# Лекція № 6

## Навчання НМ на основі методів оптимізації другого порядку

Проф. Куссуль Н.М.

# Основна ідея

- Градієнтні методи другого порядку — це методи оптимізації з використанням других похідних (матриці Гессе). Якщо методи 1-го порядку використовують **лінійне наближення** поверхні похибки, то методи 2-го порядку — **квадратичне наближення**.
- Якщо **функція похибки насправді є квадратичною**, то розв'язок може бути знайдений дуже швидко, наприклад, за методом Ньютона — за 1 крок.
- Методи 2-го порядку **ефективні в локальному околі** мінімуму (забезпечуючи тим найшвидшу збіжність до нього).
- Однак **у широкому околі** вони працюють погано, тому найчастіше для пошуку початкового наближення використовують інші методи.

# 6.1. Узагальнена постановка задачі навчання мережі

Навчальна вибірка

$$\{X_j, T_j\},\ j = \overline{1, L}.$$

$$X_j = (x_1^{(j)}, \ldots, x_{Ni}^{(j)}),\quad T_j = (t_1^{(j)}, \ldots, t_{No}^{(j)}) \qquad (6.1)$$

Вихід НМ для j-го вектора навчальної вибірки $Y_M(X_j; w)$

Задача навчання – налаштування (ідентифікація вагових коефіцієнтів *w*), щоб

$$Y_M(X_j; w) = T_j,\ j = \overline{1, L} - \textit{для навчальної вибірки}, \qquad (6.2)$$

$$Y_M(X_i; w) = T_i,\ i = \overline{1, N_t} - \textit{для тестової вибірки}$$

Застосуємо метод розв'язання систем нелінійних рівнянь, що базується на мінімізації «нев'язки» між лівими і правими частинами (6.2).

# 6.2. Типи похибок і критерії навчання мережі

Вектор похибки моделі для образу *j* навчальної вибірки

$$e^{\mathrm{T}}(j;w) = (e_1(j;w), \ldots, e_{N_L}(j;w))$$

$$e_i(j;w) = y_i^{(L)}(X_j;w) - t_i^{(L)} \qquad (6.3)$$

$$E_J = \sqrt{\sum_{j=1}^{N_L} e_i^{\,2}(j)}, \quad j = \overline{1,m} \qquad (6.4)$$

Критерій навчання кумулятивних і оптимізаційних алгоритмів навчання

$$\Phi(w) = \frac{1}{2}\|E\|^2 = \frac{1}{2}\sum_{J=1}^{m}\sum_{i=1}^{N_L} e_i^{\,2}(j) \to \min, \qquad (6.5)$$

$$E = \frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{N_L}(t_i^{\,j} - t_i^{\,j}) = \left\langle (t-y)^2 \right\rangle \qquad (6.6)$$

# 6.3. Вектор градієнта і матриця Гессе

$$\Phi(w) = \frac{1}{2}\|E\|^2 = \frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{N_L} e_i{}^2(j) = \frac{1}{2}E^T E \qquad (6.7)$$

$$w = \arg\min_{w \in B} \Phi(w) \qquad (6.8)$$

$$\operatorname{grad}_w \Phi(w) = \nabla\Phi(w) = \frac{\partial\Phi}{\partial w} = \left(\frac{\partial\Phi}{\partial w_1}, \ldots, \frac{\partial\Phi}{\partial w_p}\right) \qquad (6.9)$$

$$H_k = H(w^{(k)}) = \left\{\frac{\partial^2\Phi(w^{(k)})}{\partial w_i\,\partial w_j}\right\}_{i,\,j=\overline{1,n}} \qquad (6.10)$$

# 6.4. Загальна схема оптимізаційних алгоритмів навчання

**Алгоритм мінімізації** - ітераційний процес, що включає такі основні етапи обчислень.

1. Вибір початкового наближення w.
2. Визначення напрямку спуску *p* (специфіка конкретного методу оптимізації).
3. Спуск у заданому напрямку $\lambda_k = \arg\min_{\lambda \geq 0} E(w^{(k)} + \lambda h_k p^{(k)})$
4. Побудова наступного наближення точки мінімуму

$$w^{(k+1)} = w^{(k)} + \lambda h_k p^{(k)}$$

5. Перевірка критерію завершення і перехід до п.2

$$H_k = H(w^{(k)}) = \left\{ \frac{\partial^2 \Phi(w^{(k)})}{\partial w_i \partial w_j} \right\}_{i,j=\overline{1,n}}$$

# 7.1. Ідея методу Ньютона

- Метод Ньютона – це теоретичний стандарт для оптимізації гладких функцій, з яким порівнюються інші методи.
- Оскільки він використовує всю доступну інформацію про похідні 1-го і 2-го порядку в явній формі, то має відмінні властивості локальної збіжності.
- На жаль, найчастіше він є **непридатним** для практичного використання, оскільки у великих задачах **обчислення Гессіана є дуже ресурсомістким**.
- Представимо матриці вагових коефіцієнтів у виді узагальненого вектора ***w***

# 7.2. Метод Ньютона

Квадратична модель функції похибки – розклад у ряд Тейлора з точністю до членів 2 порядку:

$$E(w) = E_0 + g^{\mathrm{T}} w + \frac{1}{2} w^{\mathrm{T}} H w \qquad (6.11)$$

$$g = \frac{\partial E}{\partial w} \qquad (6.12)$$

$$h_{ij} = \frac{\partial^2 E}{\partial W_i \partial W_j} \qquad (6.13)$$

Передбачається $H > 0$

# 7.2. Метод Ньютона (продовження)

Розв'яжемо (6.11)

$$E(w) = E_0 + g^{\mathrm{T}} w + \frac{1}{2} w^{\mathrm{T}} H w$$

$$\frac{\partial E}{\partial w} = 0 \qquad (6.14)$$

$$g + H w = 0 \qquad (6.15)$$

$$w^* = -H^{-1} g \qquad (6.16)$$

$$w(t+1) = w(t) - \eta H^{-1} g \qquad (6.17)$$

# 7.2. Метод Ньютона (продовження)

Необхідні умови збіжності:

$$H > 0$$

$$0 < \eta < \frac{2}{\lambda_{\max}}$$

$\lambda_{\max}$ — найбільше власне значення *H*

**Проблеми метода Ньютона:**

— в реальних задачах H може не бути додатньо визначеною;

— обчислювальна складність обчислення і зберігання прямої та оберненої матриці Гессе

(для N зв'язків розмір матриці Гессе NxN. Для точного обчислення H необхідно $\approx O(N^2)$ епох);

# 7.3. Метод Гаусса-Ньютона

Метод Гауса-Ньютона оснований на лінійному наближенні Гессіана. Методи Гауса-Ньютона і Левенберга-Марквардта використовують переваги спеціальної структури Гессіана в методі найменших квадратів.

Зазвичай функція вартості $E$ — середньоквадратична похибка $E = \left\langle (d-y)^2 \right\rangle$, де $\langle \dots \rangle$ — середнє значення на епосі. Для простоти припустимо, що $y$ — скаляр. Представимо вагові коефіцієнти у вигляді єдиного для всієї мережі вектора $w \in R^N$, тоді компоненти градієнта $g \in R^N$ ( $N$ — число зв'язків) обчислюються за формулою $g_j = \frac{\partial E}{\partial w_j} = -2\left\langle (d-y)\frac{\partial y}{\partial w_j} \right\rangle$, а компоненти Гессіана $H$ $h_{ij} = \frac{\partial E}{\partial w_i \partial w_j} = 2\left\langle \frac{\partial y}{\partial w_i}\frac{\partial y}{\partial w_j} - (d-y)\frac{\partial^2 y}{\partial w_i \partial w_j} \right\rangle$.

Це співвідношення можна переписати у вигляді: $H = 2(P-Q)$, де $P = \left\langle g g^{\mathrm{T}} \right\rangle$ — векторний добуток градієнтів, а $q_{ij} = \left\langle (d-y)\frac{\partial^2 y}{\partial w_i \partial w_j} \right\rangle$ містить члени 2-го порядку.

Оскільки $P$ — векторний добуток градієнтів, то це дійсна, симетрична і невід'ємно визначена матриця, тобто $\forall \|w\| \neq 0 \qquad w^{\mathrm{T}} P w \geq 0$

Для забезпечення повного рангу розмір навчальної вибірки повинен перевищувати число зв'язків. Якщо резидуальна похибка $(d-y)$ являє собою незміщені, незалежні, рівномірно розподілені значення некорельовані з іншими похідними $\frac{\partial^2 y}{\partial w_i \partial w_j}$, а розмір навчальної вибірки досить великий, то $\langle Q \rangle \to 0$ і нею можна зневажити.

У цьому випадку можна використовувати лінійне наближення Гессіана: $H \approx 2P$

Однак зазначені припущення не завжди виконуються, наприклад, коли навчальна вибірка містить дуже мало точок, або мережа занадто мала, тому похибка не зменшується і корелює з іншими похідними.

У методі Гауса-Ньютона використовується перше наближення Гессіана, тобто розв'язок рівняння (7.2) записується у вигляді: $\Delta w = P^{-1} g$.

У цьому методі не потрібно обчисювати другі похідні, але, як і раніше, необхідно зберігати і обертати матрицю $N \times N$. Крім того, матриця, як і раніше, повинна бути додатно визначеною.

# 7.4. Метод Левенберга-Марквардта

У роботах Левенберга, Марквардта, Голдфілда та інших була запропонована обчислювальна схема забезпечення додатної визначеності наближення оберненої матриці Гессе $H^{-1}$ . Вона базується на наступній властивості матриць.

Якщо $P$ — додатно визначена матриця, то при досить великому значенні $\beta > 0$ матриця $A + \beta P$ буде додатно визначеною, незалежно від матриці $A$ .

У даному методі напрямок і величина кроку спуску визначається в такий спосіб

$$w^{(k+1)} = w^{(k)} - R_k g^{(k)}, \text{ де } R_k = (\tilde{H}_k + \beta B_k{}^2)^{-1}, \quad k = 0,1,\ldots,$$

де $P_k = B_k{}^2$, $B_k$ — діагональна матриця, $\tilde{H}_k$ — $k$ -е наближення матриці Гессе.

$$\tilde{H}_0 = \begin{pmatrix} -\frac{w_1^{(0)}}{g_1^{(0)}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -\frac{w_n^{(0)}}{g_n^{(0)}} \end{pmatrix} \quad \textbf{(7.1)}$$

Обчислювальна схема методу Левенберга-Марквардта має наступний вигляд.

1. Початкове наближення матриці $\tilde{H}_0$ при $k = 0$ оберемо у формі (7.1)

2. Утворимо матрицю $R_k = (\tilde{H}_k + \beta B_k{}^2)^{-1}$ , де $B_k{}^2 = \operatorname{diag}\{b_{jj}^2\}_{j=1,n}$ , а її діагональні елементи обчислюються в такий спосіб $b_{jj} = \begin{cases} \sqrt{|h_{jj}|}, & h_{jj} \neq 0 \\ 1, & h_{jj} = 0. \end{cases}$ Значення $\beta > 0$ можна призначити досить великим або обрати з умови $\beta > -\min_i \alpha_i$ , де $\alpha_i$ — власне значення розкладу матриці $\Pi = B_k{}^{-1} \tilde{H}_k B_k$

3. Визначаємо напрямок спуску $\nu_k$ . Для цього утворимо розклад матриці за власними векторами $\tilde{H}_k = \sum_{i=1}^{n} \alpha_i \nu_i \nu_i{}^{\mathrm{T}}$ Тут $\nu_i$ — власні вектори матриці $\tilde{H}_k$ , а $\alpha_i$ — підкреговані власні значення матриці $\tilde{H}_k$ таким чином, що усі $\alpha_i > 0$ ; – на базі цього розкладу побудуємо розклад матриці $R_k = (\tilde{H}_k + \beta B_k{}^2)^{-1} = \sum_{i=1}^{n} (\alpha_i + \beta)^{-1} \nu_i \nu_i{}^{\mathrm{T}}$ Напрямок спуску визначається співвідношенням $\nu_k = -R_k g_k$

4. Здійснюємо одновимірний спуск у напрямку $\nu_k$ .

5. Як критерій завершення алгоритму можуть бути прийняті умови завершення описаного нижче методу Флетчера-Пауелла-Девідона.

Метод Левенберга-Марквардта забезпечує ефективний перехід від методу градієнтного спуску до методу Ньютона в околі мінімуму.

# Висновок до лістингу 7.4

Таким чином, метод Левенберга-Марквардта — це компромісний варіант між методом Ньютона, що добре збігається в околі мінімуму, і методом градієнтного спуску, що збігається до локального мінімуму в будь-якій області, хоча і повільно.

Напрямок пошуку — це лінійна комбінація між напрямком градієнтного (найшвидшого) спуску $g$ і Ньютонівським напрямком $H^{-1}g$ :

$$w(t+1) = w(t) - (H + \lambda I)^{-1} g$$

де $\lambda$ — параметр керування компромісом (по ньому виконується одновимірна оптимізація). Він забезпечує додатну визначеність матриці $H + \lambda I$ шляхом масштабування одиничної матриці.

Мінімально припустиме значення $\lambda$ визначається власними числами матриці $H$. Робота алгоритму починається з великого значення $\lambda$. У процесі ітерацій $\lambda$ динамічно налаштовується так, щоб похибка на кожнім кроці зменшувалася. На початкових кроках, поки $\lambda$ велике, рух визначається напрямком градієнтного спуску. На наступних кроках, коли $\lambda \to 0$, відбувається перехід до методу Ньютона.

Тут *усувається* проблема дивергенції і необхідності лінійного пошуку, але *зберігається* потреба *збереження і обертання* матриці Гессе.

Цей метод дає гарні результати в задачах середнього розміру.

# 7.5. Квазі-ньютонівські методи

Квазі-ньютонівські методи (іноді їх називають методами із змінною метрикою) ітеративно будують наближення Гессіана з використанням інформації про перші похідні. При цьому матриця Гессе (або обернена до неї матриця) апроксимуються деякою додатно визначеною матрицею $\eta(w^{(k)})$, яка обчислюється ітеративно. Модифікація вагових коефіцієнтів здыйснюється за формулою

$$w^{(k+1)} = w^{(k)} + \lambda_k p^{(k)} = w^{(k)} - \lambda_k \eta(w^{(k)}) g^{(k)}, \qquad k = 0,1,\ldots$$

з використанням лише перших похідних. Тут $\lambda_k$ — крок спуску.

При цьому усувається необхідність явно обчислювати і обертати Гессіан, однак потрібно зберігати його наближення.

***Найбільш відомі два квазі-ньютонівських методи:***

– Флетчера-Пауэла-Девідона — DFP (Davidon-Fletcher-Powell),

– Бройдена-Флетчера-Гольдфарба-Шанно — BFGS (Brouden-Fletcher-Goldfarb-Shanno).

Вони відрізняються лише деякими деталями.

# 7.6. Метод Флетчера-Пауела-Девідона

У даному методі в якості наближення $\eta(w^{(k)})$ до оберненої матриці Гессе $H_k^{-1}$ будується матриця $G_k$. Це виключає операцію обертання матриці в самому алгоритмі. У процесі ітерацій алгоритму $G_k \to H^{-1}$, причому для функцій квадратичного вигляду $G_k$ збігається до $H^{-1}$ за $n$ кроків: $G_n \approx H^{-1}$. У якості початкового наближення $G_0$ обирається одинична матриця $G_0 = E$ (хоча в якості $G_0$ може бути обрана будь-яка додатно визначена матриця). Матриця $G_k$ уточнюється в процесі ітерацій алгоритму *Флетчера-Пауела-Девідона*, обчислювальна схема якого має наступний

1. При $k = 0$ початковий напрямок спуску відповідає напрямку антиградієнту

$$h^{(0)} = -g^{(0)}, \quad \text{де} \quad g^{(0)} = \operatorname{grad}\Phi(w^{(0)}).$$

2. Обчислюється наступне наближення вектору вагових коефіцієнтів

$$w^{(k+1)} = w^{(k)} + \lambda_k h^{(0)}, \qquad k = 0,1,\ldots,$$ корегується напрямок спуску

$$h^{(k+1)} = -G_{k+1} g^{(k+1)}, \quad \text{де} \quad g^{(k+1)} = \operatorname{grad}\Phi(w^{(k+1)})$$

і матриця $G_{k+1} = G_k + \dfrac{dw \cdot dw^{\mathrm{T}}}{dw^{\mathrm{T}} \cdot dw} - \dfrac{G_k dg \cdot dg^{\mathrm{T}} G_k}{dg^{\mathrm{T}} G_k\, dg}$ де $dg = g^{(k+1)} - g^{(k)}, \quad dw = w^{(k+1)} - w^{(k)}$ ;

3. Виконується одновимірний спуск за допомогою кубічної інтерполяції між точками

$\lambda_0 = 0$ і $\lambda_1 = \min\left\{1, \dfrac{2(\Phi(w^{(k)}) - \Phi_0)}{(g^{(k)})^{\mathrm{T}} G_k\, g^{(k)}}\right\}$, де $\Phi_0$ — мінімальне очікуване значення функції.

Критерієм завершення одновимірного пошуку є умова $\left|\dfrac{\lambda_{k+1} - \lambda_k}{\lambda_{k+1} + \lambda_k}\right| \le \varepsilon$, де $\varepsilon$ — задана точність (параметр алгоритму).

4. Умова завершення процесу має вигляд $\left\|g^{(k+1)}\right\| \le \varepsilon$

При реалізації алгоритму необхідно стежити за симетричністю та додатною визначеністю матриці Гессе.

Таким чином, Квазі-Ньютонівські методи мають наступні переваги:

**1. Швидкість їх збіжності відповідає методу спряжених градієнтів.**

**2. Вони забезпечують швидку збіжність в околі локального мінімуму.**

# 7.7. Оптимізаційні алгоритми навчання з фіксацією груп параметрів мережі

З огляду на велику розмірність вектора параметрів мережі $w=(w_1,\ldots,w_p)$, що є істотним фактором у збіжності методів навчання, а також на однакову значимість усіх компонентів $w_i$ у процесі мінімізації критерію навчання $\Phi(w)$ на різних етапах навчання, доцільно виконувати навчання по групах компонентів з урахуванням чутливості до них функції $\Phi(w)$. А саме, на різних етапах процесу навчання оптимізацію похибки $\Phi(w)$ виконувати лише для тих компонентів $w_i$, до яких $\Phi(w)$ є найбільш чуттєвою. Вибір компонентів у таку групу $\tilde{w}$ можна здійснювати з таких міркувань. Якщо на $k$-ій ітерації маємо нормований вектор градієнта

$$g_k=\frac{\operatorname{grad}\Phi(w^{(k)})}{\left\|\operatorname{grad}\Phi(w^{(k)})\right\|}=(g_1,\ldots,g_p),$$

то в групу $I=(i_1,\ldots,i_n)$ індексів чуттєвих компонентів $\tilde{w}=(\tilde{w}_1,\ldots,\tilde{w}_n)\equiv(w_{i_1},\ldots,w_{i_n})$ включаються ті з компонентів $w_i$, для яких $|g_i|\geq\varepsilon$, де $\varepsilon=\frac{1}{p}$

Значення $\varepsilon$ може бути призначено і з інших міркувань.

Наступний етап навчання здійснюється шляхом мінімізації функції $\min_{\tilde{w}}\Phi(\tilde{w})$ при фіксованих інших компонентах $w_i$, $\forall i\notin I$.

Реалізацію цього процесу можна здійснити на базі будь яких методів мінімізації без жодної їх модифікації. Для цього паралельно з вектором $w=(w_1,\ldots,w_p)$ утворимо вектор $\tilde{w}=(\tilde{w}_1,\ldots,\tilde{w}_n)$ і відповідний вектор індексів $I=(i_1,\ldots,i_n)$ (або однобітових індикаторів $I=(i_1,\ldots,i_p)$, у якому $i_k=0$, якщо компонент $w_k$ фіксований, інакше $i_k=1$). Тоді допоміжний модуль, що керує навчанням мережі з урахуванням фіксації деяких компонентів, повинен виконувати наступні функції.

– З огляду на значення компонентів нормованого вектора градієнта $g^{(k)}=(g_1,\ldots,g_p)$ на поточній ітерації $w^{(k)}$, відповідно до умови $|g_i|\geq\varepsilon$ формуються вектори $\tilde{w}=(\tilde{w}_1,\ldots,\tilde{w}_n)$ і $I=(i_1,\ldots,i_n)$

– У модуль мінімізації замість вектора $w$ передається вектор $\tilde{w}$ розмірності $n$ а в якості функції, що мінімізується — допоміжна функція $\Phi(\tilde{w})$, призначення якої — розгорнути вектор $\tilde{w}$ у вектор $w$ і звернутися до стандартної функції навчання.

Очевидно, що фактор фіксації компонентів тим або іншим чином повинен враховуватися й у модулях обчислення вектора градієнта, а також у модулях аналітичного обчислення матриці Гессе (якщо такі є).