Faisons un point maintenant… on a :

- Des données d'entrée x_i liées à des sorties y_i via un modèle $h_\Theta(x_i)$
- Les x_i, y_i sont fixés et les paramètres Θ sont notre inconnue
- On considère l'erreur de prédiction : $e_{i,\Theta} = h_\Theta(x_i) - y_i$

Faisons un point maintenant… on a :

- Des données d'entrée x_i liées à des sorties y_i via un modèle $h_\Theta(x_i)$
- Les x_i, y_i sont fixés et les paramètres Θ sont notre inconnue
- On considère l'erreur de prédiction : $e_{i,\Theta} = h_\Theta(x_i) - y_i$

On fait l'hypothèse : $e_{i,\Theta} \sim \mathcal{N}(0,\sigma)$

et on maximise la vraisemblance : $L(\Theta) = \prod_{i=1}^{n} f(x_i, y_i; \Theta, \sigma)$

Avec $f(x_i, y_i; \Theta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_{i,\Theta}^2}{2\sigma^2}\right)$

Faisons un point maintenant… on a :

- Des données d'entrée x_i liées à des sorties y_i via un modèle $h_\Theta(x_i)$
- Les x_i, y_i sont fixés et les paramètres Θ sont notre inconnue
- On considère l'erreur de prédiction : $e_{i,\Theta} = h_\Theta(x_i) - y_i$

On fait l'hypothèse : $e_{i,\Theta} \sim \mathcal{N}(0,\sigma)$

et on maximise la vraisemblance : $L(\Theta) = \prod_{i=1}^{n} f(x_i, y_i; \Theta, \sigma)$

Avec $f(x_i, y_i; \Theta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_{i,\Theta}^2}{2\sigma^2}\right)$

Plutôt que d'optimiser la vraisemblance par rapport à σ, nous allons fixer σ et optimiser :

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right)$$

Remarque : Il sera aussi possible dans certain cas d'optimiser simultanément les paramètres du modèle d'erreur (ici σ) et du modèle prédictif (ici Θ)

Remarque importante : Quand une hypothèse de normalité est faite sur les erreurs, c'est à dire que $h_\Theta(x_i) - y_i = e_{i,\Theta} \sim \mathcal{N}(0,\sigma)$), la maximisation de la vraisemblance est

$$\hat{\theta} = \arg\max_\theta \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right)$$

Il est intéressant de remarquer que ce problème d'optimisation peut être ré-écrit comme :

$$\begin{aligned}
\hat{\theta} &= \arg\max_\theta \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right) \\
&= \arg\max_\theta \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} e_i^2\right) \\
&= \arg\min_\theta \sum_{i=1}^{n} e_i^2 \\
&= \arg\min_\theta \sum_{i=1}^{n} (y_i - h_\Theta(x_i))^2
\end{aligned}$$

On a alors

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right)$$

$$= \arg\min_{\theta} \sum_{i=1}^{n} (y_i - h_\Theta(x_i))^2$$

Cette re-formulation est très pratique car :

- Elle peut être résolue de manière analytique si h_Θ est linéaire
- Elle ouvre la porte à une stratégie d'optimisation au sens des moindre carrés :

$$\nabla_\Theta e_i^2 = 2\left(y_i - h_\Theta(x_i)\right)\nabla_\Theta h_\Theta(x_i)$$

On constatera par la suite que les techniques de minimisation de l'erreur au sens des moindres carrés sont très courantes en apprentissage automatique. En faisant écho au maximum de vraisemblance, il faudra se souvenir qu'elles reposent sur une hypothèse de normalité des erreurs du modèle prédictif optimisé.