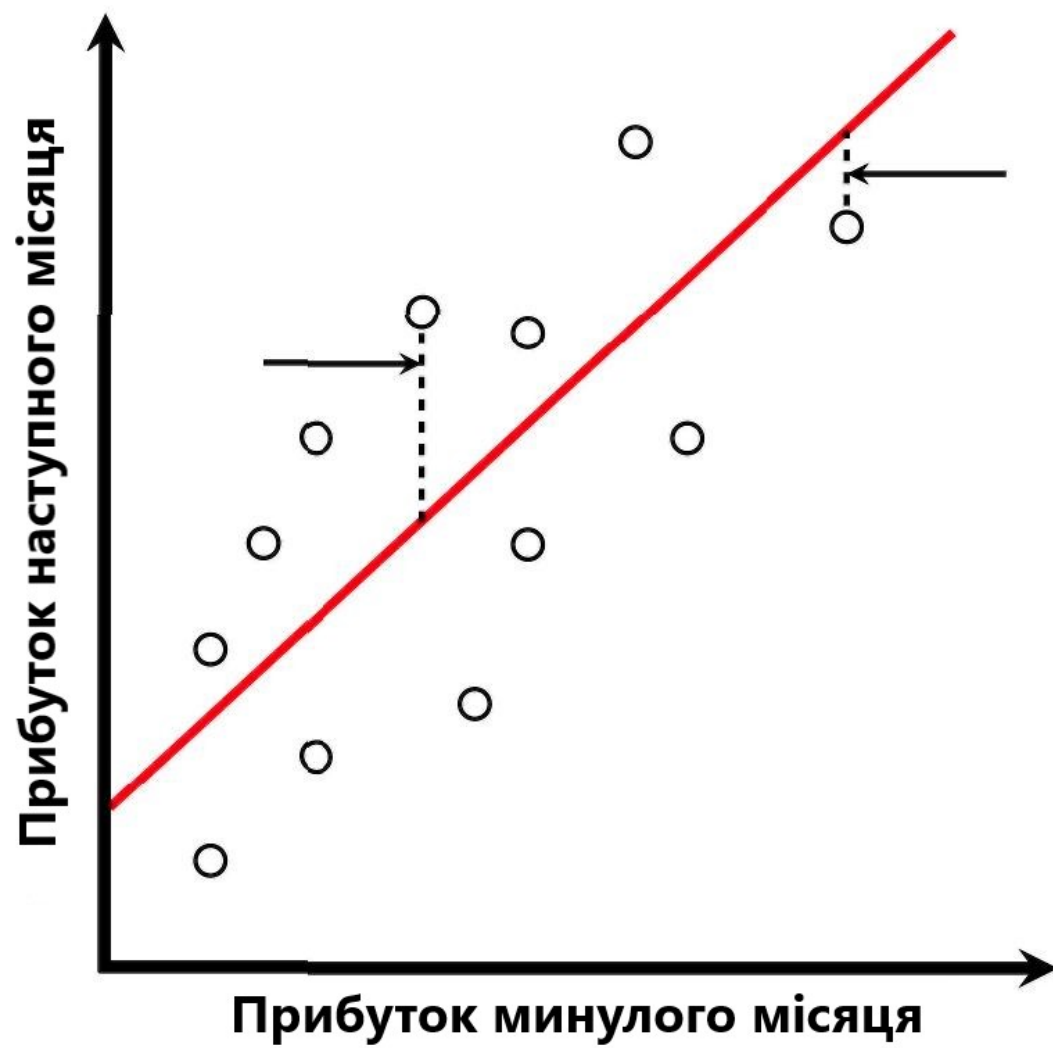
Щоб навчитися вирішувати задачу регресії, необхідно задати:

- $a(x)$ — алгоритм, модель
- $Q(a, X)$ — функціонал похибки алгоритму a на вибірці X
- Метод навчання: $a(x) = \arg \min_{a \in A} Q(a, X)$

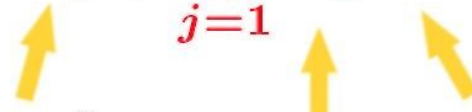
Приклад задачі регресії: прибутку магазину







Опис лінійної моделі

$$a(x) = w_0 + \sum_{j=1}^d w_j x^j$$


Вільний коефіцієнт

Ваги

Ознаки

Додамо константну ознаку

$$a(x) = \sum_{i=1}^{d+1} w_i x^i = \langle w, x \rangle$$

Опис лінійної моделі: функціонал похибки

Відхилення від прогнозу: $a(x) - y$

$a(x)$	y	відхилення
11	10	1
9	10	-1
20	10	10
1	10	-9

Як міру похибки відхилення не можна обирати, тому що в цьому випадку **мінімум функціонала не буде досягатися при правильній відповіді** $a(x) = y$ (мінімум буде, коли $a(x) - y \rightarrow -\infty$).

Відхилення від прогнозу: модуль відхилення: $|a(x) - y|$

Не є гладкою функцією, і для оптимізації такого функціонала незручно використовувати градієнтні методи

Відхилення від прогнозу:

середньоквадратична похибка алгоритму (англ. mean squared error, MSE):

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

У випадку лінійної моделі його можна переписати у вигляді функції (оскільки тепер Q залежить від вектора, а не від функції) помилок:

$$Q(\mathbf{w}, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, x_i \rangle - y_i)^2$$



Вектор дійсних чисел

§17 Навчання лінійної регресії

Як навчати модель лінійної регресії, тобто як знайти її параметри?

Нами уведено наступний вираз для якості лінійної моделі на навчальній вибірці

$$Q(\mathbf{w}, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle \mathbf{w}, x_i \rangle - y_i)^2 \rightarrow \min_{\mathbf{w}}$$

- d невідомих
- є константна ознака
- опукла функція

Перехід до матричної форми запису

Матриця «об'єкти-ознаки»:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{\ell 1} & \cdots & x_{\ell d} \end{pmatrix} \quad \text{Об'єкт}$$

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{\ell 1} & \cdots & x_{\ell d} \end{pmatrix}$$

Ознака

Вектор відповідей

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_\ell \end{pmatrix}$$

Середньоквадратична похибка може бути переписана в матричному вигляді

$$Q(w, X) = \frac{1}{\ell} \|X w - y\|^2 \rightarrow \min_w$$

Аналітичний розв'язок

$$\mathbf{w}_* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

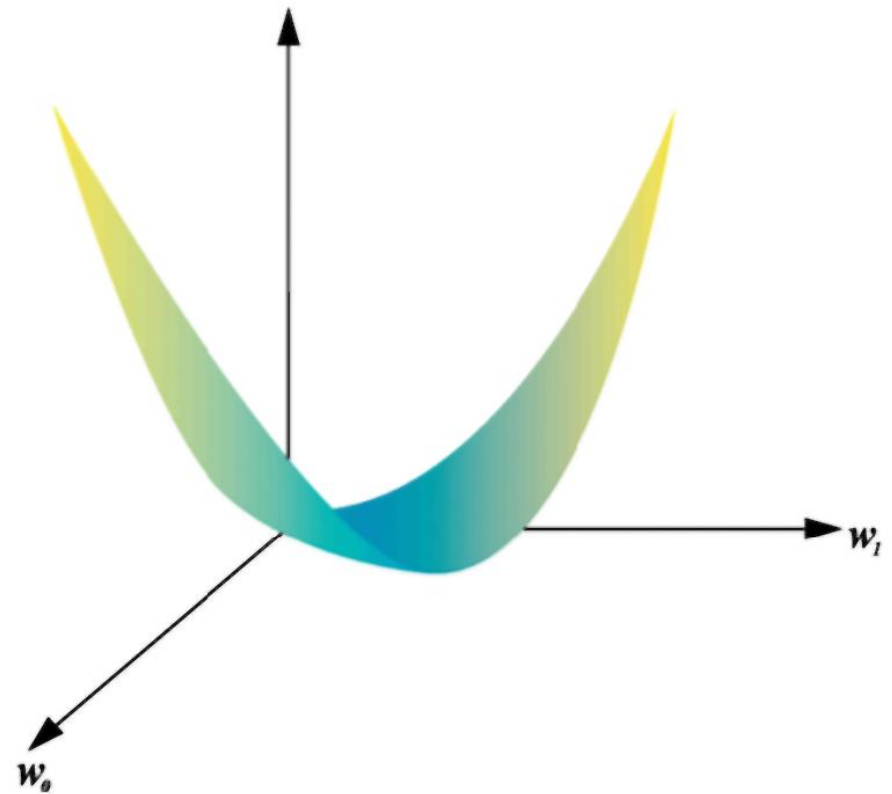
Основні складності:

- Для знаходження вирішення необхідно **обчислювати зворотну матрицю**. Операція обернення матриці вимагає, у випадку d ознак, виконання порядку d^3 операцій, і є чисельно складною вже в задачах з десятком ознак.
- Чисельний спосіб знаходження зворотної матриці не може бути застосований у деяких випадках (коли **матриця погано обумовлена**).

Оптимізаційний підхід

Використовувати чисельні методи оптимізації.

Нескладно показати, що середньоквадратична похибка – це **опукла й гладка функція**. Опуклість гарантує існування лише **одного мінімуму**, а гладкість - **існування вектора градієнта** в кожній точці. Це дозволяє використовувати **метод градієнтного спуска**.



Для використання методу градієнтного спуска необхідно вказати початкове наближення. Є багато підходів до того, найпростіший:

$$w^0 = 0.$$

На кожній наступній ітерації, $t = 1, 2, 3, \dots$, з наближення, отриманого в попередній ітерації w^{t-1} , віднімається вектор градієнта у відповідній точці w^{t-1} , помножений на деякий коефіцієнт η_t , який називається кроком:

$$w^t = w^{t-1} - \eta_t \nabla Q(w^{t-1}, X)$$

Зупинити ітерації потрібно, коли настає збіжність:

$$\| w^t - w^{t-1} \| < \varepsilon$$

§18 Градієнтний спуск для лінійної регресії

Випадок парної регресії

У випадку парної регресії ознака всього одна, а лінійна модель виглядає так:

$$a(x) = w_1 x + w_0$$

де w_1 і w_0 — два параметри.

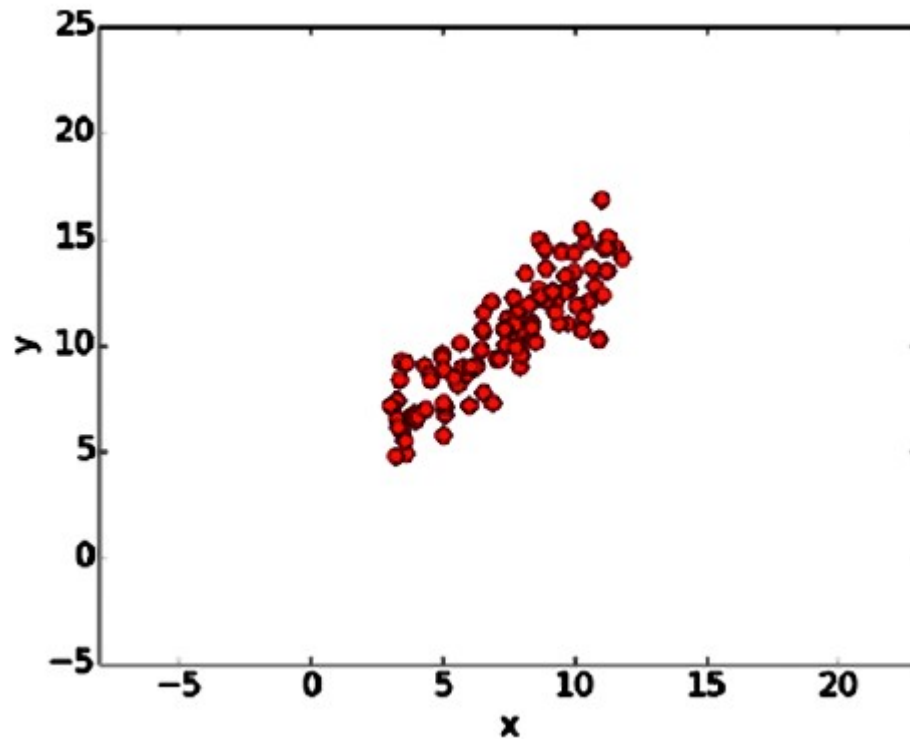
Середньоквадратична похибка приймає вигляд:

$$Q(w_0, w_1, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (w_1 x_i + w_0 - y_i)^2$$

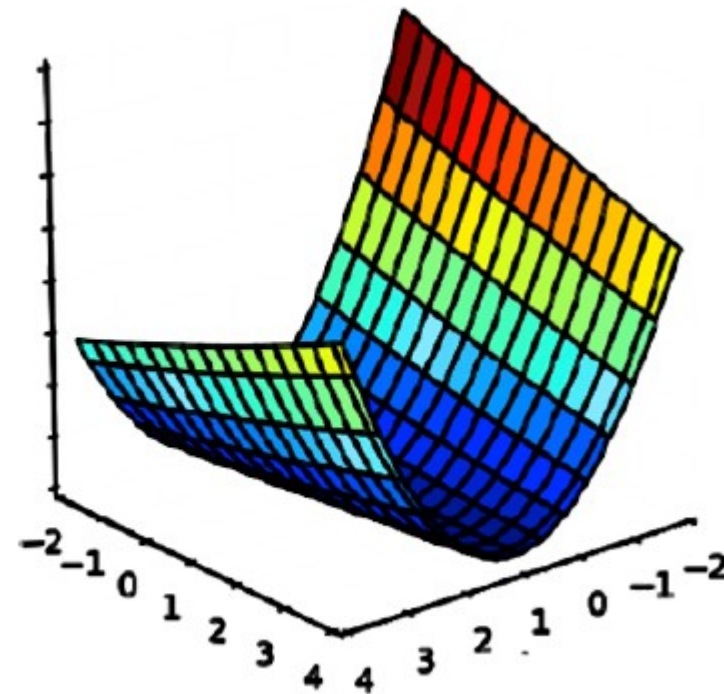
Для знаходження оптимальних параметрів буде застосовуватися метод градієнтного спуску. Щоб це зробити, необхідно спочатку обчислити частинні похідні функції похибки:

$$\frac{\partial Q}{\partial w_1} = \frac{2}{l} \sum_{i=1}^l (w_1 x_i + w_0 - y_i) x_i,$$

$$\frac{\partial Q}{\partial w_0} = \frac{2}{l} \sum_{i=1}^l (w_1 x_i + w_0 - y_i) \cdot 1.$$

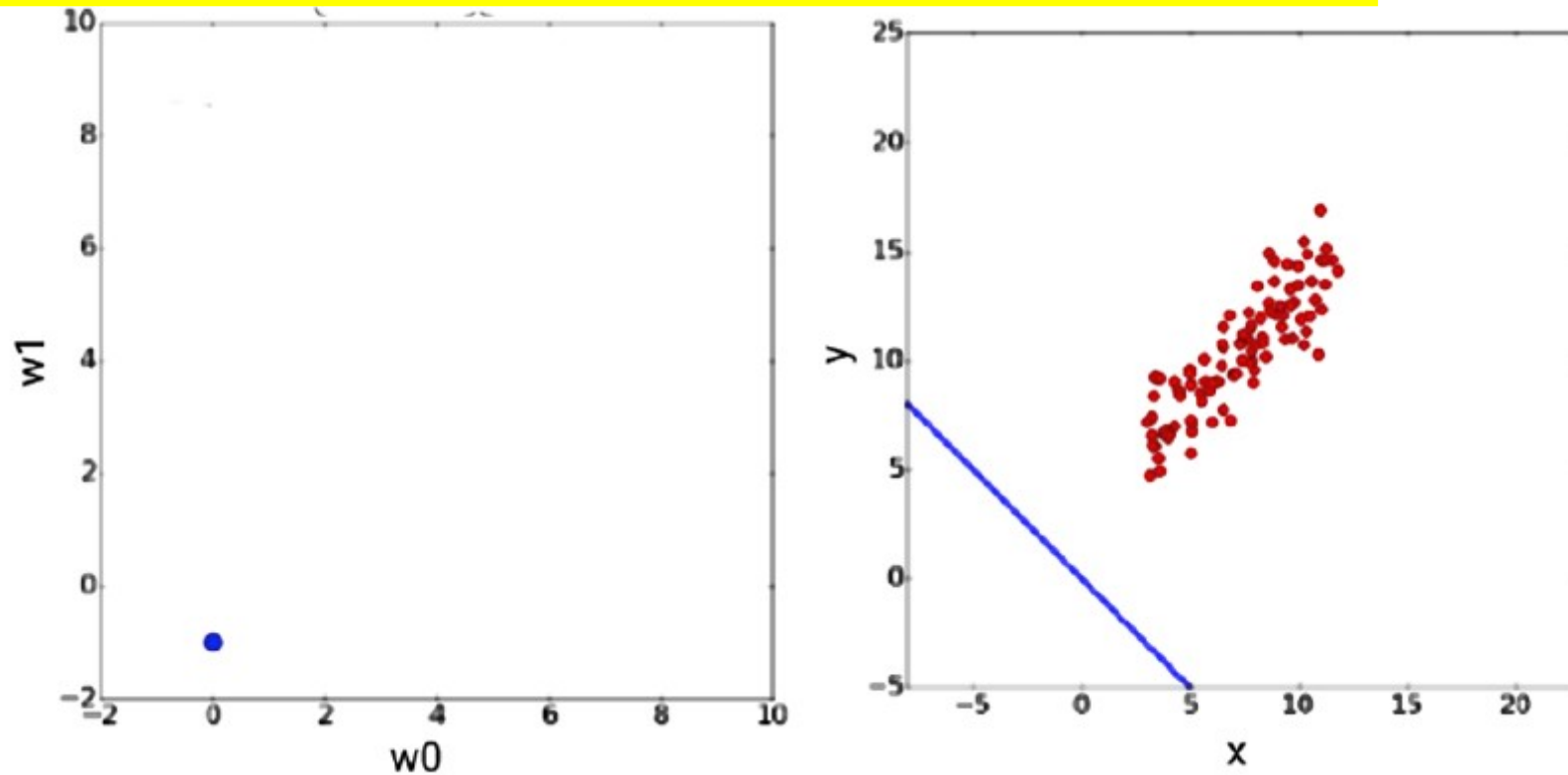


Вибірка



Функціонал якості

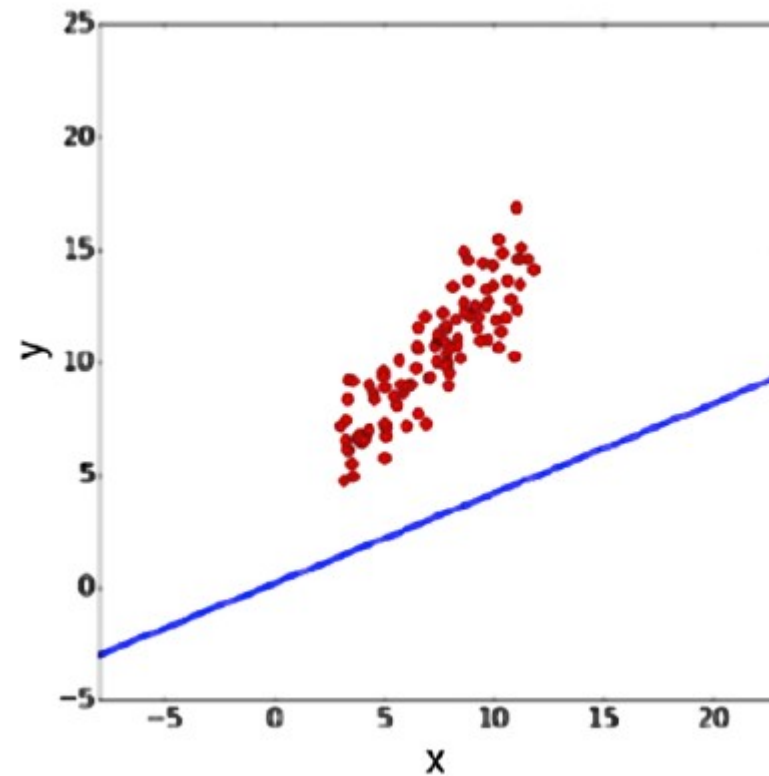
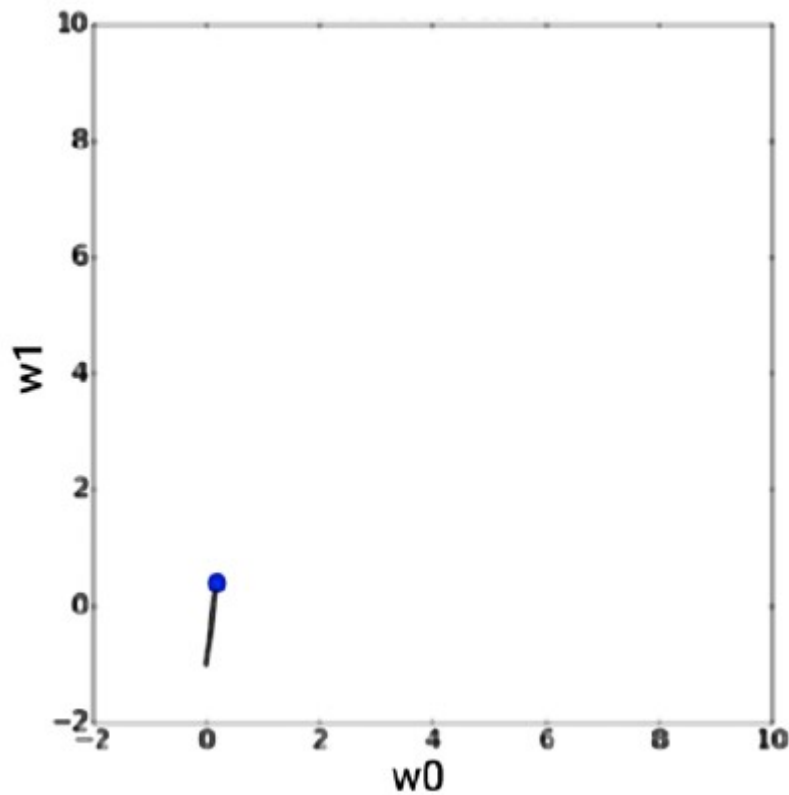
Демонстрація градієнтного спуска у випадку парної регресії



$$w^0 = (0, -1)$$

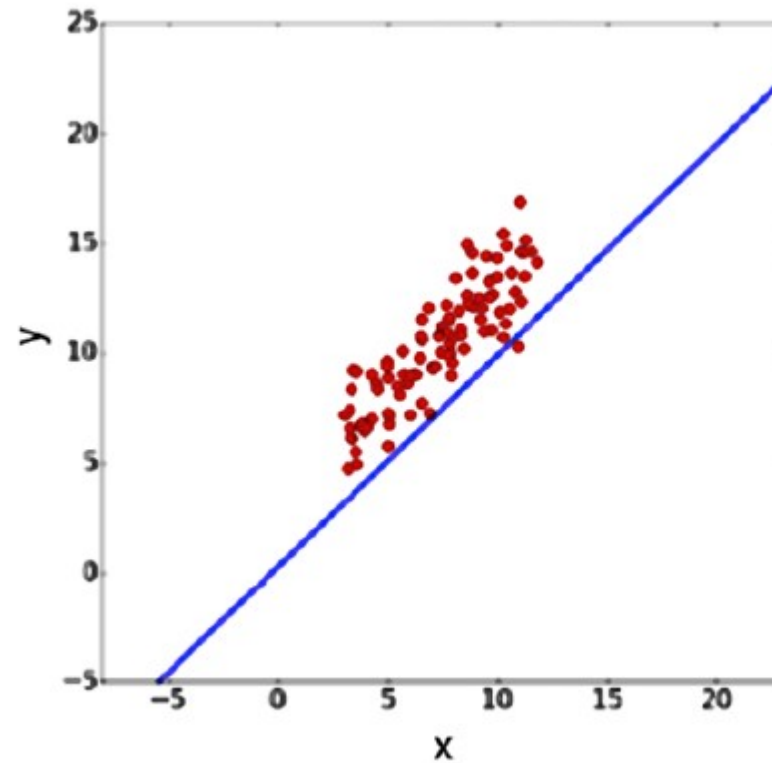
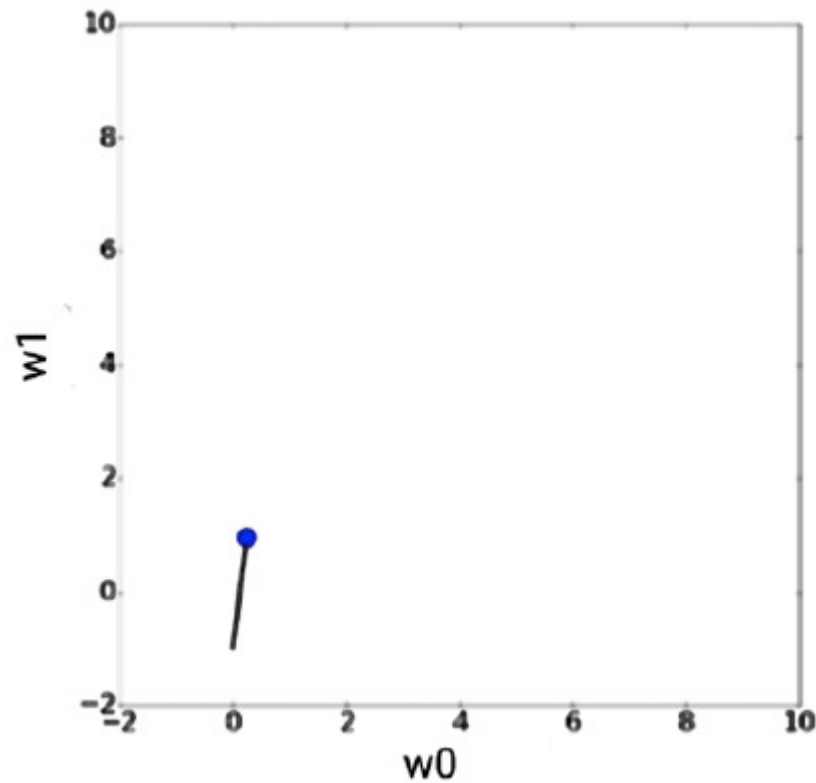
$$w_0^1 = w_0^0 - \eta_t \left. \frac{\partial Q}{\partial w_0} \right|_{w=w^0} = w_0^0 - \eta_t \frac{2}{l} \sum_{i=1}^l (w_1^0 x_i + w_0^0 - y_i) \cdot 1$$

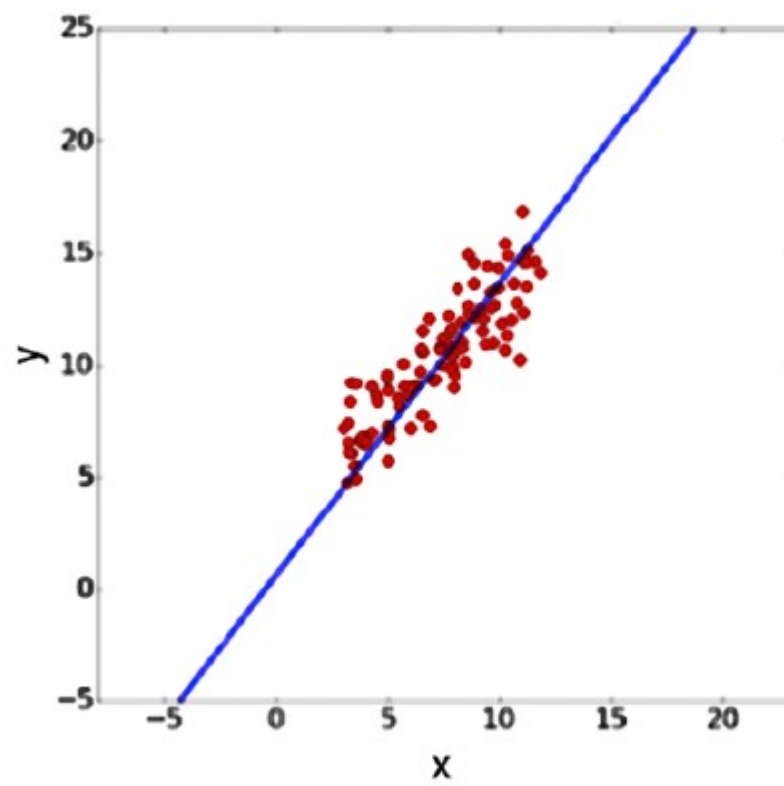
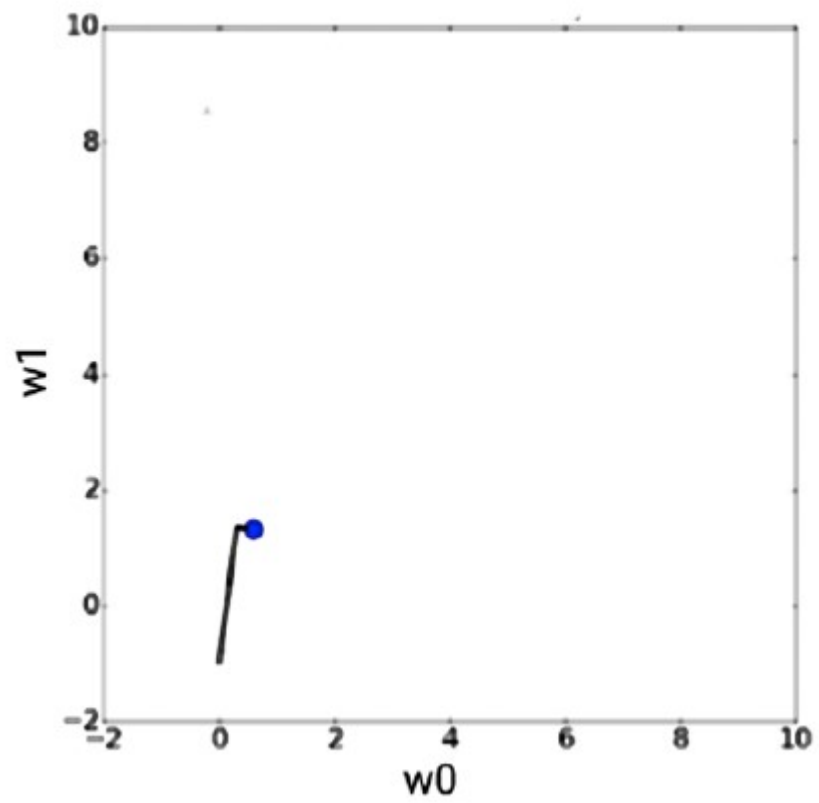
$$w_1^1 = w_1^0 - \eta_t \left. \frac{\partial Q}{\partial w_1} \right|_{w=w^0} = w_1^0 - \eta_t \frac{2}{l} \sum_{i=1}^l (w_1^0 x_i + w_0^0 - y_i) \cdot x_i$$

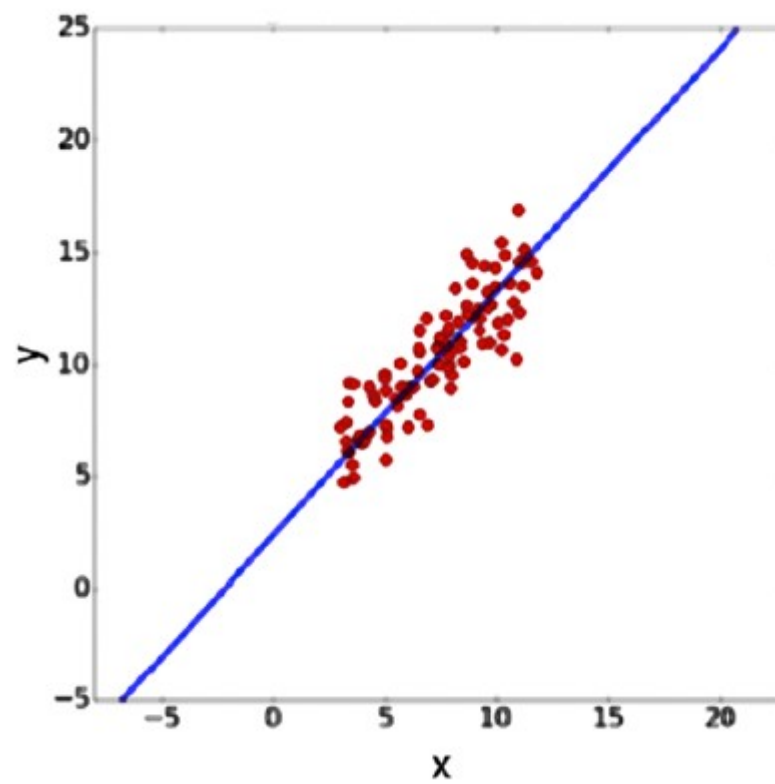
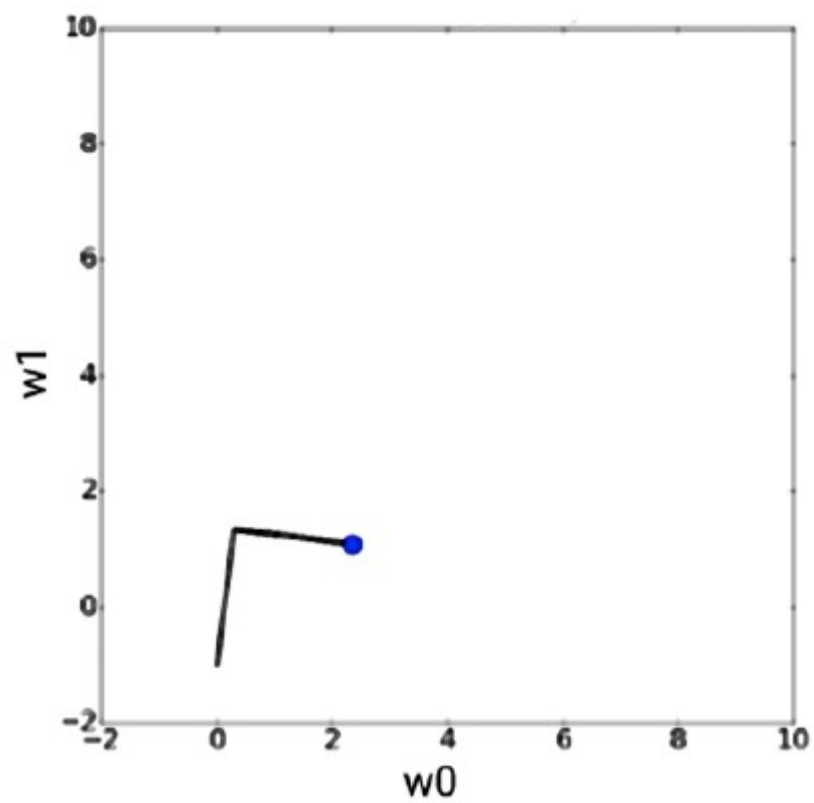


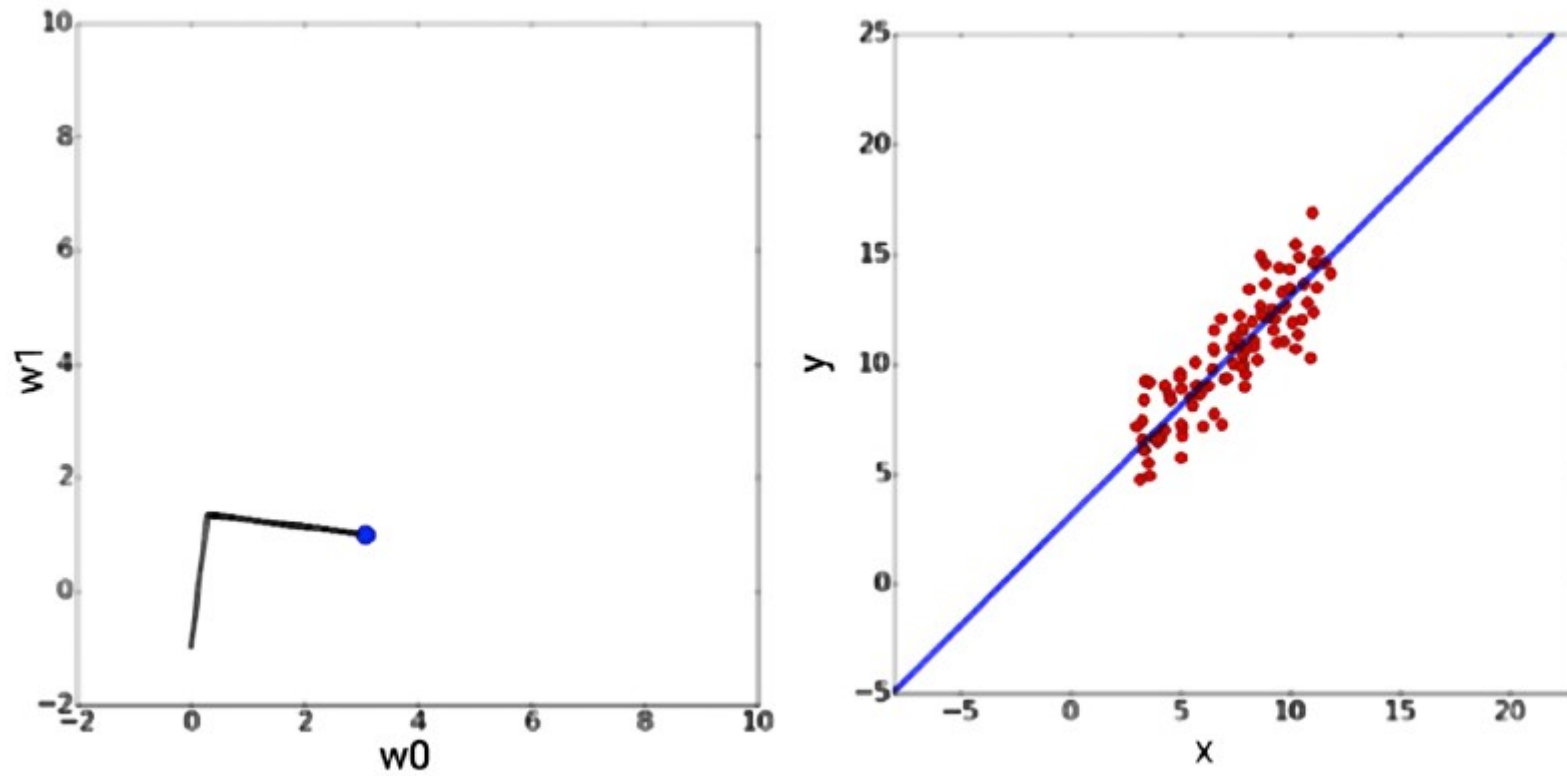
$$w_0^2 = w_0^1 - \eta_t \left. \frac{\partial Q}{\partial w_0} \right|_{w=w^1} = w_0^1 - \eta_t \frac{2}{l} \sum_{i=1}^l (w_1^1 x_i + w_0^1 - y_i) \cdot 1$$

$$w_1^2 = w_1^1 - \eta_t \left. \frac{\partial Q}{\partial w_1} \right|_{w=w^1} = w_1^1 - \eta_t \frac{2}{l} \sum_{i=1}^l (w_1^1 x_i + w_0^1 - y_i) \cdot x_i$$





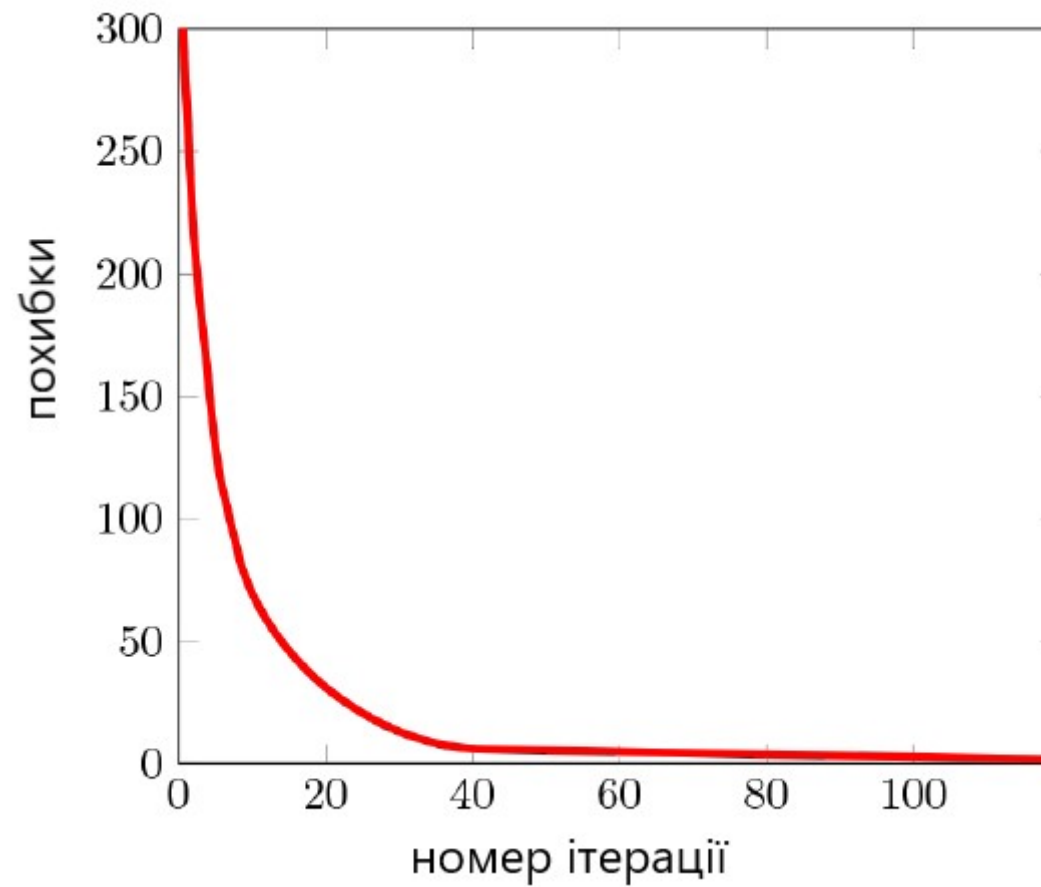




Зупинка, коли

$$\|w^t - w^{t-1}\| < \varepsilon$$

Графік залежності функції похибки від числа виконаних операцій



Вибір розміру кроку в методі градієнтного спуску

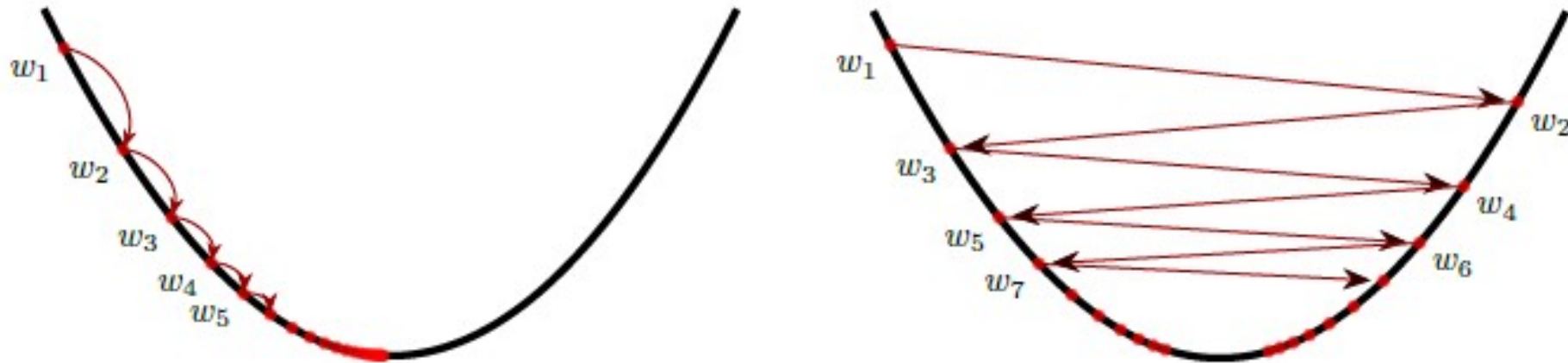


Рис. 2.4: Випадки маленького й великого кроку

Один зі способів задати розмір кроку наступний:

$$\eta_t = \frac{k}{t}$$

де k — константа, яку необхідно підібрати, а t — номер кроку.

Випадок багатомірної лінійної регресії

У випадку багатовимірної лінійної регресії використовується той же самий підхід — необхідно вирішувати задачу мінімізації:

$$Q(\mathbf{w}, X) = \frac{1}{\ell} \|X \mathbf{w} - \mathbf{y}\|^2 \rightarrow \min_{\mathbf{w}}$$

Формула для обчислення градієнта приймає наступний вигляд:

$$\nabla_{\mathbf{w}} Q(\mathbf{w}, X) = \frac{2}{\ell} X^T (X \mathbf{w} - \mathbf{y})$$

Варто відзначити, що вектор $X\mathbf{w} - \mathbf{y}$, що є присутнім у цьому виразі, є **вектором похибок**.

§19 Стохастичний градієнтний спуск

Недоліки звичайного методу градієнтного спуска

У звичайному методі градієнтного спуска на кожному кроці ітерації наступне наближення виходить із попереднього:

$$\mathbf{w}^t = \mathbf{w}^{t-1} - \eta_t \nabla Q(\mathbf{w}^{t-1}, X)$$

При цьому вираз для градієнта у матричній формі має вигляд:

$$\nabla_{\mathbf{w}} Q(\mathbf{w}, X) = \frac{2}{l} X^T (X \mathbf{w} - \mathbf{y})$$

Вираз для j -ої компоненти градієнта, таким чином, містить підсумовування за усіма об'єктами навчальної:

$$\frac{\partial Q}{\partial \mathbf{w}_j} = \frac{2}{l} \sum_{i=1}^{\ell} x_i^j (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)$$

Основний недолік методу градієнтного спуска - у випадку великої вибірки навіть **одна ітерація методу градієнтного спуска буде виконуватися довго.**

Стохастичний градієнтний спуск

Ідея стохастичного градієнтного спуску заснована на тому, що **в сумі** у виразі для j -компоненти градієнта i -й доданок вказує на те, як потрібно поміняти вагу w_j , щоб якість збільшилася для i -го об'єкта вибірки.

$$\frac{\partial Q}{\partial w_j} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_i^j (\langle \mathbf{w}, x_i \rangle - y_i)$$

У **стохастичному методі** градієнтного спуску градієнт функції якості обчислюється тільки **на одному випадково обраному об'єкті** навчальної вибірки.

Алгоритм стохастичного градієнтного спуска наступний.

Спочатку вибирається початкове наближення:

$$\mathbf{w}^0 = \mathbf{0}$$

Далі послідовно обчислюються ітерації \mathbf{w}^t :

- випадковим чином вибирається об'єкт x_i з навчальної вибірки X ;
- обчислюється вектор градієнта функції якості на цьому об'єкті

$$\frac{\partial Q}{\partial w_j} = 2x_i^j (\langle \mathbf{w}, x_i \rangle - y_i)$$

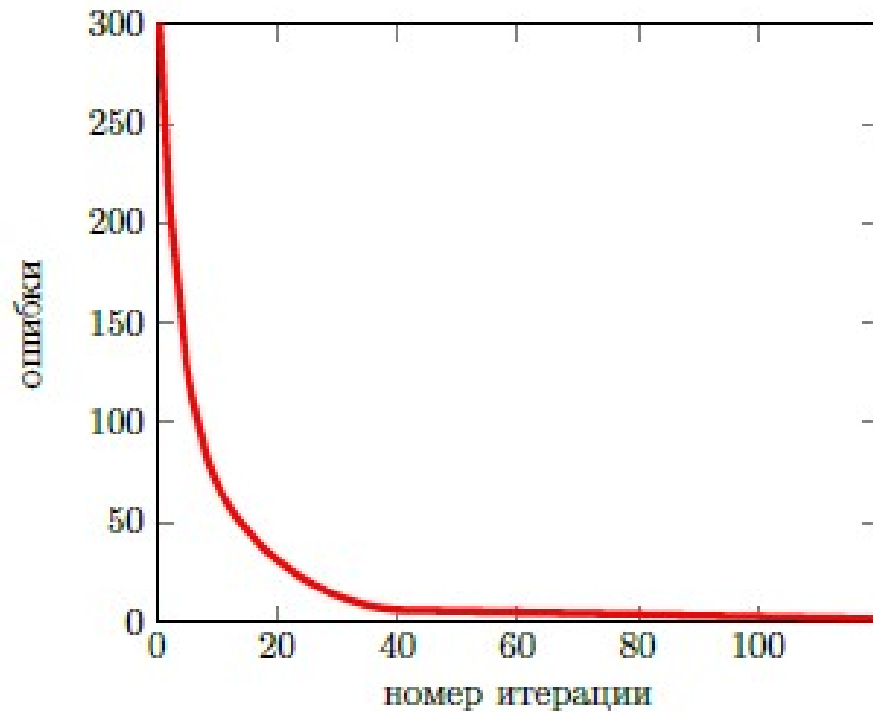
- наступне наближення отримаємо із попереднього:

$$\mathbf{w}^t = \mathbf{w}^{t-1} - \eta_t \nabla Q(\mathbf{w}, \{x_i\})$$

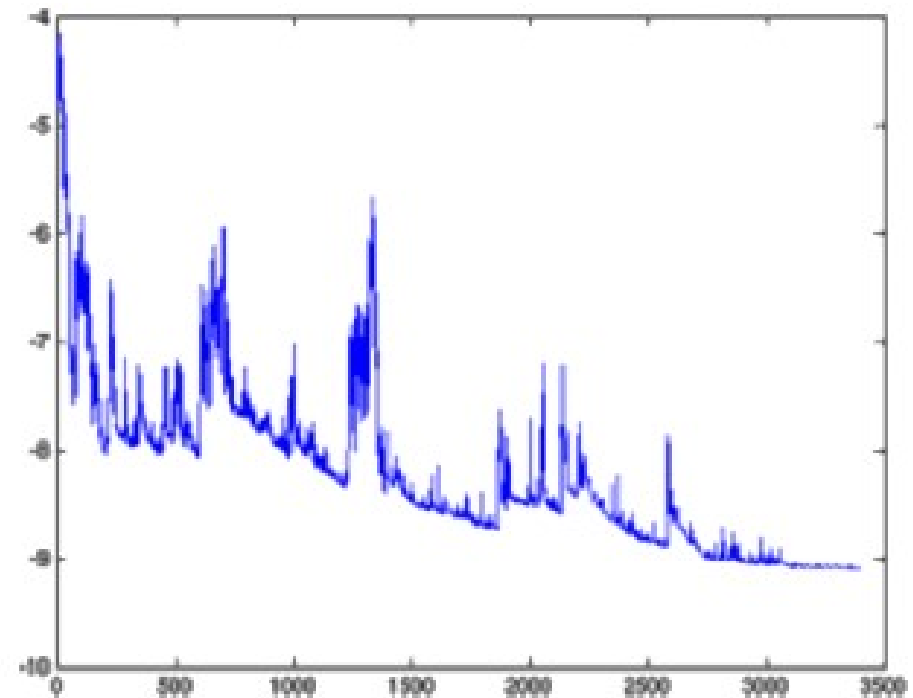
- Ітерації припиняються при досягненні певної умови, наприклад:

$$\| \mathbf{w}^t - \mathbf{w}^{t-1} \| < \varepsilon$$

Збіжність стохастичного градієнтного спуска



Градiєнтний спуск



Стохастичний градієнтний спуск

Особливості стохастичного градієнтного спуска (SGD)

- Кожний крок виконується **ЗНАЧНО ШВИДШЕ** кроку звичайного градієнтного методу
- **Не потрібно** постійно зберігати всю навчальну вибірку в пам'яті. Це дозволяє використовувати для навчання **дуже великі вибірки**
- Можна використовувати **для онлайн-навчання**, тобто в ситуації, коли на кожному кроці алгоритм одержує тільки один об'єкт і повинен урахувати його для корекції моделі.

§20 Лінійна класифікація

Задача бінарної класифікації

У випадку бінарної класифікації множина можливих значень відповідей складається із двох елементів:

$$\mathbb{Y} = \{-1, +1\}$$

Далі використовуємо стандартний підхід:

- Вибрати **функціонал (функцію) похибки**, тобто задати спосіб визначення якості роботи того або іншого алгоритму на навчальній вибірці.
- Побудувати **сімейство алгоритмів**, тобто множину алгоритмів, з якого потім буде вибиратися найкращий з погляду певного функціонала похибки.
- Увести **метод навчання**, тобто визначити спосіб вибору кращого алгоритму із сімейства.

Лінійний класифікатор

У задачі лінійної регресії алгоритм являв собою лінійну комбінацію ознак з деякими вагами й вільним коефіцієнтом.

Лінійні класифікатори схожі, але вони повинні повертати бінарні значення, а отже потрібно також брати знак від отриманого виразу:

$$a(x) = \text{sign} \left(\underset{\uparrow}{w_0} + \sum_{j=1}^d \underset{\uparrow}{w_j} \underset{\uparrow}{x^j} \right)$$

Вільний коефіцієнт

Ваги

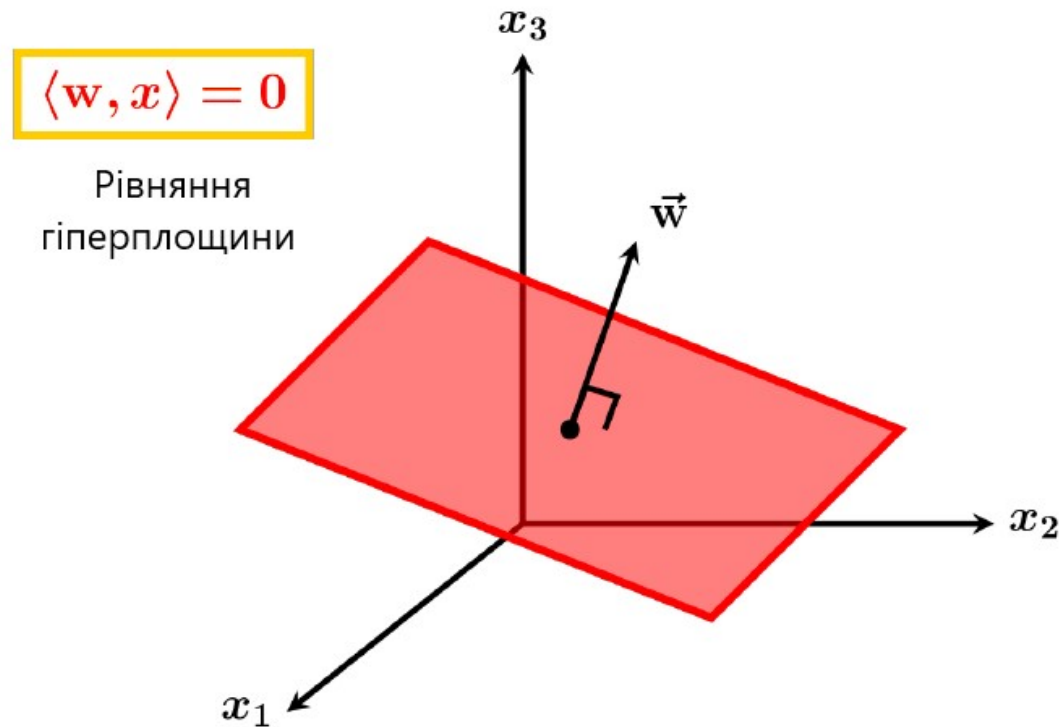
Ознаки

Як і раніше, додаванням ще одного постійного для всіх об'єктів ознаки можна привести формулу до більше однорідного виду:

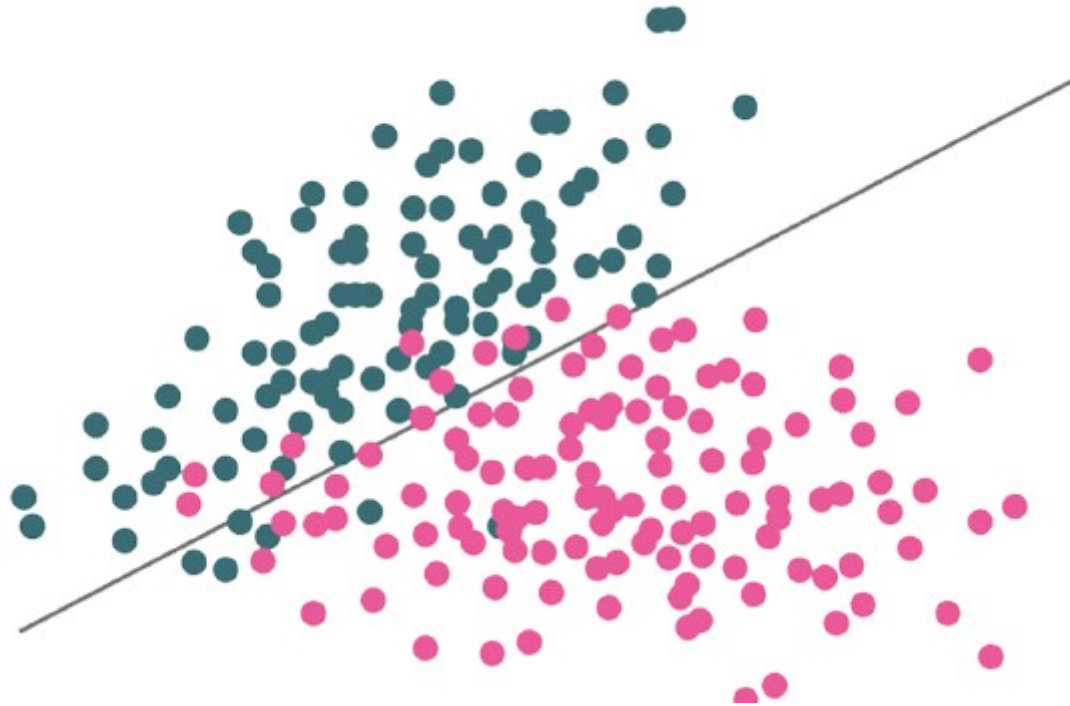
$$a(x) = \text{sign} \sum_{j=1}^{d+1} w_j x^j = \text{sign} \langle \mathbf{w}, \mathbf{x} \rangle$$

Геометричний зміст лінійного класифікатора

Вираз $\langle w \cdot x \rangle = 0$ є рівнянням деякої площини в просторі ознак



При цьому для точок по одну сторону від цієї площини скалярний добуток буде **додатнім**, а з іншого боку – **від'ємним**.

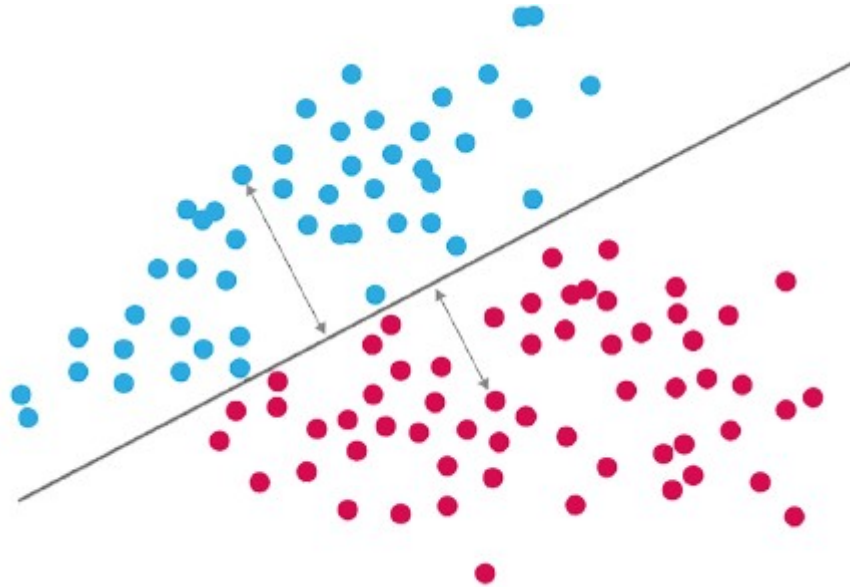


Таким чином, лінійний класифікатор **проводить площину в просторі ознак і відносить об'єкти у різні сторони від площини до різних класів.**

Відстань від конкретного об'єкта, що має ознаковий опис x , до гіперплощини $\langle w \cdot x \rangle = 0$ дорівнює

$$\frac{|\langle w, x \rangle|}{\|w\|}$$

Чим більше $\langle w, x \rangle$ тим ділі об'єкт від гіперплощини



Із цим зв'язане таке важливе поняття в задачах лінійної класифікації як **поняття відступу**:

$$M_i = y_i \langle w, x \rangle$$

Відступ є величиною, що визначає **коректність відповіді**.

- Якщо відступ більше нуля $M_i > 0$, то класифікатор дає **вірну відповідь** для i -го об'єкта,
- Якщо відступ менше нуля $M_i < 0$, то **класифікатор помиляється**.
- Причому чим **далі відступ від нуля**, тим **більше впевненість як у правильній відповіді**, так і в тому, що алгоритм помиляється.

§21 Функції втрат у задачах класифікації

Гранична функція втрат

У випадку лінійної класифікації природний спосіб визначити якість того або іншого алгоритму – обчислити для об'єктів навчальної вибірки **частку неправильних відповідей**:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) \neq y_i]$$

За допомогою уведеного раніше поняття відступу можна переписати цей вираз для випадку лінійної класифікації наступним чином:

$$Q(a, x) = \frac{1}{\ell} \sum_{i=1}^{\ell} [y_i \langle w, x_i \rangle < 0] = \frac{1}{\ell} \sum_{i=1}^{\ell} [M_i < 0]$$

Функція, що знаходиться під знаком суми, називається **функцією втрат**.

У цьому випадку це гранична функція втрат, графік якої залежно від відступу виглядає в такий спосіб:

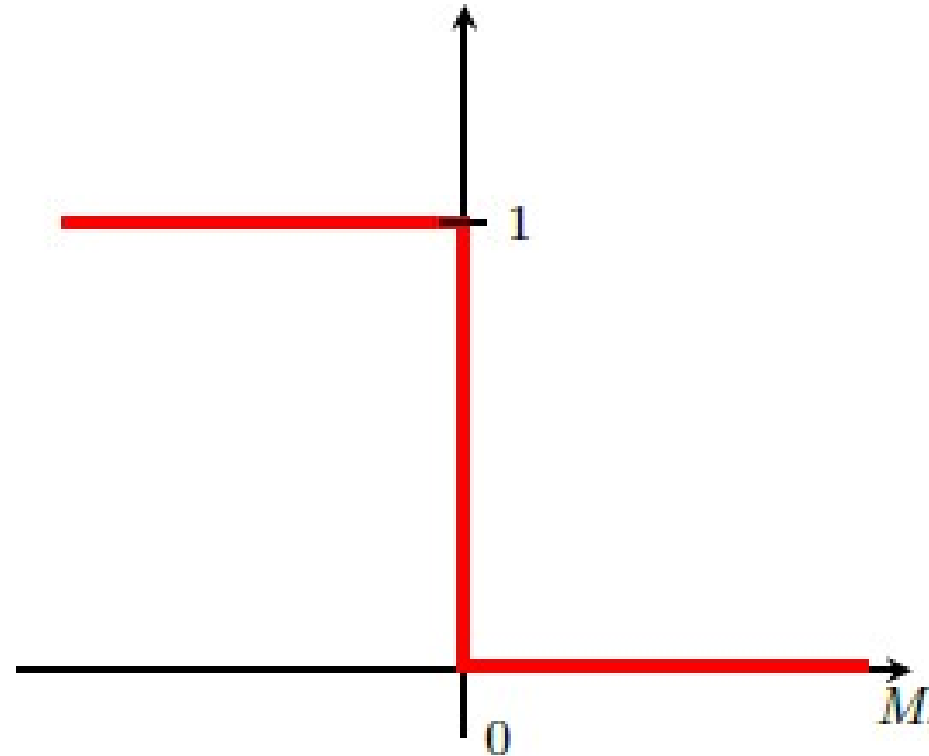


Рис. 2.7: Графік граничної функції втрат

Така функція є розривною в точці 0, що унеможлиблює застосування методу градієнтного спуска.

Оцінка функції втрат

Використовуючи будь-яку гладку оцінку граничної функції:

$$[M < 0] \leq \tilde{L}(M)$$

можна побудувати оцінку $\tilde{Q}(a, X)$ для функціонала похибки $Q(a, X)$:

$$Q(a, X) \leq \tilde{Q}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \tilde{L}(M_i)$$

У цьому випадку мінімізувати потрібно буде не частку неправильних відповідей, а деяку іншу функцію, що є оцінкою зверху:

$$Q(a, X) \leq \tilde{Q}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \tilde{L}(M_i) \rightarrow \min_a$$

Тут використовується припущення, що в точці мінімуму цієї верхньої оцінки число помилок також буде мінімальним. У загальному випадку, це не завжди так.

Приклади оцінок функції втрат

Прикладами таких оцінок функції втрат є:

- Логістична функція втрат (використовується в логістичній регресії):

$$\tilde{L}(M) = \ln(1 + \exp(-M))$$

- Експонентна функція втрат:

$$\tilde{L}(M) = \exp(-M)$$

- Кульково-лінійна функція втрат (використовується в методі опорних векторів):

$$\tilde{L}(M) = \max(0, 1 - M)$$

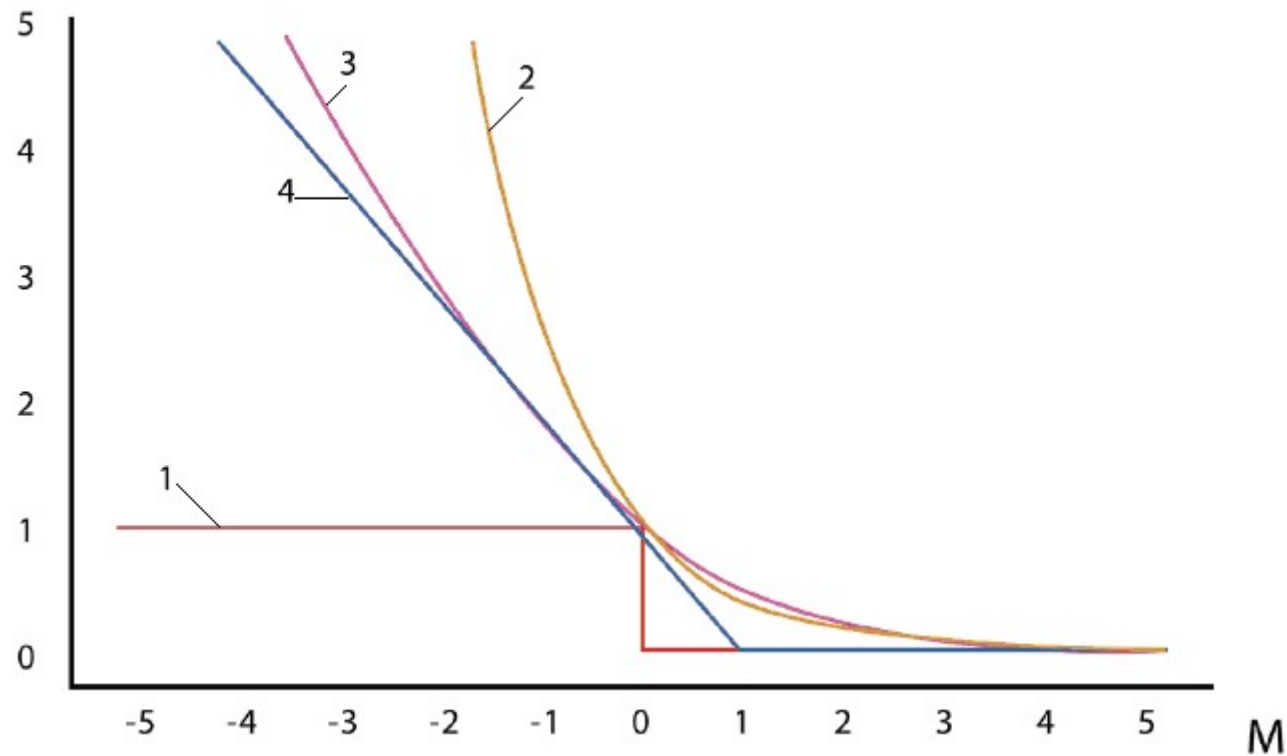


Рис. 2.8: Графіки різних функцій втрат: гранична (1, червона лінія), експонентна (2, синя), логістична (3, жовтогаряча) і кусково-лінійна (4, сіра).

Логістична функція втрат

У випадку логістичної функції втрат функціонал похибки має вигляд:

$$\tilde{Q}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \ln (1 + \exp (-M_i))$$

$$\tilde{Q}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \ln (1 + \exp (-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle))$$

Отриманий **вираз є гладким**, а, отже, можна використовувати, наприклад, метод градієнтного спуску.

Варто звернути увагу, що у випадку, якщо число помилок стало дорівнювати нулю, однаково в ході навчання алгоритму лінійної класифікації **будуть збільшуватися відступи**, тобто буде збільшуватися впевненість в отриманих результатах.

§22 Комп'ютерний проект 1: "КП01_linreg_height_weight"

§23 Комп'ютерний проект 2:

"КП02_linreg_stochastic_grad_descent"