FTML practical session 14

14 juin 2025

Table des matières

1 Convergence speed of gradient descent

1.0.1 Step 1

$\forall \theta \in \mathbb{R}^d$,

$$\begin{aligned} f(\theta) &= \frac{1}{2n}(X\theta - y)^T(X\theta - y) \\ &= \frac{1}{2n}(\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta + y^T y) \\ &= \frac{1}{2}(\theta^T H\theta - \theta^T H\theta^* - \theta^{*T} H^T \theta + \frac{1}{n} y^T y) \end{aligned} \tag{1}$$

where we used that

$$H\theta^* = \frac{1}{n}X^T y \tag{2}$$

Hence,

$$\begin{aligned} f(\theta^*) &= \frac{1}{2}(\theta^{*T} H\theta^* - \theta^{*T} H\theta^* - \theta^{*T} H^T \theta^* + \frac{1}{n} y^T y) \\ &= \frac{1}{2}(-\theta^{*T} H\theta^* + \frac{1}{n} y^T y) \end{aligned} \tag{3}$$

The difference $f(\theta) - f(\theta^*)$ leads to the result.

It is also possible to prove the result by stating that it is the Taylor expansion of order 2 of f in θ^*, and using the fact that the third order derivatives of f are always 0.

1.0.2 Step 2

$$\begin{aligned} \forall t,\ \theta_t &= \theta_{t-1} - \gamma \nabla_\theta f(\theta_{t-1}) \\ &= \theta_{t-1} - \gamma \frac{1}{n} X^T (X\theta_{t-1} - y) \\ &= \theta_{t-1} - \gamma H(\theta_{t-1} - \theta^*) \end{aligned} \tag{4}$$

1.0.3 *Step 3*

$$\begin{aligned}\forall t,\ \theta_t - \theta^* &= \theta_{t-1} - \theta^* - \gamma H(\theta_{t-1} - \theta^*) \\ &= (I - \gamma H)(\theta_{t-1} - \theta^*)\end{aligned} \tag{5}$$

By applying 5 t times starting from $t = 0$, we get

$$\theta_t - \theta^* = (I - \gamma H)^t(\theta_0 - \theta^*) \tag{6}$$

1.0.4 *Step 4*

EIGENVALUES OF $(I - \gamma H)^t$ Let X be an eigenvector of H with eigenvalue λ. Then,

$$\begin{aligned}(I-\gamma H)^{2t}X &= (I-\gamma H)^{2t-1}(I-\gamma H)X \\ &= (I-\gamma H)^{2t-1}(X-\gamma HX) \\ &= (I-\gamma H)^{2t-1}(1-\gamma\lambda)X \\ &= (1-\gamma\lambda)(I-\gamma H)^{2t-1}X \\ &= (1-\gamma\lambda)^{2t}X\end{aligned} \tag{7}$$

Hence $(1-\gamma\lambda)^{2t}$ is an eigenvalue of $(I-\gamma H)^{2t}$. However this does not show the inverse property. It is better to exploit the fact that H is symmetric and real, hence there exists $P \in GL_n(\mathbb{R})$ and a diagonal matrix D containing the eigenvalues of H, such that

$$H = PDP^{-1} \tag{8}$$

Hence

$$\begin{aligned}I - \gamma H &= I - \gamma PDP^{1} \\ &= P(I-\gamma D)P^{-1}\end{aligned} \tag{9}$$

and

$$\begin{aligned}(I-\gamma H)^{2t} &= \left(P(I-\gamma D)P^{-1}\right)^{2t} \\ &= P(I-\gamma D)P^{-1}P(I-\gamma D)P^{-1}\ldots P(I-\gamma D)P^{-1} \\ &= P(I-\gamma D)^{2t}P^{-1}\end{aligned} \tag{10}$$

But $(I-\gamma D)^{2t}$ is a diagonal matrix with values of the form $(1-\gamma\lambda)^{2t}$ on the diagonal. We can conclude that the eigenvalues of $(I-\gamma D)^{2t}$ are exactly the $(1-\gamma\lambda)^{2t}$.

BOUNDING THE EIGENVALUES OF $(I - \gamma H)^t$ If λ is an eigenvalue of H, then

$$\mu \leqslant \lambda \leqslant L \tag{11}$$

Hence

$$1 - \gamma L \leqslant 1 - \gamma\lambda \leqslant 1 - \gamma\mu \tag{12}$$

And

$$-(1-\gamma\mu) \leqslant -(1-\gamma\lambda) \leqslant -(1-\gamma L) \tag{13}$$

We have $|1-\gamma\lambda| = \max\Big((1-\gamma\lambda), -(1-\gamma\lambda)\Big)$.
With 12,

$$(1-\gamma\lambda) \leqslant 1-\gamma\mu \leqslant |1-\gamma\mu| \tag{14}$$

With 13,

$$-(1-\gamma\lambda) \leqslant 1-\gamma L \leqslant |1-\gamma L| \tag{15}$$

Finally,

$$|1-\gamma\lambda| \leqslant \max\left(|1-\gamma\mu|, |1-\gamma L|\right) \tag{16}$$

With $\gamma = \frac{1}{L}$,

— $|1-\gamma\mu| = |1-\frac{\mu}{L}| = (1-\frac{\mu}{L})$

— $|1-\gamma L| = 0$.

and

$$\max_{\lambda\in[\mu,L]} |1-\gamma\lambda| \leqslant (1-\frac{\mu}{L}) = (1-\frac{1}{\kappa}) \tag{17}$$

and all absolute values of the eigenvalues of $(I-\gamma H)^t$ are smaller than $(1-\frac{1}{\kappa})^t$.

CONCLUSION We note $A = (I-\gamma H)^t$

We note $e = (e_i)_{i\in[1,d]}$ an orthonormal basis of eigenvectors of A, with eigenvalues μ_i. Using the previous results, we already know that

$$\forall i, |\mu_i| \leqslant (1-\frac{1}{\kappa})^t \tag{18}$$

As e is a basis of $\mathbb{R}^d$, there exists $(\alpha_1, \dots, \alpha_d) \in \mathbb{R}^d$ such that :

$$\theta_0 - \theta^* = \sum_{i=1}^{d} \alpha_i e_i \tag{19}$$

Using 6 and 19,

$$\begin{aligned} \theta_t - \theta^* &= A(\theta_0 - \theta^*) \\ &= A\sum_{i=1}^{d} \alpha_i e_i \\ &= \sum_{i=1}^{d} \alpha_i A e_i \\ &= \sum_{i=1}^{d} \alpha_i \mu_i e_i \end{aligned} \tag{20}$$

By properties of orthonormal bases, we have that

$$||\theta_0 - \theta^*||^2 = \sum_{i=1}^{d} \alpha_i^2 \tag{21}$$

and

$$\begin{aligned} ||\theta_t - \theta^*||^2 &= \sum_{i=1}^{d} \alpha_i^2 \mu_i^2 \\ &\leqslant max_i(|\mu_i|)^2 \sum_{i=1}^{d} \alpha_i^2 \end{aligned} \tag{22}$$

Using 18 and 21, we conclude that

$$||\theta_t - \theta^*||^2 \leqslant (1-\frac{1}{\kappa})^{2t} ||\theta_0 - \theta^*||^2 \tag{23}$$