Introduction to Artificial Neural Networks

Filippo Gatti[†] [1]

[†]Laboratoire de Mécanique Paris-Saclay UMR 9026
Université Paris-Saclay, CentraleSupélec, ENS Paris-Saclay, CNRS

Laboratoire de Mécanique Paris-Saclay

1. filippo.gatti@centralesupelec.fr

Outline

Learning framework

Supervised learning : labelled database

$$\mathcal{D}_{XY} = \left\{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{X} \times \mathcal{Y}\right\}_{i=1}^{N},\ (\boldsymbol{x}_i, \boldsymbol{y}_i) \sim p \text{ and i.i.d. } p(\mathcal{D}_{XY}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i, \boldsymbol{y}_i)$$

Statistical model : parametric probability distribution

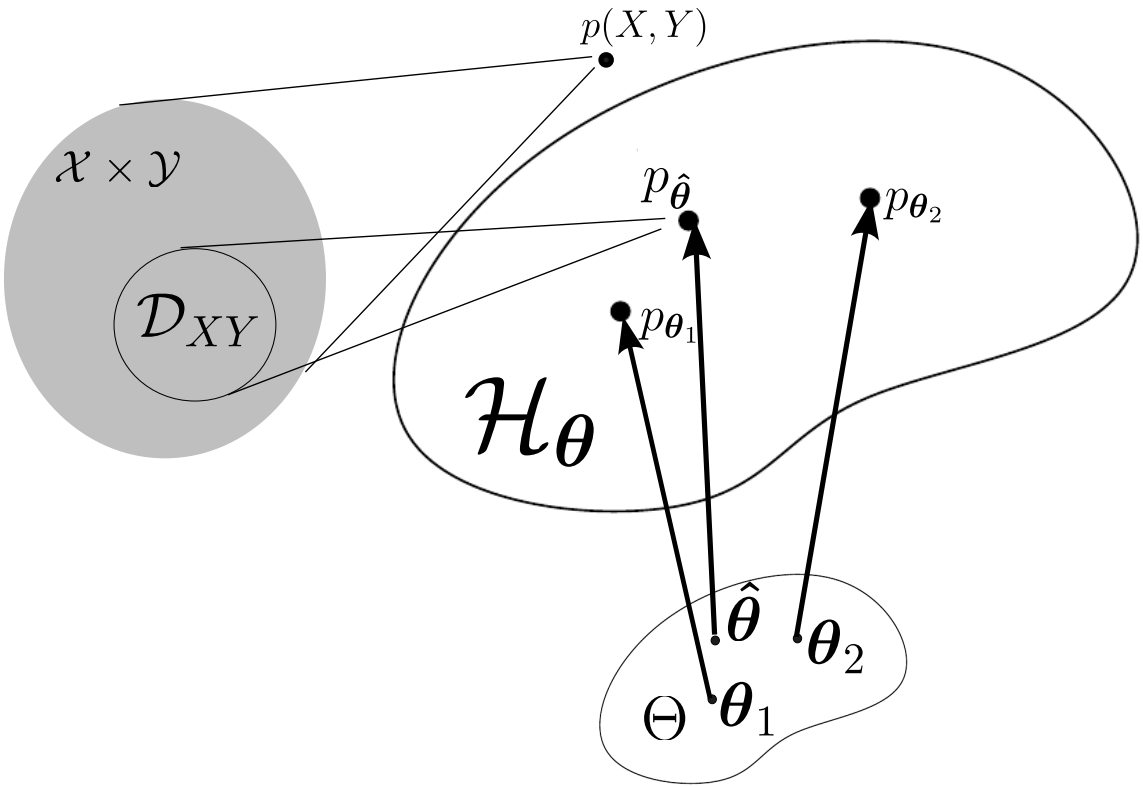
$$\mathcal{H}_\Theta := \left\{p_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{d_\Theta}\right\}$$

- Injective : $\exists! \boldsymbol{\theta}^* \iff \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 \longrightarrow p_{\boldsymbol{\theta}_1} = p_{\boldsymbol{\theta}_2}$
- Regular : $\exists g \in L^1_{loc}(\mathcal{X}) : \left|\frac{\partial p_{\boldsymbol{\theta}}}{\partial \theta_k}\right| \leq g$ μ-a.e. in open neighborhood $U \in N(\boldsymbol{x})$
- ? $\exists \boldsymbol{\theta}^* \in \Theta : p_{\boldsymbol{\theta}^*} = p$

Idea : find $\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}^*$ minimizing the *Negative Log-Likelihood* $\mathcal{NLL}$

$$\hat{\boldsymbol{\theta}}(\mathcal{D}_{XY}) = \underset{\boldsymbol{\theta} \in \Theta}{\arg\min}\ -\ln p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x}) = \underset{\boldsymbol{\theta} \in \Theta}{\arg\min}\ -\sum_{i=1}^{N} \ln p_{\boldsymbol{\theta}}(\boldsymbol{y}_i|\boldsymbol{x}_i), \qquad p_{\hat{\boldsymbol{\theta}}} \in \mathcal{H}_{\boldsymbol{\theta}}$$

Visualize the learning framework



Minimizing $\mathbb{E}\left[\mathcal{NLL}\right] \Longrightarrow$ minimizing Kullback-Leibler divergence $\mathbb{D}_{KL}\left(p \| p_{\boldsymbol{\theta}}\right) = \mathbb{E}_{\boldsymbol{x}\sim p}\left[\ln \frac{p}{p_{\boldsymbol{\theta}}}\right] < +\infty$

$$\min_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\boldsymbol{x}\sim p}\left[-\ln p_{\boldsymbol{\theta}}\left(\boldsymbol{X}\right)\right] = \min_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\boldsymbol{x}\sim p}\left[-\ln \frac{p_{\boldsymbol{\theta}}}{p}\right] \underbrace{-\mathbb{E}_{\boldsymbol{x}\sim p}\left[\ln p\right]}_{\geq 0} \geq \min_{\boldsymbol{\theta}\in\Theta} \mathbb{D}_{KL}\left(p \| p_{\boldsymbol{\theta}}\right)$$

Likelihood vs Log-Likelihood

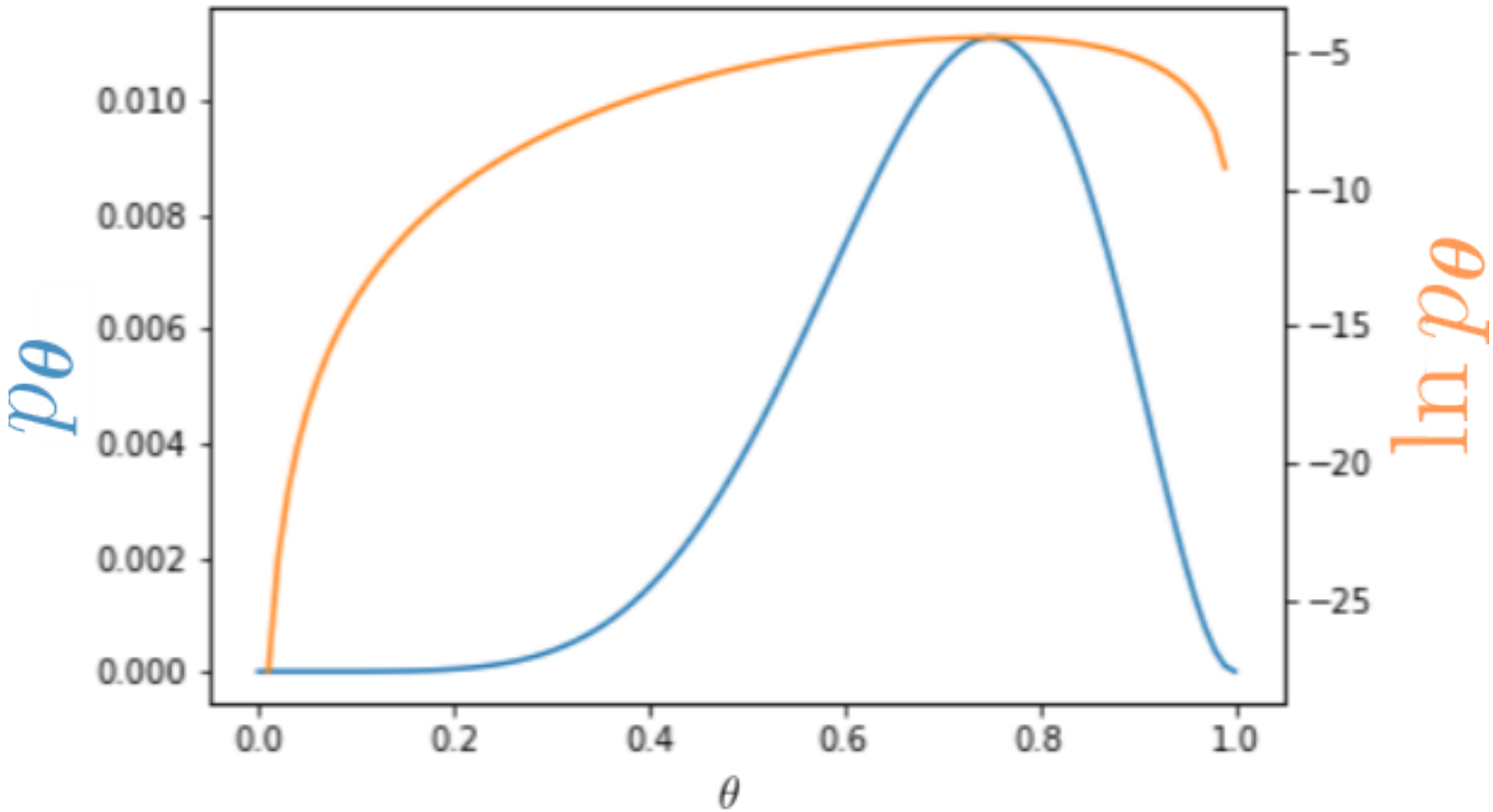


Figure – Bernoulli distribution $p_\theta(\mathcal{D}_X) = \prod_{i=1}^{N} \theta^{1_{x_i=1}} (1-\theta)^{1_{x_i=0}}$

`stats.stackexchange.com/questions/486007/measuring-predictive-uncertainty-with-negative-log-likelihood-nll`

Negative Log-Likelihood

For a "true" Gaussian distribution $p(Y|X) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y_i|x_i-\mu}{2\sigma^2}}$ with $\boldsymbol{\theta} = (\mu, \sigma)$:

$$\frac{\mathcal{NLL}(\mathcal{D}_{XY})}{N} = \frac{1}{2}\ln(2\pi) + \ln\sigma + \frac{1}{2\sigma^2 N}\sum_{i=1}^{N}(y_i|x_i - \mu)^2$$

Idea : use a model $y_i|x_i = h_{\boldsymbol{\theta}}(x_i)$

$$\frac{\mathcal{NLL}(\boldsymbol{\theta}; \mathcal{D}_{XY})}{N} = \frac{1}{2}\ln(2\pi) + \ln\sigma + \frac{1}{2\sigma^2 N}\sum_{i=1}^{N}\left(h_{\boldsymbol{\theta}}(x_i) - \mu\right)^2$$

with $\frac{1}{N}\sum_{i=1}^{N}\left(h_{\boldsymbol{\theta}}(x_i) - \mu\right)^2$ =MSE

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{\mathcal{NLL}(\boldsymbol{\theta}; \mathcal{D}_{XY})}{N} \rightarrow h_{\hat{\boldsymbol{\theta}}} = \mu$$

- Teacher (=source) sends a message to a pupil (=receiver)
- The message is meaningful if the receiver had no a priori knowledge of it
- Deterministic messages bare zero information
- Shannon's information = uncertainty or surprise associated to a random variable

$$\mathbb{I}(\omega) = f(P(\omega)) = -\log_b P(\omega) = \frac{-\ln P(\omega)}{\ln b} \geq 0, \quad b > 1$$

quantifying the "surprise"

- $\mathbb{I}(\omega) = 0$ if $P(\omega) = 1$
- $\mathbb{I}(\omega) = \infty$ if $P(\omega) = 0$
- $\mathbb{I}(\omega_1 \cap \omega_2) = \mathbb{I}(\omega_1) + \mathbb{I}(\omega_2)$ with $P(\omega_1 \cap \omega_2) = P(\omega_1) \cdot P(\omega_2)$

$$\mathbb{H}(p(\omega_1), \ldots, p(\omega_N)) = \cdot \sum_{\omega_i \in \omega}^{n} p(\omega_i) \cdot \mathbb{I}(\omega_i) = \mathbb{E}_{\omega_i \sim p_i}\left[\mathbb{I}(\omega)\right]$$

Example of Shannon's entropy

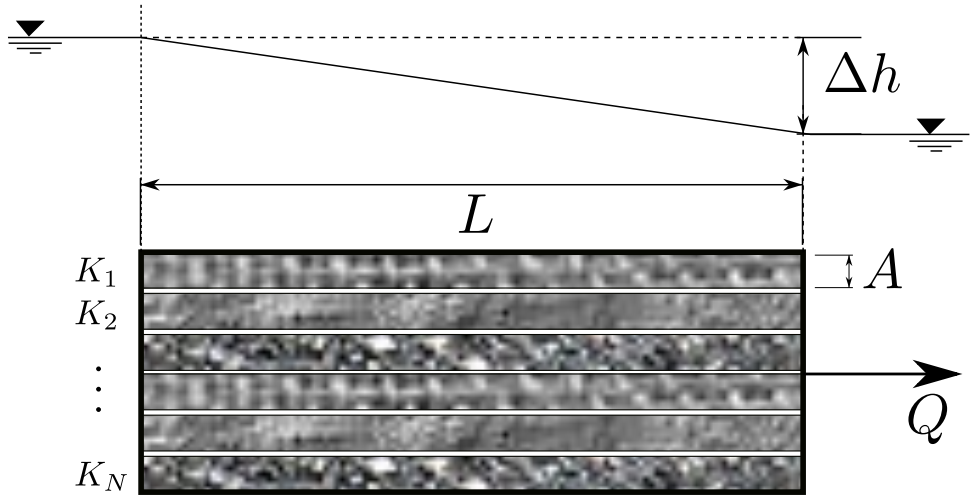


Figure – Darcy's law in a set of N parallel channels with random hydraulic conductivity K_i.

- Total discharge $Q = \left(\sum_{i=1}^{N} K_i\right) \frac{\Delta h \cdot A}{L}$
- Collection of values $\mathcal{D}_K = \{K_1, \ldots, K_b\}$, $\mathrm{card}(\mathcal{D}_K) = b$ and $p(\mathcal{D}_K) = (p_1, \ldots, p_b)$
- Pick N independent channels with $p_i = \frac{1}{b} : p\left(\{K_i\}_{i=1}^{N}\right) = b^{-N}$
- Entropy = average number of channel that deliver Q

$$\mathbb{E}_{n\sim p}[N] = -\mathbb{E}_{n\sim p}\left[\log_b p\left(\{K_i\}_{i=1}^{N}\right)\right] = \mathbb{H}(p_1, \ldots, p_N)$$

Example of Shannon's entropy

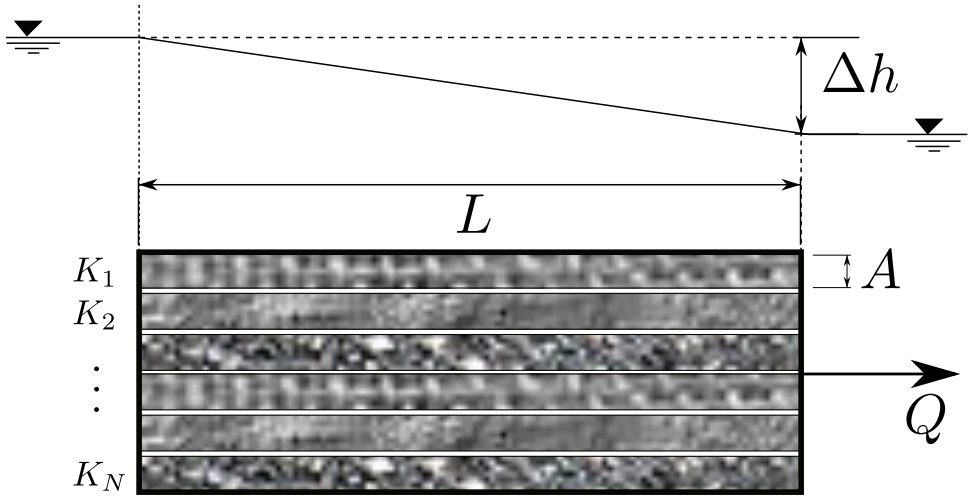


Figure – Darcy's law in a set of N parallel channels with random hydraulic conductivity K_i.

- Pick N independent channels with empirical frequency $f_i = \frac{b_i}{N}$, $\sum_{i=1}^{b} f_i = 1$

$$p\left(\{K_i\}_{i=1}^{N}\right) = \prod_{i=1}^{N} p_i^{b_i} = b^{N\left(\sum_{i=1}^{N} f_i \log_b p_i\right)} = e^{-N\cdot\left(\mathbb{D}_{KL}\left((f_1,\ldots,f_N)\|(p_1,\ldots,p_N)\right)+\mathbb{H}(f_1,\ldots,f_N)\ln b\right)}$$

- $b_i \sim \mathcal{B}(N, p_i) = \binom{N}{b_i} p_i^{b_i}(1-p_i)^{N-b_i}$, $\mathbb{E}_{b_i\sim\mathcal{B}(N,p_i)}\left[B_i\right] = N\cdot p_i \Longrightarrow \mathbb{E}_{f_i}\left[f_i\right] = p_i$
- $f_i \underset{N\to+\infty}{\longrightarrow} p_i \Longrightarrow p\left(\{K_i\}_{i=1}^{N}\right) \underset{N\to+\infty}{\longrightarrow} e^{-N\cdot\mathbb{H}(p(K))\ln b}$

Example of Shannon's entropy

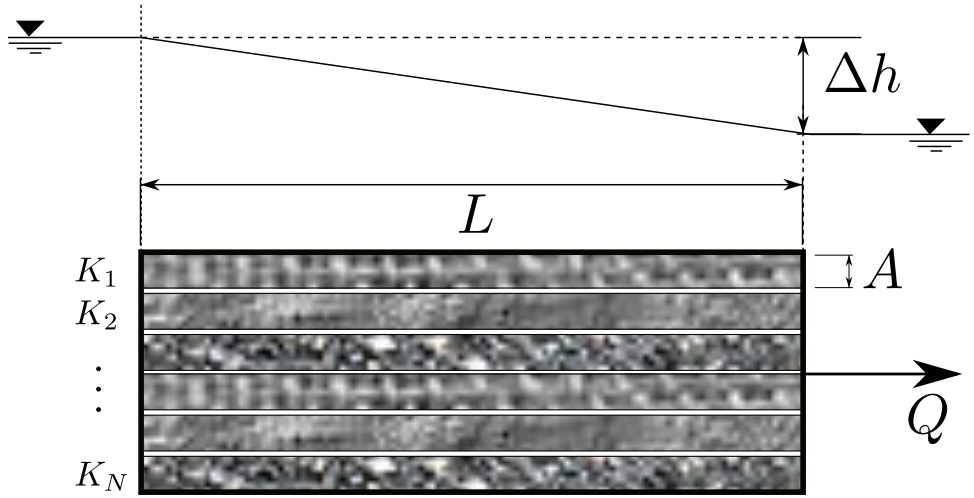


Figure – Darcy's law in a set of N parallel channels with random hydraulic conductivity K_i.

- possible combinations (with repetitions) of values of hydraulic conductivity
  $N_b = \frac{(N+b-1)!}{b!\cdot(N-1)!} \Longrightarrow Q = \left(\sum_{i=1}^{N} f_i K_i\right) \frac{\Delta h \cdot A}{L}$
- Q is attained $N_Q = \frac{N!}{\prod_{i=1}^{N} b_i!} \approx \frac{e^{\mathbb{H}(f_1,\ldots,f_N)}}{\sqrt{(2\pi N)^{N-1}\prod_{i=1}^{N} f_i}}$ unique times with probability
  $p(Q) = N_Q \cdot p\left(\{K_i\}_{i=1}^{N}\right) \approx \frac{e^{-N\cdot\mathbb{D}_{KL}\left((f_1,\ldots,f_N)\|(p_1,\ldots,p_N)\right)}}{\sqrt{(2\pi N)^{N-1}\prod_{i=1}^{N} f_i}} \underset{N\to+\infty}{\longrightarrow} 1 \approx N_Q \cdot e^{-N\cdot\mathbb{H}(p(K))\ln b}$
- $\frac{\ln N_Q}{N \ln b} \approx \mathbb{H}(p(K))$

Gibbs-Boltzmann theorem

$$\begin{cases} \max\limits_{q\in\mathcal{H}} \mathbb{H}_d(q) = \max\limits_{q\in\mathcal{H}} \left(-\int_{\mathcal{X}} \ln q\left(\boldsymbol{x}\right) \cdot q\left(\boldsymbol{x}\right) d\boldsymbol{x} \right) \\ c_k\left(q\right) = \mu_{y_k} - \int_{\mathcal{X}} y_k\left(\boldsymbol{x}\right) \cdot q\left(\boldsymbol{x}\right) d\boldsymbol{x} = 0, \quad c_k : \mathbb{R}^n \rightarrow \mathbb{R}, 1 \leq k \leq K \end{cases}$$

if it exists it is defined as $p_{\boldsymbol{\theta}} \in \mathcal{H}$, which reads :

$$p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = \underset{q\in\mathcal{H}, \boldsymbol{c}:\boldsymbol{c}=\boldsymbol{0}}{\arg\max}\ \mathbb{H}_d(q) = \frac{e^{-\sum_{k=1}^{K} y_k(\boldsymbol{x})\cdot\theta_k}}{Z}, \qquad \mathbb{H}_d\left(p_{\boldsymbol{\theta}}\right) \geq \mathbb{H}_d\left(p\right)$$

with $Z = \int_{\mathcal{X}} e^{-\sum_{k=1}^{K} \theta_k \cdot y_k(\boldsymbol{x})} \cdot \mu\left(d\boldsymbol{x}\right)$

$$\hat{\boldsymbol{\theta}} = \underset{\theta\in\Theta}{\arg\min}\ \max_{p_{\boldsymbol{\theta}}\in\mathcal{H}_\Theta}\ \mathbb{H}_d\left(p_{\boldsymbol{\theta}}\left(Y|\boldsymbol{X}\right)\right)$$

$$\mathcal{H}_{\boldsymbol{\theta}} \ni p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = \exp\left(-\left\langle \boldsymbol{y}\left(\boldsymbol{x}\right), \boldsymbol{\theta}\right\rangle\right)/Z$$

Log-Likelihood

$$\ell\left(\boldsymbol{\theta}; \boldsymbol{x}\right) = \ln p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = -\ln Z - \left\langle \boldsymbol{y}\left(\boldsymbol{x}\right), \boldsymbol{\theta}\right\rangle$$

$$\boldsymbol{s}_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}}\left[\boldsymbol{y}\left(\boldsymbol{X}\right)\right] - \boldsymbol{y}\left(\boldsymbol{x}\right) = \boldsymbol{\mu}_{\boldsymbol{y}} - \boldsymbol{y}\left(\boldsymbol{x}\right)$$

$$\mathbf{H}_{\ell}(\boldsymbol{\theta}; \boldsymbol{X}) = -\mathbb{C}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}}\left(\boldsymbol{y}\left(\boldsymbol{X}\right)\right)$$

- Unbiased score : $\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}}\left[\boldsymbol{s}_{\boldsymbol{\theta}}\left(\boldsymbol{X}\right)\right] = \mathbf{0} \longrightarrow$ stationarity
- Fisher Information Matrix (under regularity conditions)
  $\mathbb{I}_F(\boldsymbol{\theta}; \boldsymbol{X}) = -\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\mathbf{H}_{\ell}(\boldsymbol{\theta}; \boldsymbol{X})\right] = \mathbb{C}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}}\left(\boldsymbol{s}_{\boldsymbol{\theta}}\left(\boldsymbol{X}\right)\right)$
- $\exists(\hat{\boldsymbol{\theta}}_n)_{n\in\mathbb{N}}$ convergent in $\mathcal{N}\left(\mathbf{0}, \mathbb{I}_F^{-1}\left(\boldsymbol{\theta}^*; \boldsymbol{X}\right)\right) \in \mathcal{H}_{\boldsymbol{\theta}}$

$$\boldsymbol{\theta}^* = \lim_{n\to\infty} \hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \max_{p_{\boldsymbol{\theta}}\in\mathcal{H}_{\Theta}} \mathbb{H}_d\left(p_{\boldsymbol{\theta}}\left(Y|\boldsymbol{X}\right)\right)$$

Sketching the Fisher Information

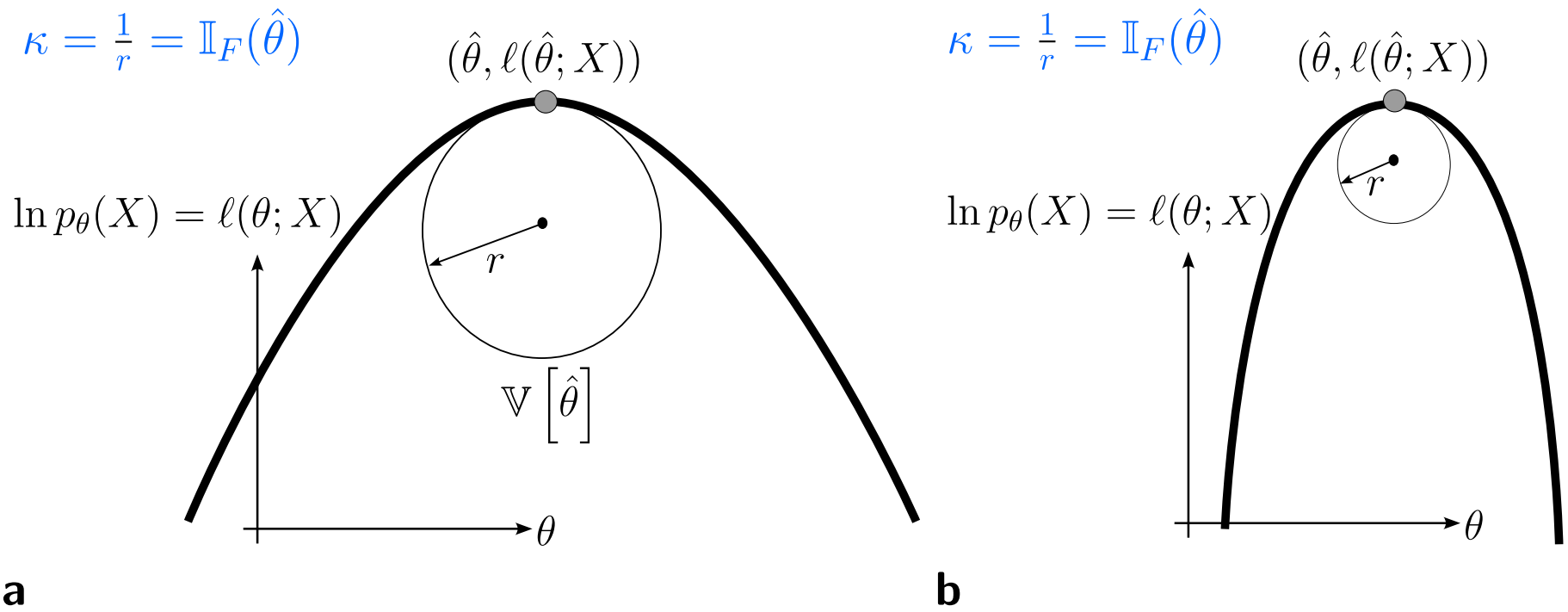


Figure – (a) Small FI/curvature (flat peak) → large variance (small accuracy); (b) Large FI/curvature (sharp peak) → small variance (high accuracy). Inspired by F. Nielsen presentation on Introduction to Geometry of Information

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}}\left[\ell(\boldsymbol{\theta};\boldsymbol{X})\right] = \mathbb{E}_{\boldsymbol{x}\sim p_{\hat{\boldsymbol{\theta}}}}\left[\ell(\hat{\boldsymbol{\theta}};\boldsymbol{X})\right] + \cancelto{0}{\left\langle \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}}\left[\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{X})\right], \left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)\right\rangle} - \frac{1}{2}\left\langle \boldsymbol{\theta}-\hat{\boldsymbol{\theta}}, \mathbb{I}_F\left(\boldsymbol{\theta}^*;\boldsymbol{X}\right).\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)\right\rangle$$

Empirical Loss Minimization ($\mathcal{P}$)

Given a labeled dataset $\mathcal{D}_{XY} = \left\{\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right)\right\}_{i=1}^{N} \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^{d_X + d_Y}$

Find $\boldsymbol{h}_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$, $\boldsymbol{h}_{\boldsymbol{\theta}} \in \mathcal{H}_{\boldsymbol{\theta}}$ such that : $\boldsymbol{h}_{\hat{\boldsymbol{\theta}}}\left(\mathcal{D}_{XY}\right) = \underset{\boldsymbol{h}_{\boldsymbol{\theta}} \in \mathcal{H}_{\boldsymbol{\theta}}}{\arg\min}\, L_{\mathcal{D}_{XY}}(\boldsymbol{h}_{\boldsymbol{\theta}})$

$$L_{\mathcal{D}_{XY}}\left(\boldsymbol{h}_{\boldsymbol{\theta}}\right) = \frac{1}{N}\sum_{i=1}^{N} \ell\left(\boldsymbol{h}_{\theta}\left(\boldsymbol{x}_k\right), \boldsymbol{y}_k\right), \quad \ell\left(\boldsymbol{y}, \boldsymbol{y}\right) = 0$$

Empirical Loss

- $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^{+}$: "distance" between "true" label and prediction
- $\ell\left(y, h_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\right) = -\ln p_{\boldsymbol{\theta}}\left(y|\boldsymbol{x}\right) = \mathcal{NLL}\left(y, h_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\right)$
- $L_{\mathcal{D}_{XY}}$ is an approximation of the "true" loss $L = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p}\left[\ell\left(\boldsymbol{h}_{\theta}\left(\boldsymbol{X}\right), \boldsymbol{Y}\right)\right]$, but $p(\boldsymbol{X}, \boldsymbol{Y})$ is not known *a priori*

Loss landscape : `https://losslandscape.com/explorer`

Optimize the statistical model

Existence of minimizing sequence (Gibbs-Boltzmann theorem)

$\exists(\hat{\boldsymbol{\theta}}_n)_{n\in\mathbb{N}}$ convergent in $\mathcal{N}\left(\mathbf{0}, \mathbb{I}_F^{-1}\left(\boldsymbol{\theta}^*; \boldsymbol{X}\right)\right) \in \mathcal{H}_{\boldsymbol{\theta}}$

"Delta rule"

The quest for $\hat{\boldsymbol{\theta}}$ is iteratively conducted, following the direction of $-\boldsymbol{\nabla}_{\boldsymbol{\theta}} L_{\mathcal{D}_{XY}}$, according to the following update rule from iteration i to iteration $+i+1$ (the so called *delta rule* :

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \eta^{(i)} \boldsymbol{\nabla}_{\boldsymbol{\theta}} L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(i)}\right), \quad \eta^{(i)} \in \mathbb{R}^+$$

with $\eta^{(i)} \in \mathbb{R}^+$ being the *learning rate*, and

$$L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(i)}\right) \geq L_{\mathcal{D}_{XY}}\left(\hat{\boldsymbol{\theta}}\right)$$

Euler's inequality : Necessary condition for minimizer on convex sets

Given a function $f : K \to \mathbb{R}$ defined over a non-empty convex set $K \subset H$ (Hilbert), f proper on K and $f \in C^1(K)$,

$$\left\langle \nabla_x f(\hat{\boldsymbol{x}}), \boldsymbol{y} - \hat{\boldsymbol{x}} \right\rangle \geq 0, \quad \forall \boldsymbol{y} \in K \quad \Longrightarrow \quad \hat{\boldsymbol{x}} \text{ local minimizer of } f \text{ on } K$$

Why Gradient Descent ?

- Taylor expansion : $T_{\hat{\boldsymbol{\theta}}} L_{\mathcal{D}_{XY}}(\boldsymbol{\theta}) = L_{\mathcal{D}_{XY}}\left(\hat{\boldsymbol{\theta}}\right) + \left\langle \nabla_\theta L_{\mathcal{D}_{XY}}\left(\hat{\boldsymbol{\theta}}\right), \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right\rangle + o\left(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|\right)$
- Euler's inequality holds if $\left\langle \nabla_\theta L_{\mathcal{D}_{XY}}\left(\hat{\boldsymbol{\theta}}\right), \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right\rangle \geq 0, \quad \forall \boldsymbol{\theta} \in \Theta$
- $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}} + \eta \nabla_\theta L_{\mathcal{D}_{XY}}\left(\hat{\boldsymbol{\theta}}\right)$
  - $T_{\hat{\boldsymbol{\theta}}} L_{\mathcal{D}_{XY}}(\boldsymbol{\theta}(\eta)) \geq L_{\mathcal{D}_{XY}}\left(\hat{\boldsymbol{\theta}}\right), \quad \eta > 0$
  - $\xi : (\eta, \boldsymbol{\theta}) \mapsto L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta} - \eta \cdot \nabla_\theta L_{\mathcal{D}_{XY}}(\boldsymbol{\theta})\right)$
  - $T_0 \xi(\eta; \boldsymbol{\theta}) = L_{\mathcal{D}_{XY}}(\boldsymbol{\theta}) - \eta \| \nabla_x L_{\mathcal{D}_{XY}}(\boldsymbol{\theta}) \|^2 + o(\eta)$

$$-\frac{\nabla_\theta L_{\mathcal{D}_{XY}}(\boldsymbol{\theta})}{\| \nabla_\theta L_{\mathcal{D}_{XY}}(\boldsymbol{\theta}) \|} = \frac{1}{r} \underset{\|\delta\boldsymbol{\theta}\| = r}{\arg\min}\, L_{\mathcal{D}_{XY}}(\boldsymbol{\theta} + \delta\boldsymbol{\theta})$$

Convergence analysis of functions with Lipschitz gradients

A function $f : K \to \mathbb{R}$, non-empty convex set $K \subset H$ (Hilbert), f proper and strictly convex on K, $f \in C^1(K)$ and a gradient $\nabla_x f \in \mathrm{Lip}(K)$ of constant $\beta > 0$. $\exists (\boldsymbol{x}_k)_{k\in\mathbb{N}} \in K$ such that

$$\eta_k \in \left[\frac{1}{\beta}, \frac{2}{\beta}\right[\qquad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \nabla_x f(\boldsymbol{x}_k)$$

$$\exists A > 0 \text{ such that : } f\left(\boldsymbol{x}_k\right) - f(\hat{\boldsymbol{x}}) \leq \frac{A}{k+1} \to \lim_{k\to\infty} \boldsymbol{x}_k = \hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in K} f(\boldsymbol{x})$$

If f is strongly convex of coefficient α, $\exists \rho = \sqrt{\frac{\beta}{\alpha+\beta}} \leq \sqrt{\frac{1}{1+\kappa(\mathbf{H}_f)}}$:

$$\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}\| \leq \rho^k \|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}\|, \quad \forall k \in \mathbb{N}$$

$$\kappa\left(\mathbf{H}_f(\boldsymbol{x})\right) = \frac{\max\limits_{1\leq i\leq d_K} \lambda_i(\mathbf{H}_f(\boldsymbol{x}))}{\min\limits_{1\leq i\leq d_K} \lambda_i(\mathbf{H}_f(\boldsymbol{x}))} \text{ and } 0 < \kappa\left(\mathbf{H}_f(\boldsymbol{x})\right) \leq \frac{\beta}{\alpha}$$

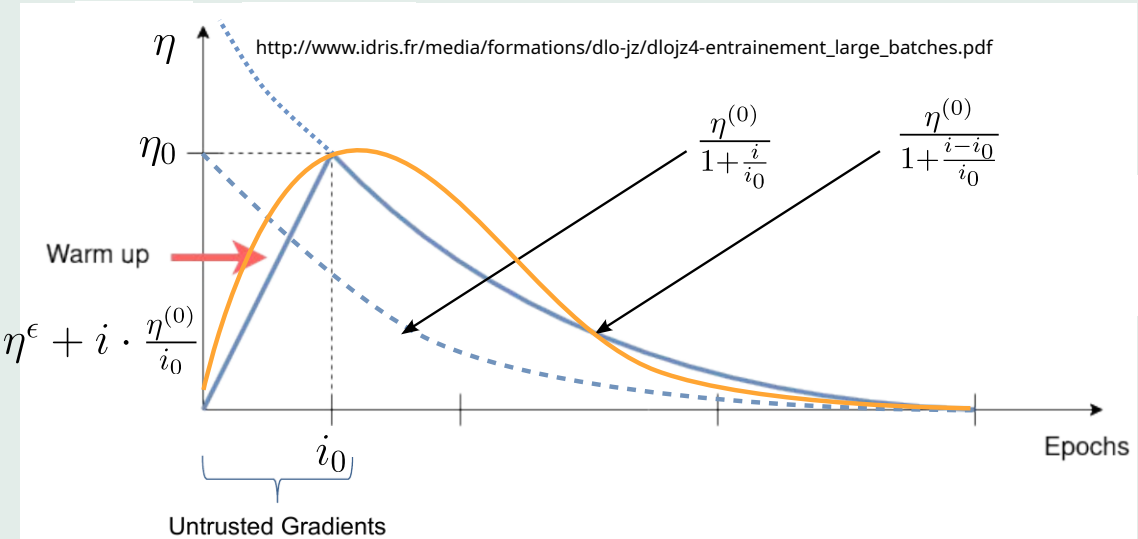
Learning rate scheduler

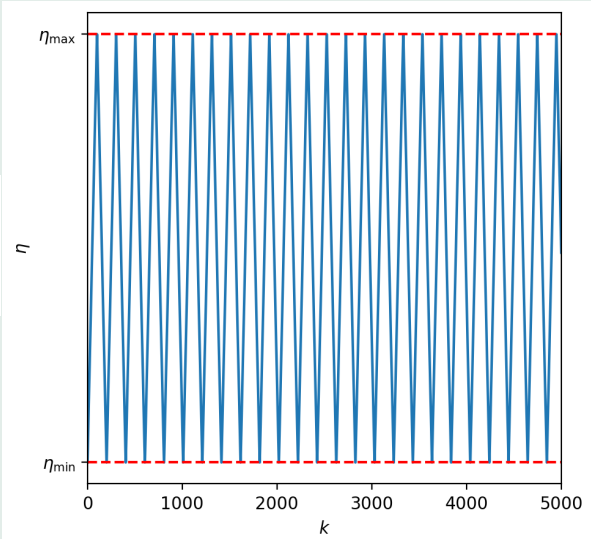
Classical scheduling

$$\eta^{(i)} = \frac{\eta^{(0)}}{1 + i \cdot \eta_d} \leq \eta^{(0)}$$

η_d learning rate decay (often $\eta_d = \frac{1}{i_0}$, $i_0 > 0$)

Warm-up and cyclic scheduler





Different learning rate schedulers

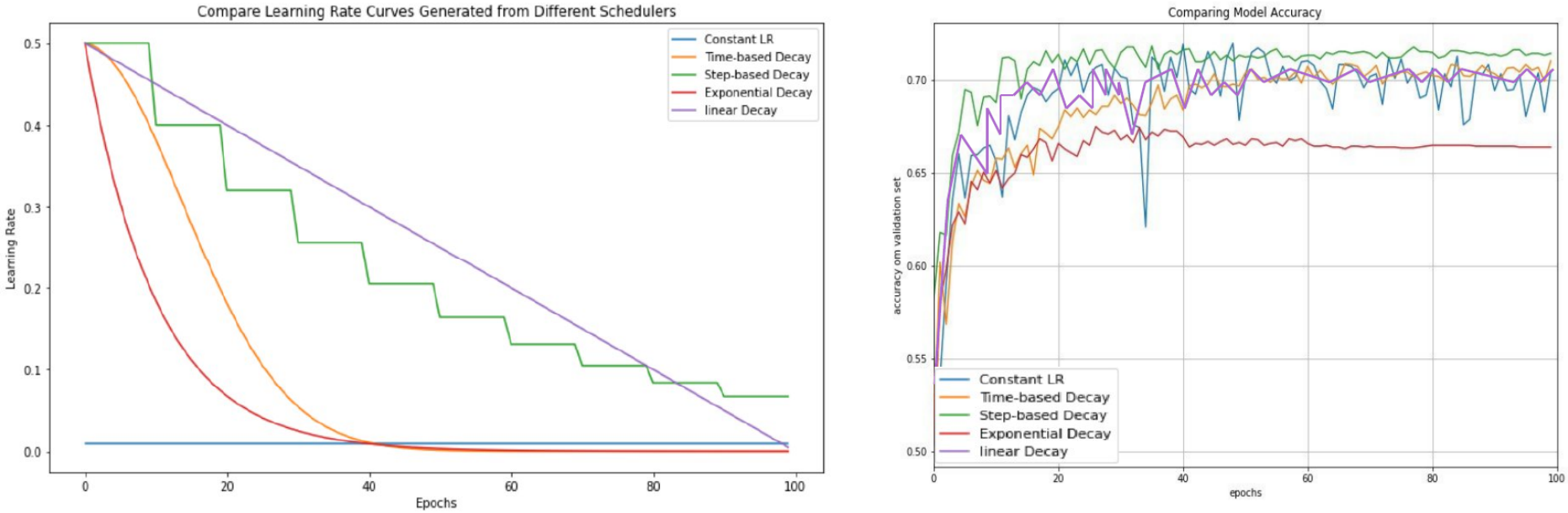


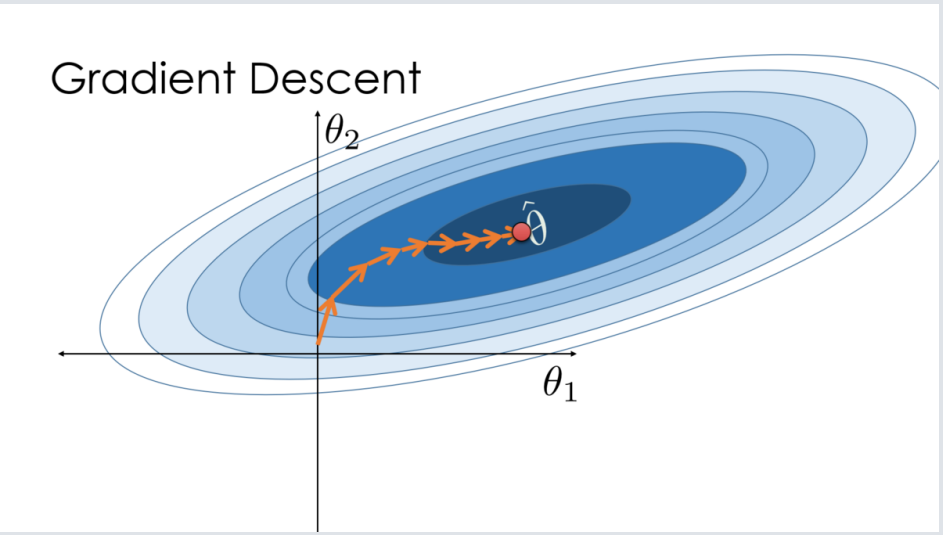
Figure – Courtesy of IDRIS : “Deep Learning Optimisé - Jean Zay”

Computing the gradient

$$\nabla_\theta L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(i)}\right) = \frac{1}{N} \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{D}_{XY}} \nabla_\theta \ell\left(\boldsymbol{h}_\theta\left(\boldsymbol{x}_i\right), \boldsymbol{y}_i\right)$$

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \frac{\eta^{(i)}}{N} \sum_{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{D}_{XY}} \nabla_\theta \ell\left(\boldsymbol{h}_\theta\left(\boldsymbol{x}_i\right), \boldsymbol{y}_i\right)$$

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \eta^{(i)} \nabla_{\boldsymbol{\theta}} \ell\left(\boldsymbol{h}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right), \boldsymbol{y}_i\right)$$



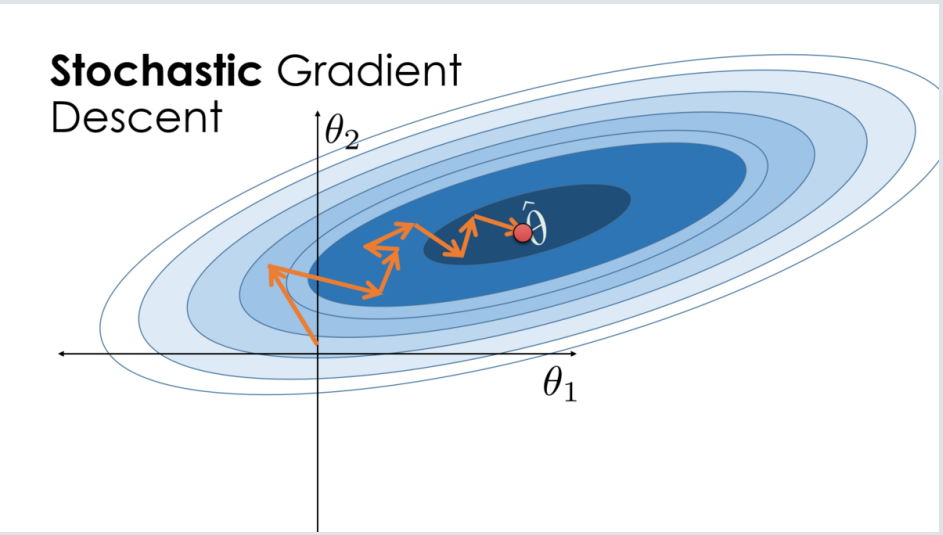


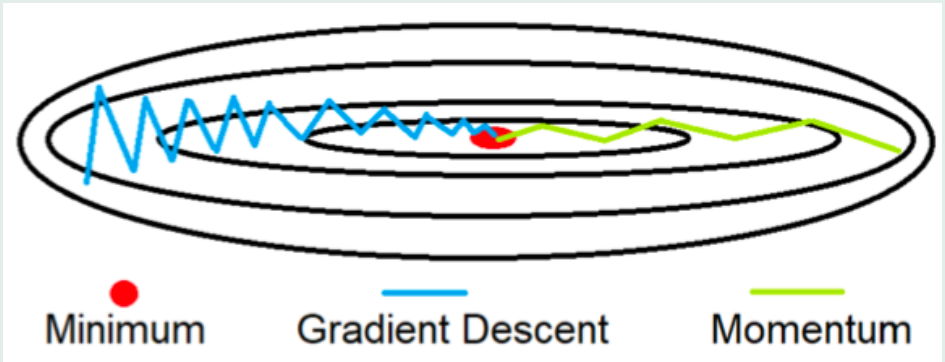
Figure – Lau, S., Gonzalez, J., Nolan, D. (2020)
`https://www.textbook.ds100.org/ch/11/gradient_stochastic.html`,
`https://optimization.cbe.cornell.edu`

Use mini-batches

Algorithm – Stochastic Gradient Descent on mini-batches

$i = 0$
Initialize $\boldsymbol{\theta}^{(0)}$
$\eta^{(0)} = \eta_0$
while $i < n_e$ **do**
  for $j \sim \mathcal{U}\left(\{i\}_{i=1}^{\frac{N}{N_b}}\right)$ **do**
    $\boldsymbol{g}_{\boldsymbol{\theta}} = 0$
    for $t \sim \mathcal{U}\left(\{j_i\}_{i=1}^{N_b}\right)$ **do**
      $\boldsymbol{g}_{\boldsymbol{\theta}} += \nabla_{\boldsymbol{\theta}} \ell\left(\boldsymbol{h}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right), \boldsymbol{y}^{(t)}; \boldsymbol{\theta}^{(i)}\right)$
    end for
    $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \frac{\eta^{(i)}}{N_b} \boldsymbol{g}_{\boldsymbol{\theta}}, \quad \eta^{(i)} \in \mathbb{R}^+$
  end for
end while

The role of momentum : "pushing a ball down a hill"



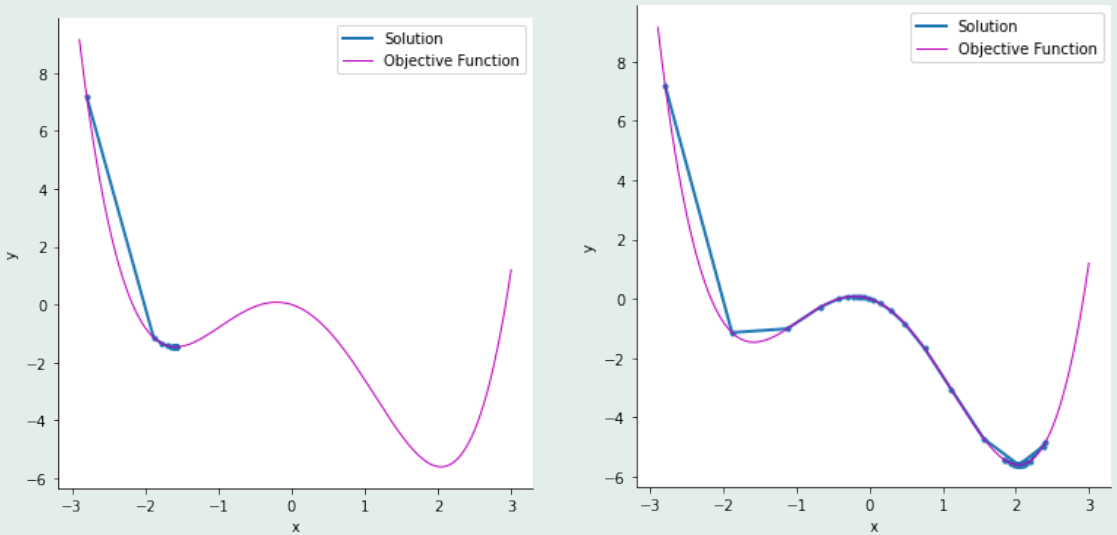


Figure – Srihari, S. (n.d.). Basic Optimization Algorithms. Deep learning. `https://cedar.buffalo.edu/~srihari/CSE676/8.3%20BasicOptimizn.pdf`, `https://optimization.cbe.cornell.edu/index.php?title=Momentum`

The role of momentum

Algorithm – SGD on mini-batches with Classical Momentum

$i = 0$
Initialize $\boldsymbol{\theta}^{(0)}$
$\eta^{(0)} = \eta_0$
$\boldsymbol{v}^{(0)} = \mathbf{0}$
while $i < n_e$ **do**
  for $j \sim \mathcal{U}\left(\{i\}_{i=1}^{\frac{N}{N_b}}\right)$ **do**
    $\boldsymbol{g}_\theta = 0$
    for $t \sim \mathcal{U}\left(\{j_i\}_{i=1}^{N_b}\right)$ **do**
      $\boldsymbol{g}_\theta += \nabla_\theta \ell\left(\boldsymbol{h}_\theta\left(\boldsymbol{x}_i\right), \boldsymbol{y}^{(t)}; \boldsymbol{\theta}^{(i)}\right)$
    end for
    $\boldsymbol{m}^{(i+1)} = \left(1 - \langle \tau^{(i)} - 1 \rangle\right) \boldsymbol{g}_\theta + \gamma^{(i)} \boldsymbol{m}^{(i)}$
    $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \frac{\eta^{(i)}}{N_b} \boldsymbol{m}^{(i+1)}$
  end for
end while

$\eta^{(i)} \in \mathbb{R}^+$, $\gamma^{(i)} \in [0, 1)$ (usually $\gamma^{(i)} = 0.85$ or 0.95)

$-\langle \tau^{(i)} - 1 \rangle$ dampens the gradient only if $\tau^{(i)} > 1$ ("air resistance")

AdaGrad and RMSprop (Hinton's Lecture 6a-Coursera)

- Newton's method : $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \mathbf{H}^{-1}_{L_{\mathcal{D}_{XY}}}\left(\boldsymbol{\theta}^{(i)}\right).\boldsymbol{\nabla}_{\boldsymbol{\theta}} L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(i)}\right)$
- $\operatorname{diag}\left(\mathbb{I}_F(\boldsymbol{\theta};\boldsymbol{X})\right) = \operatorname{diag}\left(\mathbb{V}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left(s_{\boldsymbol{\theta},k}\left(\boldsymbol{X}\right)\right)\right) = \operatorname{diag}\left(\mathbf{H}_{L_{\mathcal{D}_{XY}}}\right)$
- Gradient memory : $\mathbb{G}^{(i)} = \sum_{k=0}^{i-1} \boldsymbol{\nabla}_{\boldsymbol{\theta}} L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(k)}\right) \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(k)}\right)$
- $\operatorname{Tr}\left(\mathbb{G}^{(i)}\right) = \sum_{k=1}^{i-1} \|\boldsymbol{\nabla}_{\boldsymbol{\theta}} L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(k)}\right)\|_2^2 \approx \mathbb{E}\left[\|\boldsymbol{\nabla}_{\boldsymbol{\theta}} L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}\right)\|^2\right]$

$$\text{AdaGrad} : \boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \eta\left(\operatorname{diag}\mathbb{G}^{(i)} + \epsilon\mathbb{I}\right)^{-\frac{1}{2}} \boldsymbol{\nabla}_{\boldsymbol{\theta}} L_{\mathcal{D}_{XY}}$$

- Average gradient memory on short window :
  $\operatorname{diag}\tilde{\mathbb{G}}^{(i)} = \alpha_r \operatorname{diag}\mathbb{G}^{(i-1)} + (1-\alpha_r)\left(\operatorname{diag}\mathbb{G}^{(i)} - \operatorname{diag}\mathbb{G}^{(i-1)}\right)$, $\alpha_r = 0.9$

$$\text{RMSprop} : \boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \eta\left(\operatorname{diag}\tilde{\mathbb{G}}^{(i)} + \epsilon\mathbb{I}\right)^{-\frac{1}{2}} \boldsymbol{\nabla}_{\boldsymbol{\theta}} L_{\mathcal{D}_{XY}}$$

ADADELTA

- AdaGrad's η decay too aggressive !
- Introduce momentum vectors $\boldsymbol{U}^{(i+1)} = \alpha_u \boldsymbol{U}^{(i)} + (1 - \alpha_u)\, \boldsymbol{u}^{(i)}$,
  $\mathbb{U}^{(i)} = \left(\boldsymbol{U}^{(i)} \otimes \boldsymbol{U}^{(i)} \right)^{\frac{1}{2}}$
- Compute vectors
  $\boldsymbol{u}^{(i+1)} = \text{diag}^{\frac{1}{2}} \left(\mathbb{U}^{(i)} + \epsilon \mathbb{I} \right) \text{diag}^{-\frac{1}{2}} \left(\tilde{\mathbb{G}}^{(i)} + \epsilon \mathbb{I} \right) . \boldsymbol{\nabla}_\theta L_{\mathcal{D}_{XY}} \left(\boldsymbol{\theta}^{(i)} \right)$
- Approximate Hessian : $\mathbf{H}_{L_{\mathcal{D}_{XY}}} \left(\boldsymbol{\theta}^{(i)} \right) \approx \text{diag}^{\frac{1}{2}} \left(\mathbb{U}^{(i)} + \epsilon \mathbb{I} \right)$

$$\text{ADADELTA} : \boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \boldsymbol{u}^{(i)}$$

- No need to set a default learning rate

Adam : "heavy ball with friction" [Heu+]

- Decay average of past gradients :
  $\boldsymbol{m}^{(i)} = \beta_1 \boldsymbol{m}^{(i-1)} + (1-\beta_1)\, \boldsymbol{\nabla}_\theta L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}^{(i)}\right)$
- Decay average of past squared gradients :
  $\mathrm{diag}\mathbb{V}^{(i)} = \beta_2 \mathbb{V}^{(i-1)} + (1-\beta_2)\, \mathrm{diag}\mathbb{G}^{(i)}$
- Normalization : $\hat{\boldsymbol{m}}^{(i)} = \frac{\boldsymbol{m}^{(i)}}{1-\beta_1}, \qquad \hat{\mathbb{V}}^{(i)} = \frac{\mathbb{V}^{(i)}}{1-\beta_2}$

$$\text{Adam} : \boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \eta \cdot \mathrm{diag}^{-\frac{1}{2}}\left(\hat{\mathbb{V}}^{(i)} + \epsilon \mathbb{I}\right) \hat{\boldsymbol{m}}^{(i)}$$

- $\beta_1 = 0.9, \beta_2 = 0.999$

Create a statistical model

Empirical loss minimization

Find $\boldsymbol{h}_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}, \boldsymbol{h}_{\boldsymbol{\theta}} \in \mathcal{H}_{\boldsymbol{\theta}}$ such that :

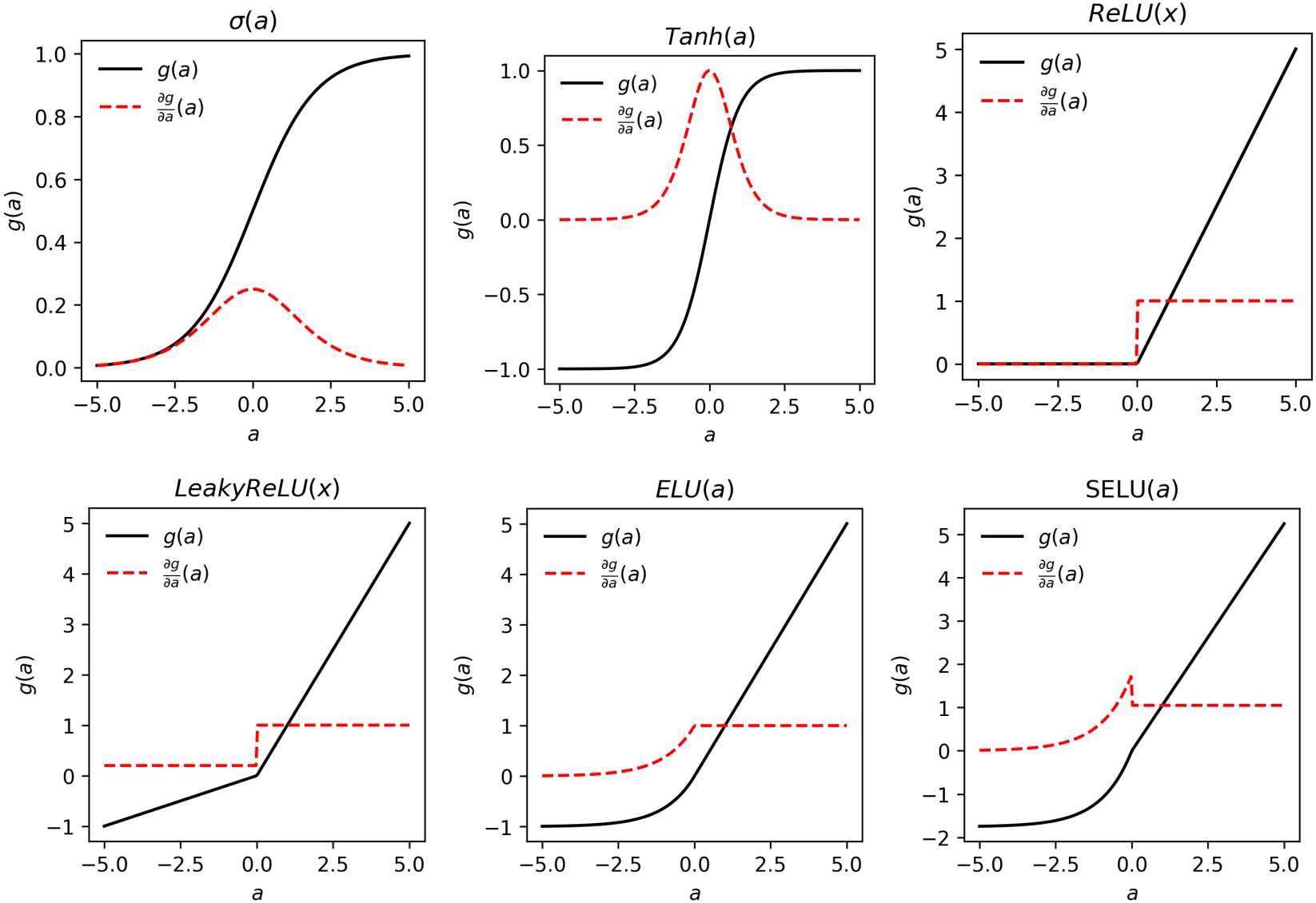
$$\boldsymbol{h}_{\hat{\boldsymbol{\theta}}}\left(\mathcal{D}_{XY}\right) = \underset{\boldsymbol{h}_{\boldsymbol{\theta}} \in \mathcal{H}_{\boldsymbol{\theta}}}{\arg\min}\, L_{\mathcal{D}_{XY}}(\boldsymbol{h}_{\boldsymbol{\theta}})$$

$$\ell\left(y, h_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\right) = -\ln p_{\boldsymbol{\theta}}\left(y|\boldsymbol{x}\right) = \mathcal{NLL}\left(y, h_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right)\right)$$

The simplest statistical model [MP43 ; Ros57]

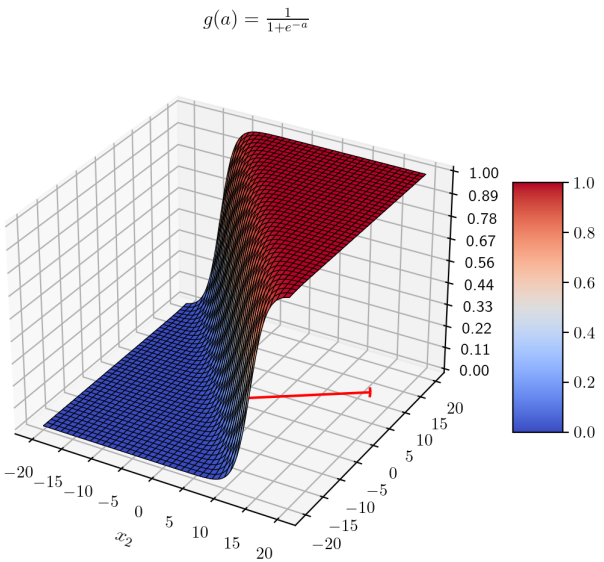
$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = g(a(\boldsymbol{x})) = g(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) = g\left(\sum_{c=1}^{d_X} w_c \cdot x_c + b\right); \qquad \theta := \{\boldsymbol{w}; b\}$$

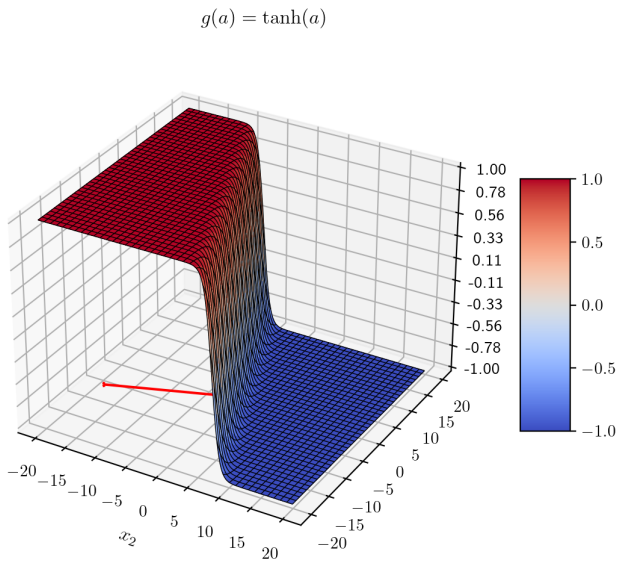
Activation functions



Decision boundary

- Binary classification $y \in \mathcal{Y} := \{0, 1\}$ (if $g(\cdot) = \sigma(\cdot)$) or $y \in \mathcal{Y} := \{-1, 1\}$ (if $g(\cdot) = \tanh(\cdot)$)
- hyperplane that separates the samples of class $y = 0$ (or $y = -1$) if $h_{\boldsymbol{\theta}}(\boldsymbol{x}) > 0.5$ from those labeled as $y = 1$, if $h_{\boldsymbol{\theta}}(\boldsymbol{x}) \leq 0.5$





Unsolved decision boundary

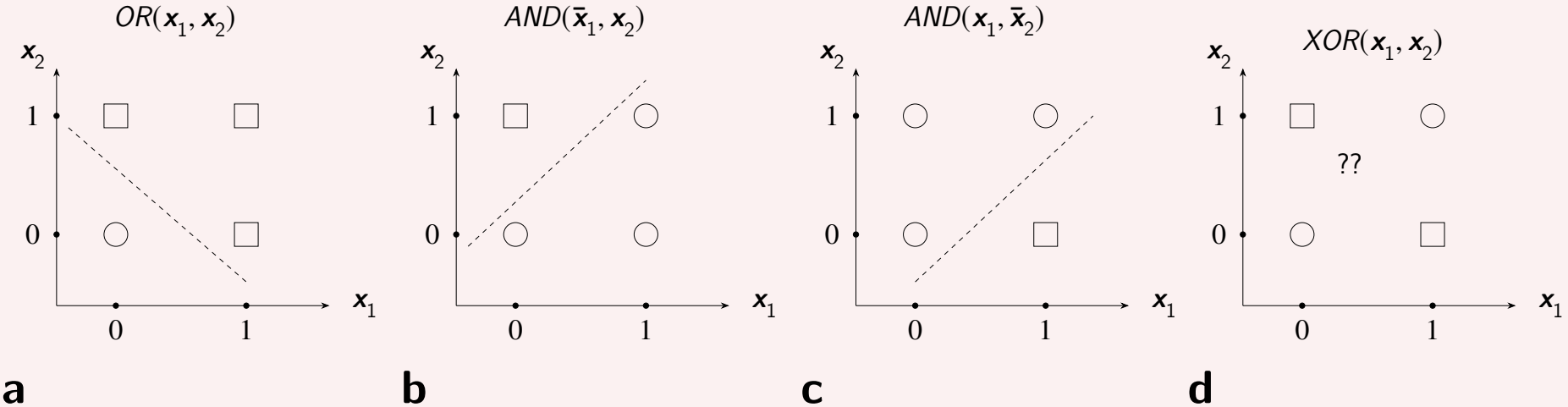


Figure – Solved and unsolved decision boundaries for different binary operations. (a) $f_{OR}(\boldsymbol{x}) = \{x_1\} \cup \{x_2\}$, (b) $f_{AND1}(\boldsymbol{x}) = \{x_1\} \cap \{\{0,1\}/\{x_2\}\}$, (c) $f_{AND2}(\boldsymbol{x}) = \{\{0,1\}/(x_1)\} \cap \{x_2\}$, (d) $f_{XOR}(\boldsymbol{x}) = \{\{x_1\} \cup \{x_2\}\}/\{\{x_1\} \cap \{x_2\}\}$ (unsolved). Reprinted from the video-lecture notes by Hugo Larochelle (`https://www.youtube.com/watch?v=iT5P4z6Fzj8&list=PL6Xpj9I5qXYEcOhn7TqghAJ6NAPrNmUBH&index=3`).

Add intermediate feature=hidden layers

$$h_{\theta}(\boldsymbol{x}) = g(a(\boldsymbol{\phi}(\boldsymbol{x}))) = g\left(\sum_{c=1}^{d_{\phi}} w_c \cdot \phi_c(\boldsymbol{x}) + b\right); \qquad \boldsymbol{\theta} := \{\boldsymbol{w}; b\}$$

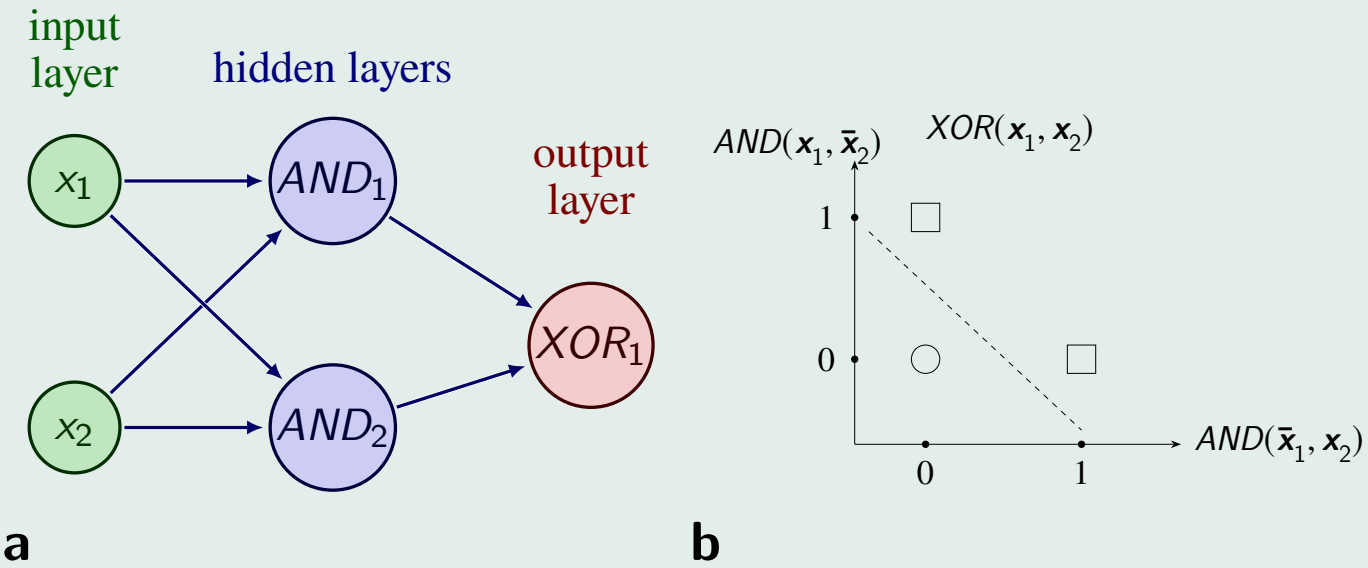


Figure – Simple example of Multi-Layer Perceptron (1 hidden layers $\boldsymbol{h}^{(1)}$ and 1 output $y \in \mathbb{R}$). layer. The figure was drawn with tikz.net.

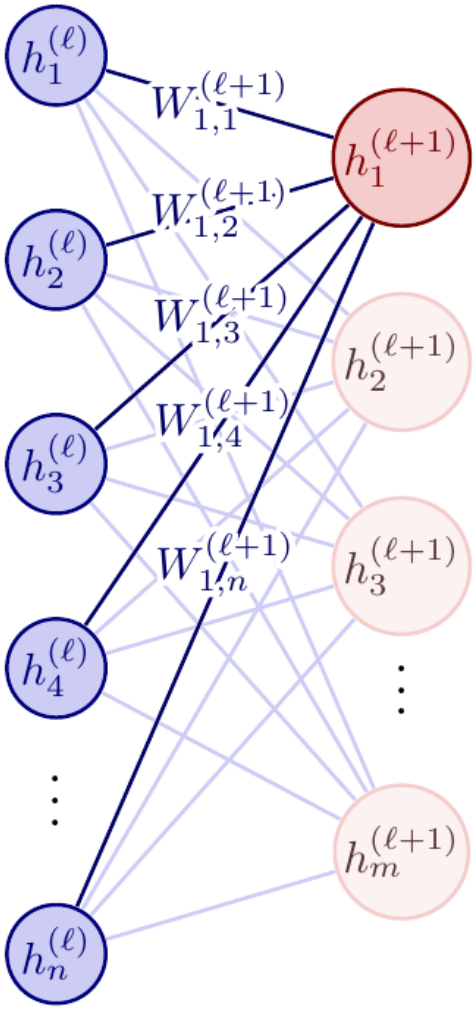
Stacking layers of neurons

$$h_\theta(\boldsymbol{x}) = h^{(o)} \circ \boldsymbol{h}^{(1)}(\boldsymbol{x}) = g^{(o)}(a^{(o)}(\boldsymbol{h}^{(1)}(\boldsymbol{x}))) = g\left(\left\langle \boldsymbol{w}^{(o)}, \boldsymbol{h}^{(1)}(\boldsymbol{x}) \right\rangle + b^{(o)}\right) =$$

$$= g^{(o)}\left(\sum_{c=1}^{d_{h_1}} w_c^{(o)} \cdot h_c^{(1)}(\boldsymbol{x}) + b^{(o)}\right); \quad \theta := \left\{\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(o)}; b^{(1)}, b^{(o)}\right\}$$

with

$$h_c^{(1)} = g^{(1)}(a_c^{(1)}(\boldsymbol{x})) = g^{(1)}\left(\left\langle \boldsymbol{w}^{(1)}, \boldsymbol{x} \right\rangle + b^{(1)}\right) = g^{(1)}\left(\sum_{c=1}^{d_X} w_c^{(1)} \cdot x_c + b^{(1)}\right)$$



$$= g\left(w_{1,1}^{(\ell+1)} h_1^{(\ell)} + w_{1,2}^{(\ell+1)} h_2^{(\ell)} + \ldots + w_{1,n}^{(\ell+1)} h_n^{(\ell)} + b_1^{(\ell+1)}\right)$$

$$= g^{(\ell+1)}\left(\sum_{i=1}^{n} w_{1,i}^{(\ell+1)} h_i^{(\ell)} + b_1^{(\ell+1)}\right)$$

$$\begin{pmatrix} h_1^{(\ell+1)} \\ h_2^{(\ell+1)} \\ \vdots \\ h_m^{(\ell+1)} \end{pmatrix} = \boldsymbol{g}^{(\ell+1)}\left[\begin{pmatrix} w_{1,1}^{(\ell+1)} & w_{1,2}^{(\ell+1)} & \cdots & w_{1,n}^{(\ell+1)} \\ w_{2,1}^{(\ell+1)} & w_{2,2}^{(\ell+1)} & \cdots & w_{2,n}^{(\ell+1)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,1}^{(\ell+1)} & w_{m,2}^{(\ell+1)} & \cdots & w_{m,n}^{(\ell+1)} \end{pmatrix}\begin{pmatrix} h_1^{(\ell)} \\ h_2^{(\ell)} \\ \vdots \\ h_n^{(\ell)} \end{pmatrix} + \begin{pmatrix} b_1^{(\ell+1)} \\ b_2^{(\ell+1)} \\ \vdots \\ b_m^{(\ell+1)} \end{pmatrix}\right]$$

$$h_i^{(\ell+1)} = g^{(\ell+1)}\left(\mathbf{W}_{i,:}^{(\ell+1)} \boldsymbol{h}^{(\ell)} + \mathbf{b}^{(\ell+1)}\right) = g^{(\ell+1)}\left(\left\langle \boldsymbol{w}_1^{(\ell+1)}, \boldsymbol{h}^{(\ell)} \right\rangle + b_1^{(\ell+1)}\right)$$

How many layers and neurons ?

Lower bounds for approximation by $\mathcal{MLP}$ neural networks [Mai99]

Given a function $f : [-T, T]^{d_X} \in \mathcal{X} \rightarrow \mathbb{R}$, $T > 0$, with $f \in W^{k,2}$, then it exists a 1-hidden-layer $\mathcal{MLP}$ h_θ, with sigmoid activation function and N_K hidden neurons, such that the L^2-error $e = \|f - h_\theta\|_{L^2([-T,T]^{d_X})}$ is lower than a tolerance ε if :

$$N_K \approx \varepsilon^{\frac{1-d_X}{k}} \tag{1}$$

d_X =dim($\mathcal{X}$)

For $f \in L^2([-T, T]^{d_X})$, to reduce the error by an order or magnitude, the number of neurons must be multiplied by $10^{\frac{d_X-1}{2}}$!

Approximate radial functions [ES16]

$$\varphi(\boldsymbol{x}) = \left(\frac{R_{d_X}}{\|\boldsymbol{x}\|}\right)^{\frac{d_X}{2}} J_{\frac{d_X}{2}}\left(2\pi R_{d_X}\|\boldsymbol{x}\|\right) = \int_{\boldsymbol{w}:\|\boldsymbol{w}\|\leq R_d} e^{-2\pi i\langle \boldsymbol{x},\boldsymbol{w}\rangle}\, d\boldsymbol{w}$$

$J_{\frac{d_X}{2}}\left(2\pi R_{d_X}\|\boldsymbol{x}\|\right)$: Bessel function of first kind of order $\frac{d_X}{2}$

- Is 1 hidden layer enough ?
- How many neurons ?
- Approximation error ?
- Try with two hidden layers !

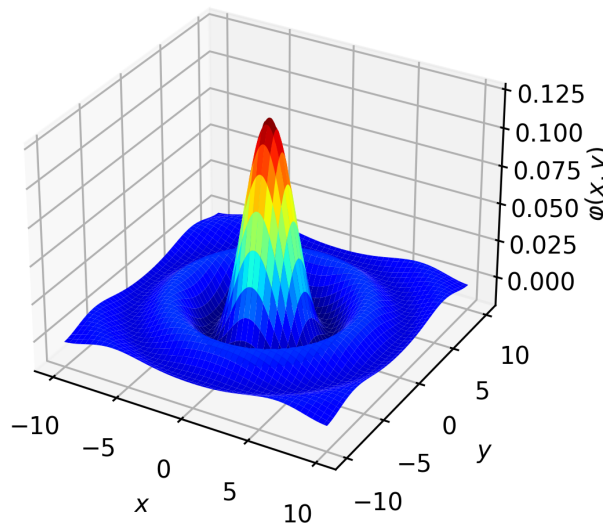


Figure – $d_X = 2$ and $R_d = 0.2$.

Main strategies

- Reduce the dimensionality : $\boldsymbol{\phi} : \boldsymbol{x} \mapsto \boldsymbol{z} \in S, \dim(S) < \dim(\mathcal{X})$ (PCA, ROM)
- Separate the interactions (disentanglement) :

$$\exists \{f_n\}_{n\in I}, f_n : \mathcal{X}_n \to \mathbb{R}, \mathcal{X}_n \subset \mathcal{X} \quad f(\boldsymbol{x}) = \sum_{n\in I} f_n(\boldsymbol{x})|_{\boldsymbol{x}\in\mathcal{X}_n}, \forall \boldsymbol{x} \in \mathcal{X}$$

$$\varepsilon = \sum_{n\in I} \varepsilon_n \approx \mathrm{card}(I) \cdot N_K^{\frac{k}{1-d_S}} \qquad N_K \approx \left(\frac{\varepsilon}{\mathrm{card}(I)}\right)^{\frac{1-d_S}{k}}$$

$$\exists \{f_n\}_{n=1}^{d_X}, f_n : \mathbb{R} \to \mathbb{R} \quad f(\boldsymbol{x}) = \sum_{n=1}^{d_X} f_n(w_n \cdot x_n), \forall \boldsymbol{x} \in \mathcal{X}$$

- Promote sparsity : dictionary learning $D = \left\{\left(k, v(\boldsymbol{x})\right)_n\right\}_{n\le N_D}$ (k is the key and v the corresponding value

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{n\in I} \theta_n[k_n] v_n(\boldsymbol{x}), \quad \mathrm{card}(\mathrm{I}) \le N_D$$

Train a statistical model

Chain rule

$$\nabla_{\boldsymbol{\theta}} L_{\mathcal{D}_{XY}} = \frac{1}{N} \sum_{k=1}^{N} \nabla_{\boldsymbol{\theta}} \ell \left(\boldsymbol{h}_{\boldsymbol{\theta}} \left(\boldsymbol{x}_k \right), \boldsymbol{y}_k \right)$$

$$\nabla_{\boldsymbol{\theta}} \ell \left(\boldsymbol{h}_{\boldsymbol{\theta}} \left(\boldsymbol{x} \right) \right) = \mathbb{J}_h^T . \nabla_{\boldsymbol{h}_{\boldsymbol{\theta}}} \ell \left(\boldsymbol{h}_{\boldsymbol{\theta}} \right) \qquad \mathbb{J}_h^{(k)} = \left[\begin{array}{ccc} \frac{\partial \boldsymbol{h}_{\boldsymbol{\theta}}^{(k)}}{\partial \theta_1^{(k)}} & \frac{\partial \boldsymbol{h}_{\boldsymbol{\theta}}^{(k)}}{\partial \theta_2^{(k)}} & \cdots \frac{\partial \boldsymbol{h}_{\boldsymbol{\theta}}^{(k)}}{\partial \theta_{u^{(k+1)}}^{(k)}} \end{array} \right]$$

Example : Autodiff of a classifier $\boldsymbol{h}_{\boldsymbol{\theta}} \left(\boldsymbol{x} \right) = \sigma_y \left(\sum_{c=1}^{N_y} h_c^{(N_\ell)} \left(\boldsymbol{x} \right) \boldsymbol{e}_c \right)$

- $h^{(o)} = g^{(o)} \left(\boldsymbol{a}^{(o)} \right) = \sigma_y \left(\boldsymbol{a}^{(o)} \right) = \mathrm{softmax}_y \left(\boldsymbol{a}^{(o)} \right) = \frac{e^{a_y^{(o)}(\boldsymbol{x})}}{\sum_{c=1}^{N_y} e^{a_c^{(o)}(\boldsymbol{x})}}$
- $a_c^{(o)} \left(\boldsymbol{x} \right) = h_c^{(N_\ell)} = g_c^{(N_\ell)} \left(a_c^{(N_\ell)} \left(\boldsymbol{x} \right) \right), \quad 1 \leq c \leq N_y$
- $\ell \left(h_{\boldsymbol{\theta}} \left(\boldsymbol{x} \right), y \right) = - \sum_{c=1}^{N_y} \ln \left(\sigma_c \left(\boldsymbol{a}^{(o)} \left(\boldsymbol{x} \right) \right) \right) \chi_{(y=c)}$

1. Derivative at output activation : $\frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y\right)}{\partial \sigma_c} = -\frac{\chi_{(y=c)}}{\sigma_c\left(\boldsymbol{a}^{(o)}(\boldsymbol{x})\right)}$

2. Gradient at output activation :

$$\boldsymbol{\nabla}_{\sigma} \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y\right) = \sum_{c=1}^{N_y} \frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y\right)}{\partial \sigma_c} \boldsymbol{e}_c = -\sum_{c=1}^{N_y} \frac{\chi_{(y=c)}}{\sigma_c\left(\boldsymbol{a}^{(o)}(\boldsymbol{x})\right)} \boldsymbol{e}_c = -\frac{\boldsymbol{e}_y}{\sigma_y\left(\boldsymbol{a}^{(o)}(\boldsymbol{x})\right)}$$

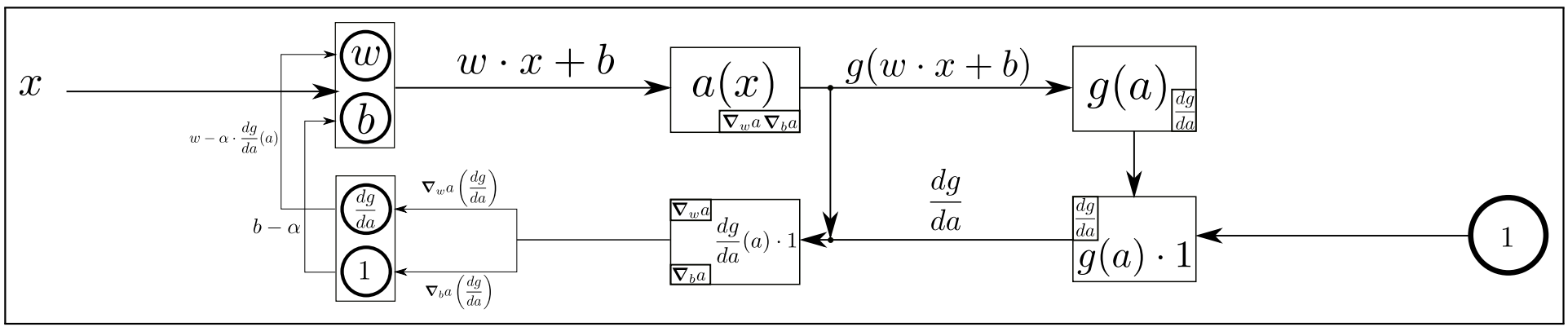
3. Derivative at N_ℓ−layer activation : $\frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y\right)}{\partial h_c^{(N_\ell)}} = \sum_{c'=1}^{N_y} \frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y\right)}{\partial \sigma_{c'}} \cdot \frac{\partial \sigma_{c'}}{\partial h_c^{(N_\ell)}}\Big|_{\boldsymbol{a}^{(o)}(\boldsymbol{x})}$

with : $\frac{\partial \sigma_{c'}}{\partial h_c^{(N_\ell)}}\Big|_{\boldsymbol{a}^{(o)}(\boldsymbol{x})} = \frac{\partial \sigma_{c'}}{\partial a_c^{(o)}} = \sigma_{c'}\left(\boldsymbol{a}^{(o)}(\boldsymbol{x})\right) \cdot \left(\chi_{(c'=c)} - \sigma_c\left(\boldsymbol{a}^{(o)}(\boldsymbol{x})\right)\right)$

4. Gradient at N_ℓ−layer activation :

$$\boldsymbol{\nabla}_{h^{(N_\ell)}} \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y\right) = \sum_{c=1}^{N_y} \frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y\right)}{\partial h_c^{(N_\ell)}} \boldsymbol{e}_c = \sum_{c=1}^{N_y} \sum_{c'=1}^{N_y} \frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y\right)}{\partial \sigma_{c'}} \cdot \frac{\partial \sigma_{c'}}{\partial h_c^{(N_\ell)}}\Big|_{\boldsymbol{a}^{(o)}(\boldsymbol{x})} =$$

$$= -\sum_{c=1}^{N_y} \left(\chi_{(y=c)} - \sigma_c\right) \boldsymbol{e}_c = -\left(\boldsymbol{e}_y - \boldsymbol{\sigma}\left(\boldsymbol{a}^{(o)}\right)(\boldsymbol{x})\right)$$



5. Derivative at N_ℓ−layer pre-activation : $\frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\right)}{\partial a_c^{(N_\ell)}} = \sum_{c'=1}^{N_y} \frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\right)}{\partial h_{c'}^{(N_\ell)}} \cdot \frac{\partial h_{c'}^{(N_\ell)}}{\partial a_c^{(N_\ell)}}$ with :

$$\frac{\partial h_{c'}^{(N_\ell)}(\boldsymbol{x})}{\partial a_c^{(N_\ell)}} = \frac{\partial g_{c'}^{(N_\ell)}}{\partial a_c^{(N_\ell)}}\left(a_{c'}^{(N_\ell)}(\boldsymbol{x})\right)\chi_{(c'=c)} \quad \frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\right)}{\partial a_c} =$$

$$-\sum_{c'=1}^{N_y}\left(\chi_{(y=c')} - \sigma_{c'}\right)\frac{\partial g_{c'}}{\partial a_c^{(N_\ell)}}\left(a_{c'}^{(N_\ell)}(\boldsymbol{x})\right)\chi_{(c'=c)} = -\left(\chi_{(y=c)} - \sigma_c\right)\frac{\partial g_c^{(N_\ell)}}{\partial a_c^{(N_\ell)}}\left(a_c^{(N_\ell)}(\boldsymbol{x})\right)$$

6. Gradient at N_ℓ−layer pre-activation (provided that $g_c^{(N_\ell)} = g \quad \forall c$)

$$\boldsymbol{\nabla}_{a^{(N_\ell)}}\ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\right) = \sum_{c=1}^{N_y}\frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\right)}{\partial a_c^{(N_\ell)}}\boldsymbol{e}_c = = \boldsymbol{\nabla}_h\ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\right)\odot\boldsymbol{\nabla}_a g = -\left(\boldsymbol{e}_y - \boldsymbol{\sigma}\left(\boldsymbol{a}^{(o)}(\boldsymbol{x})\right)\right)\odot\boldsymbol{\nabla}_a g$$

7. Derivative at N_ℓ-layer weights : $a_c^{(N_\ell)} = \sum_{c'=1}^{N_y} W_{cc'}^{(N_\ell)} h_{c'}^{(N_\ell-1)} + b_c^{(N_\ell)}$

$$\frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\right)}{\partial W_{cc'}^{(N_\ell)}} = \sum_{n=1}^{N_y}\frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\right)}{\partial a_n^{(N_\ell)}}\frac{\partial a_n^{(N_\ell)}}{\partial W_{cc'}^{(N_\ell)}} = \frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\right)}{\partial a_c^{(N_\ell)}}h_{c'}^{(N_\ell-1)}$$

8. Derivative at N_ℓ-layer bias : $\frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\right)}{\partial b_c^{(N_\ell)}} = \sum_{n=1}^{N_y}\frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\right)}{\partial a_n^{(N_\ell)}}\frac{\partial a_n^{(N_\ell)}}{\partial b_c^{(N_\ell)}} = \frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\right)}{\partial a_c^{(N_\ell)}}\cdot 1$

9. Gradient at N_ℓ-layer weight $W_{cc'}^{(N_\ell)}$: $\boldsymbol{\nabla}_{W^{(N_\ell)}}\ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\right) = \boldsymbol{\nabla}_a\ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\right)\otimes\boldsymbol{h}^{(N_\ell-1)}$

10. Gradient at N_ℓ-layer bias : $\boldsymbol{\nabla}_{b^{(N_\ell)}}\ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\right) = \boldsymbol{\nabla}_a\ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\right)$

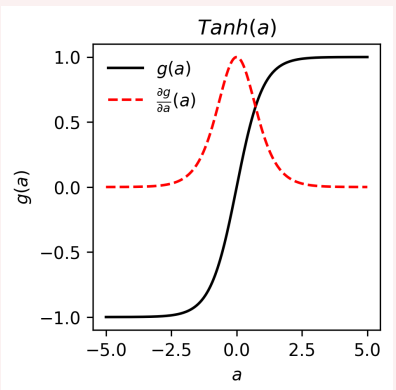
11. Derivative at hidden layer $N_{\ell-1}$ activation :
$$\frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y\right)}{\partial h_c^{(N_{\ell-1})}} = \sum_{n=1}^{N_y} \frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y\right)}{\partial a_n^{(N_\ell)}} \frac{\partial a_n^{(N_\ell)}}{\partial h_c^{(N_{\ell-1})}} = \sum_{n=1}^{N_y} \frac{\partial \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y\right)}{\partial a_n^{(N_\ell)}} W_{nc}^{(N_\ell)}$$
12. Gradient at hidden layer $N_{\ell-1}$ activation $\boldsymbol{\nabla}_{h^{(N_{\ell-1})}} \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y\right) = \boldsymbol{W}^{(N_\ell)\,T} . \boldsymbol{\nabla}_a \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y\right)$
13. Reiterate $\ell = 1 \ldots N_\ell - 1$

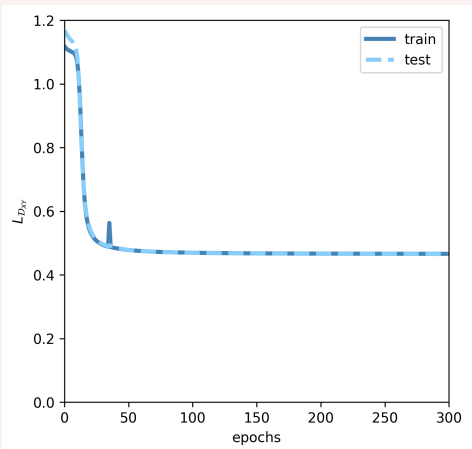
Vanishing gradients

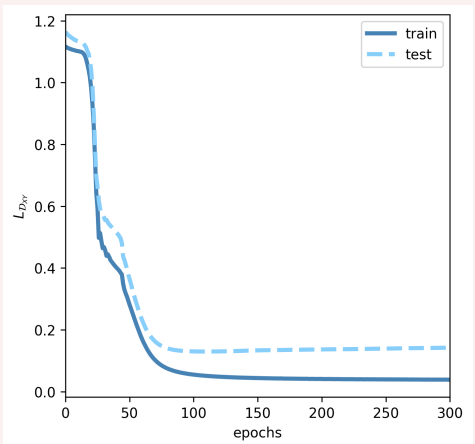
$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = g^{(o)} \circ a^{(o)} \circ g^{(N_\ell)} \circ a^{(N_\ell)} \circ \ldots \circ g^{(\ell)} \circ a^{(\ell)} \circ \ldots g^{(1)} \circ a^{(1)}(\boldsymbol{x})$$

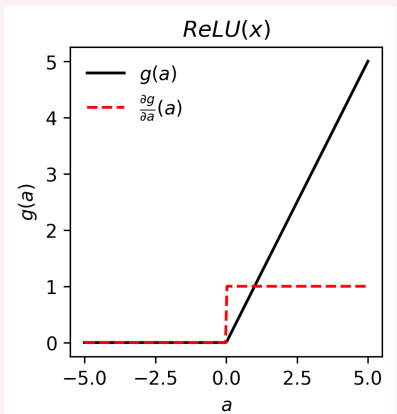
$$\frac{\partial L_{\mathcal{D}_{XY}}}{\partial w_c^{(1)}} = \frac{\partial g^{(o)}}{\partial a^{(o)}} \cdot \frac{\partial a^{(o)}}{\partial h^{(N_\ell)}} \left(\prod_{\substack{\ell=1 \\ N_\ell > 1}}^{N_\ell - 1} \frac{\partial g^{(N_\ell+1-\ell)}}{\partial a^{(N_\ell+1-\ell)}} \cdot \frac{\partial a^{(N_\ell+1-\ell)}}{\partial h^{(N_\ell-\ell)}} \right) \frac{\partial g^{(1)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial w_c^{(1)}}$$

$$\nabla_{w^{(1)}} L_{\mathcal{D}_{XY}} \approx 0 \ldots \frac{\partial g^{(N_\ell+1-\ell)}}{\partial a^{(N_\ell+1-\ell)}} \approx 0$$

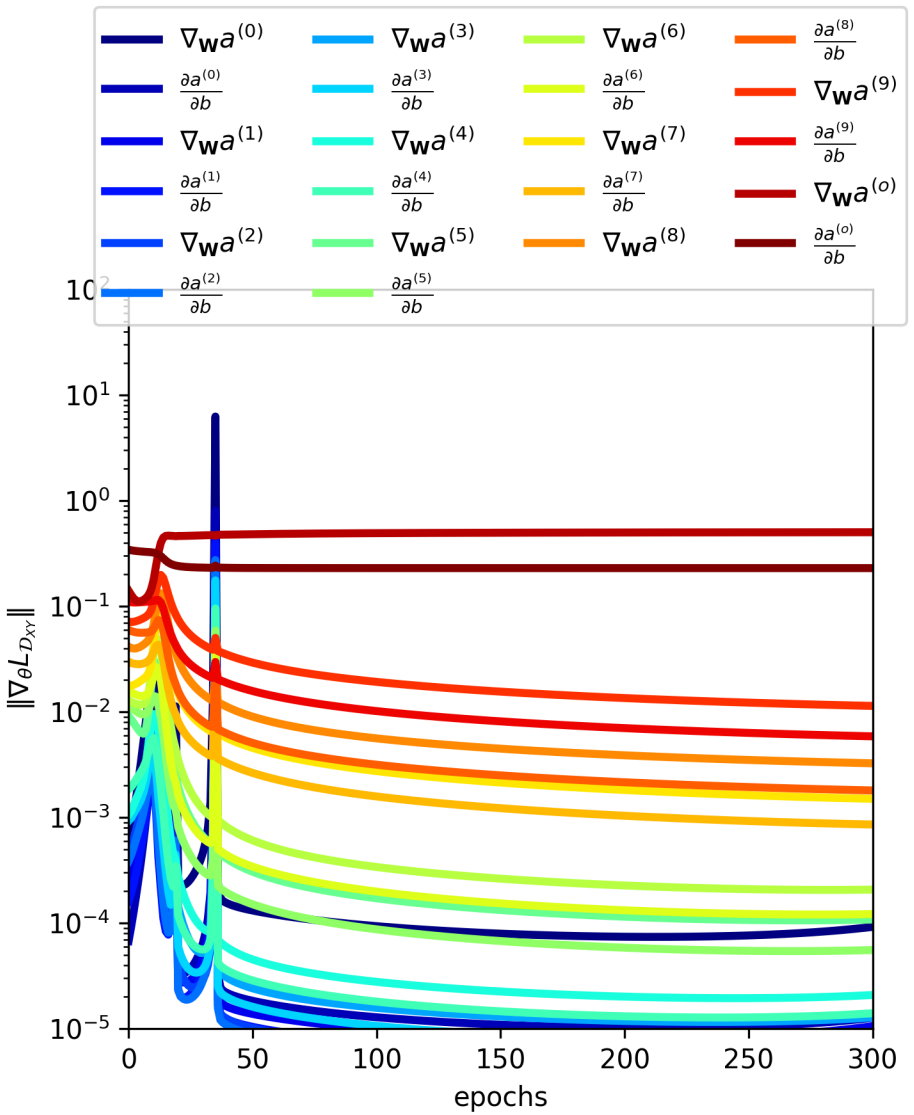


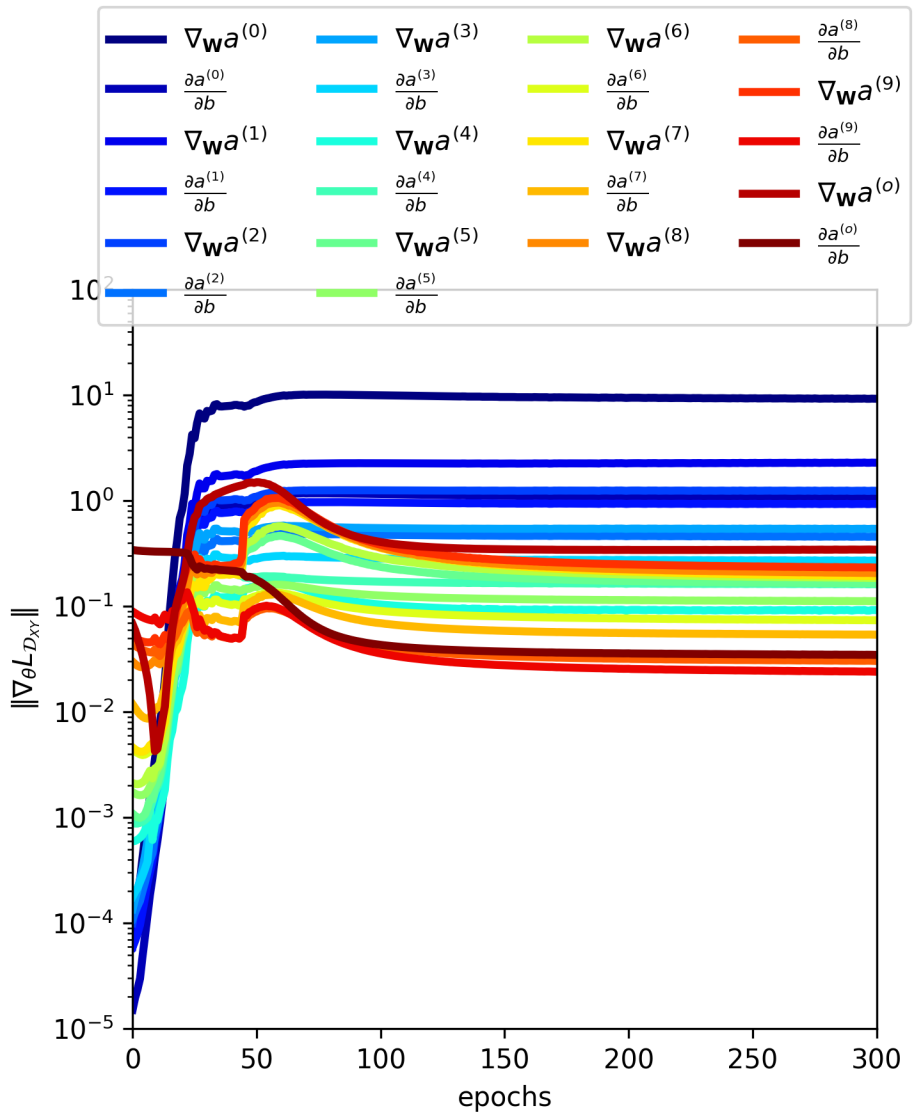






Vanishing gradients





Initialize weights

Example : 2-hidden layers $\mathcal{MLP}$

$$\frac{\partial L_{\mathcal{D}XY}}{\partial w_{1c}^{(1)}} = \frac{\partial g^{(o)}}{\partial a^{(o)}} \cdot \frac{\partial a^{(o)}}{\partial h_1^{(1)}} \frac{\partial h_1^{(1)}}{\partial a_1^{(1)}} \cdot x_c \qquad \frac{\partial L_{\mathcal{D}XY}}{\partial w_{2c}^{(1)}} = \frac{\partial g^{(o)}}{\partial a^{(o)}} \cdot \frac{\partial a^{(o)}}{\partial h_2^{(1)}} \frac{\partial h_2^{(1)}}{\partial a_2^{(1)}} \cdot x_c$$

$$\mathbb{V}(a_i^{(k)}) = \mathbb{V}\left(\sum_{j=1}^{u^{(k)}} W_{ij}^{(k)} h_j^{(k-1)}\right) + \mathbb{V}\left(b_i^{(k)}\right)$$

$$\mathbb{V}\left(\sum_{j=1}^{u^{(k)}} W_{ij}^{(k)} h_j^{(k-1)}\right) = \sum_{j=1}^{u^{(k)}} \mathbb{V}\left(W_{ij}^{(k)}\right) \cdot \left(\mathbb{E}\left[h_j^{(k-1)}\right]\right)^2 + \sum_{j=1}^{u^{(k)}} \left(\mathbb{E}\left[W_{ij}^{(k)}\right]\right)^2 \cdot \mathbb{V}\left(h_j^{(k-1)}\right) + \sum_{j=1}^{u^{(k)}} \mathbb{V}\left(W_{ij}^{(k)}\right) \cdot \mathbb{V}\left(h_j^{(k-1)}\right)$$

How to initialize weights ?

- $\theta^{(0)} = cst.$
  - If $N_\ell = 1$: no vanishing gradients
  - Symmetric evolution :
    $a_1^{(1)} = a_2^{(1)} = 0 \rightarrow h_1^{(1)} = g(a_1^{(1)}) = h_1^{(1)} = g(a_1^{(1)}) \rightarrow w_{1c}^{(1)(i)} = w_{2c}^{(1)(i)}$

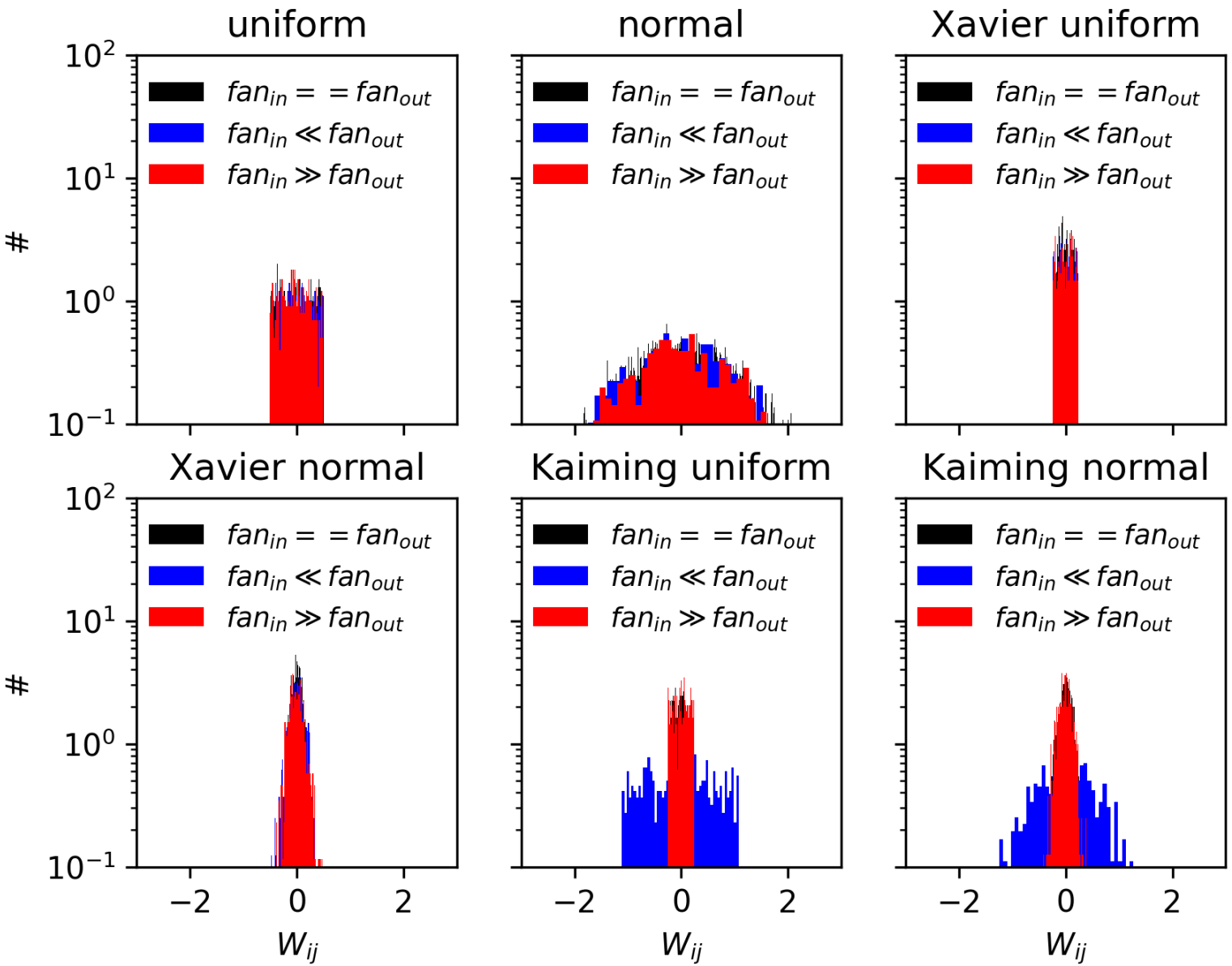
Initialize weights

Example : 2-hidden layers $\mathcal{MLP}$

$$\frac{\partial L_{\mathcal{D}XY}}{\partial w_{1c}^{(1)}} = \frac{\partial g^{(o)}}{\partial a^{(o)}} \cdot \frac{\partial a^{(o)}}{\partial h_1^{(1)}} \frac{\partial h_1^{(1)}}{\partial a_1^{(1)}} \cdot x_c \qquad \frac{\partial L_{\mathcal{D}XY}}{\partial w_{2c}^{(1)}} = \frac{\partial g^{(o)}}{\partial a^{(o)}} \cdot \frac{\partial a^{(o)}}{\partial h_2^{(1)}} \frac{\partial h_2^{(1)}}{\partial a_2^{(1)}} \cdot x_c$$

How to initialize weights ?

- $\theta_i \sim \mathcal{N}(\mathbf{0}, \sigma_\theta \mathbb{I})$ or $\theta \sim \mathcal{U}(-\sqrt{3\sigma_\theta}, \sqrt{3\sigma_\theta})$
  - which σ_θ ?
  - $\mathbb{V}\left(a_i^{(k)}\right) = \left(u^{(k)} \cdot \mathbb{V}(W^{(k)})\right)^i \cdot \mathbb{V}(h^{(k-1)}) + \mathbb{V}\left(b_i^{(k)}\right)$
  - $\left(u^{(k)} \cdot \mathbb{V}(W^{(k)})\right)^i$ can vanish
  - Decorrelate : $\mathbb{V}(W^{(k)}) = \frac{1}{u^{(k)}} \Rightarrow \mathbb{V}\left(a_i^{(k)}\right) = \mathbb{V}(h^{(k-1)}) + \mathbb{V}\left(b_i^{(k)}\right)$
  - Xavier/Glorot : $\mathbb{V}(W^{(k)}) = \frac{2}{u^{(k)}+u^{(k+1)}}$
  - He/Kaiming : $\mathbb{V}(W^{(k)}) = \frac{2}{u^{(k)}}$


uniform
normal
Xavier uniform
Xavier normal
Kaiming uniform
Kaiming normal
$fan_{in} = = fan_{out}$
$fan_{in} \ll fan_{out}$
$fan_{in} \gg fan_{out}$
#
10^2
10^1
10^0
10^{-1}
−2
0
2
W_{ij}

Neural networks for time series

Time-histories

$\mathcal{D}_{XY} = \left\{\left(\mathbb{X}_i, \mathbb{Y}_i\right)\right\}_{i=1}^{N}$: i.i.d. discrete input $\mathbb{X} = \left\{\boldsymbol{x}\left[t\right]\right\}_{t=1}^{N_t}$ and output signals $\mathbb{Y} = \left\{\boldsymbol{y}\left[t\right]\right\}_{t=1}^{N_t}$, with $\boldsymbol{x}\left[t\right] \in \mathbb{R}^{d_X}$ and $\boldsymbol{y}\left[t\right] \in \mathbb{R}^{d_Y}$, both of length N_t

Recurrent Neural Networks $\mathcal{RNN}$

$$\begin{cases} \boldsymbol{a}_t\left(\boldsymbol{h}, \boldsymbol{x}\right) = \boldsymbol{W}^{(h)}\boldsymbol{h}_{t-1} + \boldsymbol{W}^{(x)}\boldsymbol{x}\left[t\right] + \boldsymbol{b}^{(h)} \\ \boldsymbol{h}_t\left(\boldsymbol{a}\right) = \boldsymbol{g}^{(h)}\left(\boldsymbol{a}_t\right) \\ \boldsymbol{a}_t^{(o)} = \boldsymbol{W}^{(o)}\boldsymbol{h}_t + \boldsymbol{b}^{(o)} \\ \boldsymbol{z}_t\left(\boldsymbol{a}^{(o)}\right) = \boldsymbol{g}^{(o)}\left(\boldsymbol{a}_t^{(o)}\right) \\ \boldsymbol{h}_{\boldsymbol{\theta}}\left(\boldsymbol{x}\left[t\right]\right) = \boldsymbol{z}_t \circ \boldsymbol{h}_t \circ \boldsymbol{a}_t\left(\cdot, \boldsymbol{h}_{t-1}\right) \circ \boldsymbol{x}\left[t\right] \end{cases}$$

hidden state $\boldsymbol{h}\left[t\right] \in \mathbb{R}^{d_h}$: "memory" variable that keeps track of past states

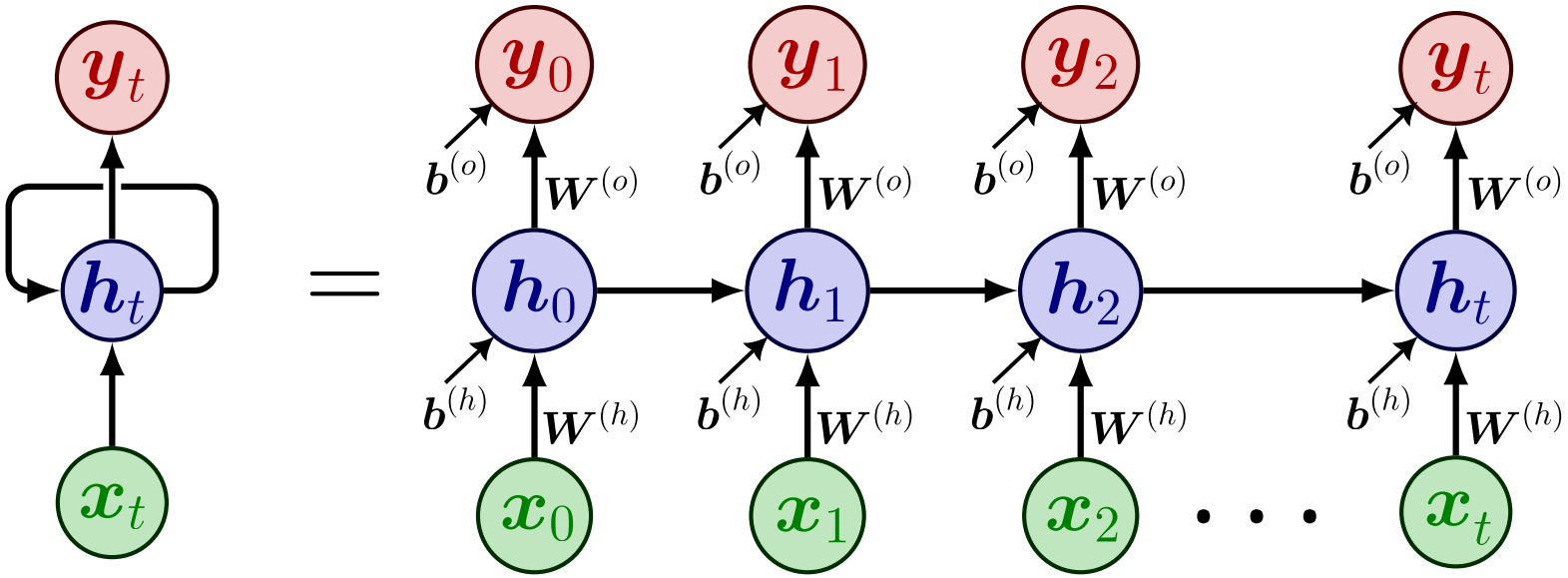


Figure – Sketch of $\mathcal{RNN}$ scheme according to **??**

- Feedback : $\boldsymbol{a}_t\left(\boldsymbol{h}, \boldsymbol{x}\right) = \boldsymbol{W}^{(h)}\boldsymbol{h}_{t-1} + \boldsymbol{W}^{(x)}\boldsymbol{x}\left[t\right] + \boldsymbol{b}^{(h)}$
- Biases $\sim$ exogenous variables
- Loss function $L_{\mathcal{D}_{XY}}\left(\boldsymbol{\theta}\right) = \frac{1}{N}\sum_{t=1}^{N_t}\sum_{i=1}^{N}\ell\left(\boldsymbol{y}_i\left[t\right], \boldsymbol{h}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\left[t\right]\right)\right)$

- Training instability because of long-term time-dynamics

$$\frac{\partial \ell}{\partial W_{ij}^{(x)}}\left(\boldsymbol{y}_i\left[t\right], \boldsymbol{h}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\left[t\right]\right)\right) = \sum_{k,l,m,n} \frac{\partial \ell}{\partial z_{t,k}} \frac{\partial g^{(o)}}{\partial h_{t,l}} \frac{\partial g^{(h)}}{\partial a_{t,m}} \left(\frac{\partial a_{t,m}}{\partial h_{t-1,n}} \frac{\partial h_{t-1,n}}{\partial W_{ij}^{(x)}} + \frac{\partial a_{t,m}}{\partial W_{ij}^{(x)}} \right)$$

$$\begin{cases} \dfrac{\partial g^{(o)}}{\partial h_{t,l}} = \displaystyle\sum_p \frac{\partial g^{(o)}}{\partial a_{t,p}^{(o)}} W_{lp}^{(o)} \\ \dfrac{\partial a_{t,m}}{\partial h_{t-1,n}} = W_{mn}^{(h)} \\ \dfrac{\partial a_{t,m}}{\partial W_{ij}^{(x)}} = \delta_{mi} x_j[t-1] \\ \dfrac{\partial h_{t-1,n}}{\partial W_{ij}^{(x)}} = \displaystyle\sum_{q,r} \frac{\partial g^{(h)}}{\partial a_{t,q}} W_{q,r}^{(h)} \frac{\partial h_{t-2,r}}{\partial W_{ij}^{(x)}} + \frac{\partial g^{(h)}}{\partial a_{t,i}} x_j[t-1] \end{cases}$$

Non-local gradient graph.

- Back-Propagation Through Time (BPTT [Sal+18])

$$\begin{cases} \nabla_{W^{(x)}} \ell \left(y_i \left[t \right], h_{\theta} \left(x_i \left[t \right] \right) \right) = \nabla_{h_t} \ell . \left(\sum_{\tau=1}^{t} \left(h_t \otimes \nabla_{h_\tau} \right) \otimes \nabla_{W^{(x)}} h_\tau \right) \\ \nabla_{h_\tau} h_t = \prod_{s=\tau+1}^{t} \frac{\partial h_s}{\partial h_{s-1}} \\ \frac{\partial h_s}{\partial h_{s-1}} = W^{(h)\,T} \left(\sum_k \frac{\partial g^{(o)}}{\partial h_{s-1,k}} e_k \otimes e_k \right) \end{cases}$$

- "immediate" $\nabla_{W^{(x)}} h_\tau$
- "short-term" $W^{(h)\,T} \left(\sum_k \frac{\partial g^{(o)}}{\partial h_{s-1,k}} e_k \otimes e_k \right), \left\| \sum_k \frac{\partial g^{(o)}}{\partial h_{s-1,k}} e_k \otimes e_k \right\| \leq \gamma$
- $\rho \left(W^{(h)} \right) = \max_i |\lambda_i| < \frac{1}{\delta}, \lambda_i \in \mathrm{Spec} \left(W^{(h)} \right)$
- "long-term" $\prod_{s=\tau+1}^{t} \frac{\partial h_s}{\partial h_{s-1}} \rightarrow \| \| \frac{\partial h_s}{\partial h_{s-1}} \| \leq \frac{\gamma}{\delta}$
- $\left\| \nabla_{W^{(x)}} \ell \left(y_i \left[t \right], h_{\theta} \left(x_i \left[t \right] \right) \right) \right\| \leq \left(\frac{\gamma}{\delta} \right)^{t-\tau} \left\| \nabla_{h_t} \ell \right\|$ Vanishing gradients !

Other types of $\mathcal{RNN}$: input-to-hidden

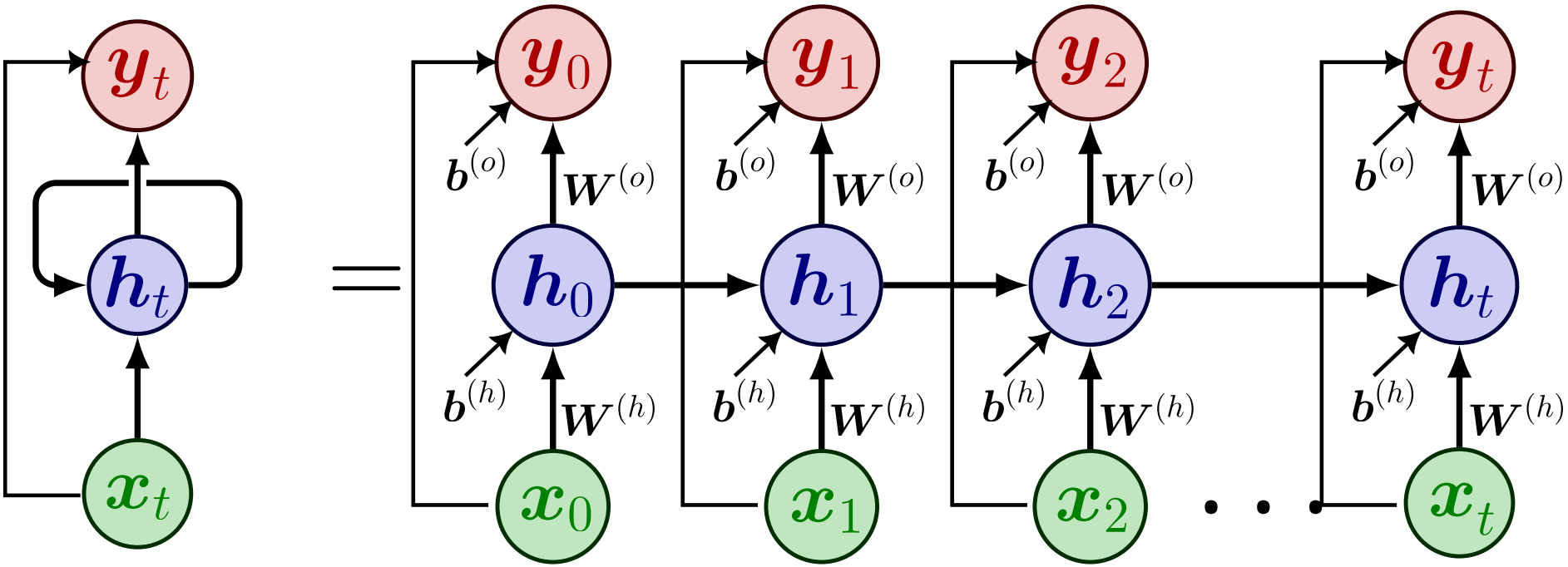


Figure – Sketch of the $\mathcal{RNN}$ version with input-to-hidden connection.

Other types of $\mathcal{RNN}$: hidden-to-hidden

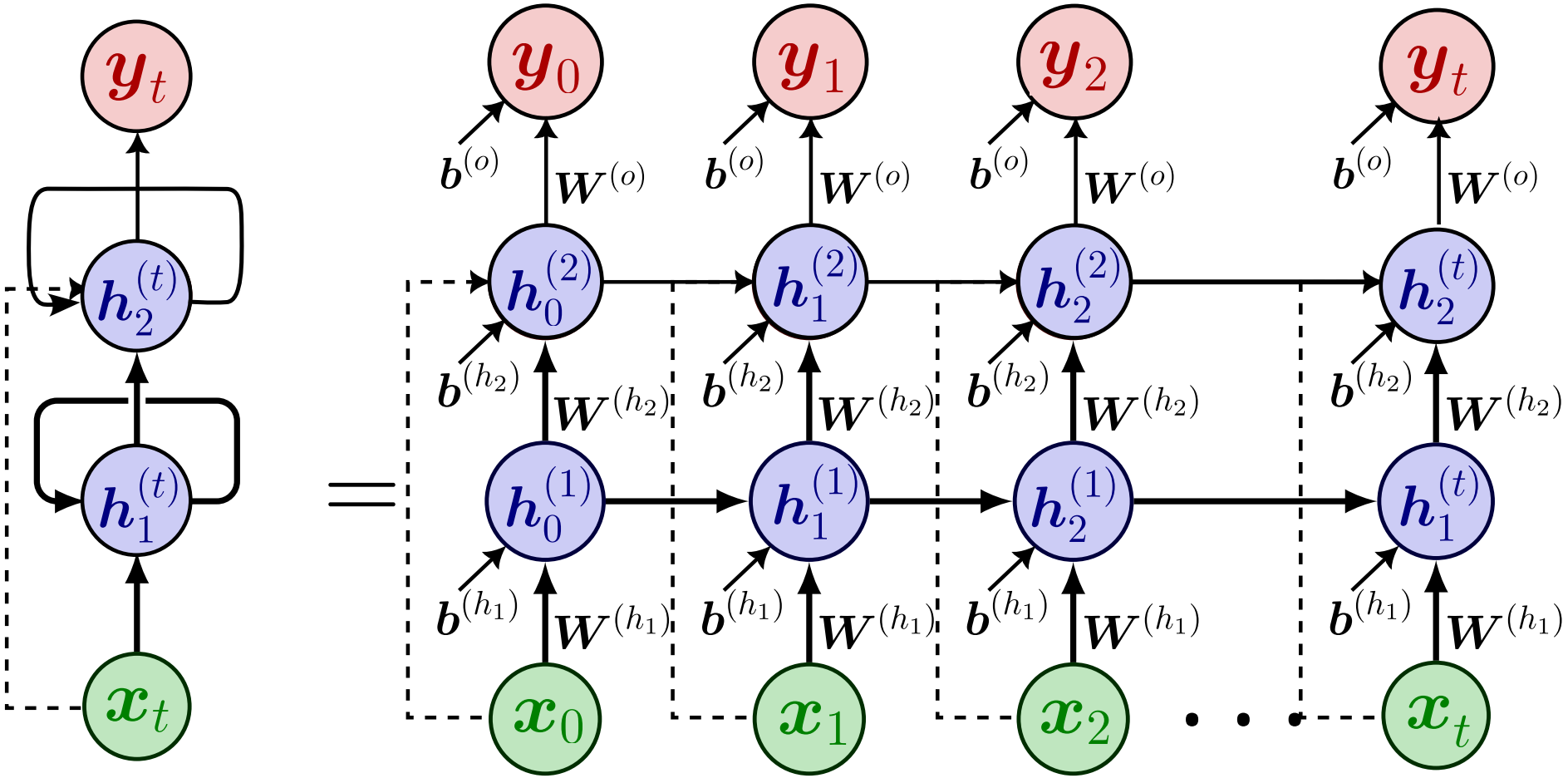


Figure – Sketch of the $\mathcal{RNN}$ version with hidden-to-hidden connection.

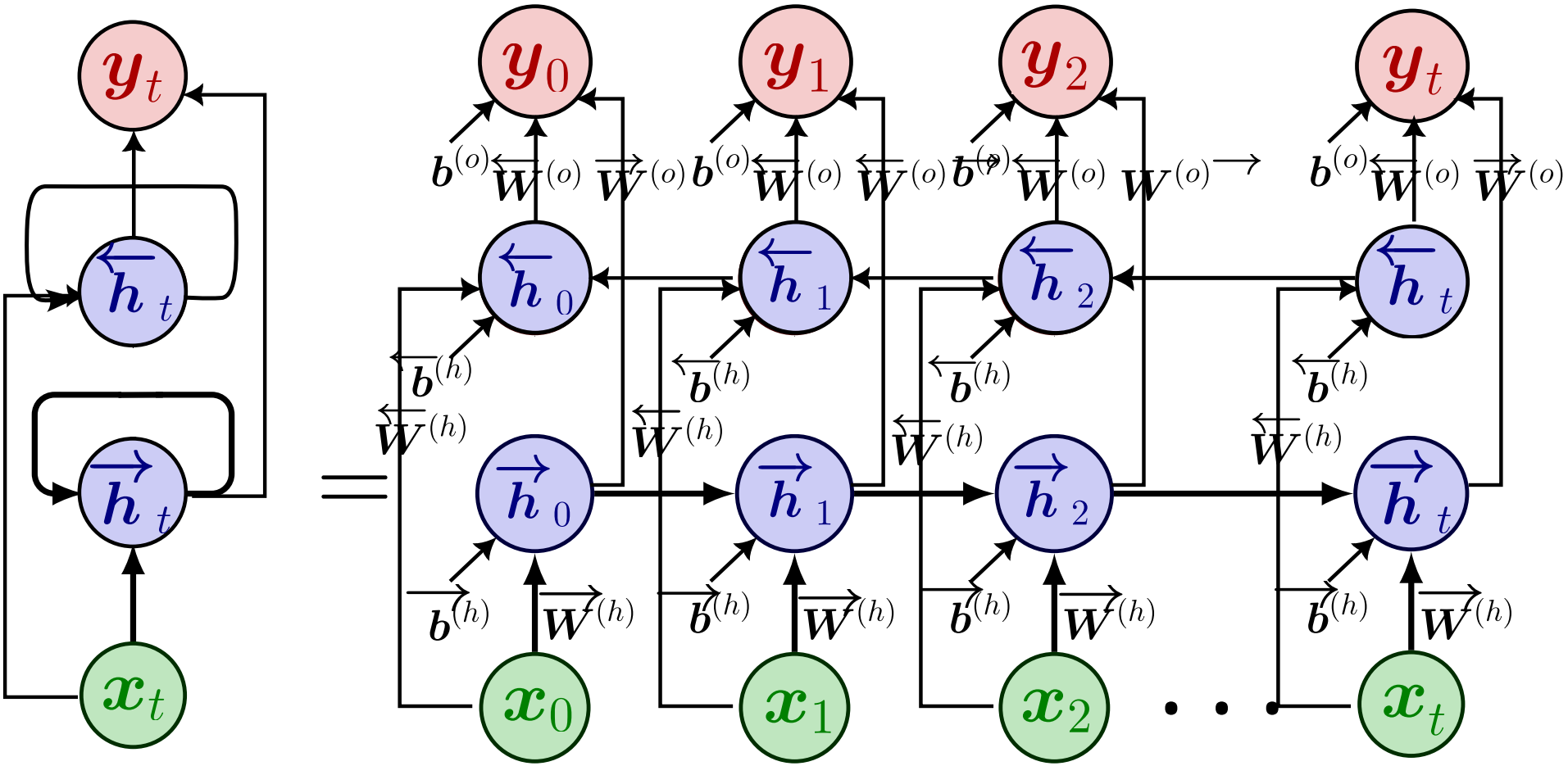


Figure – Sketch of the $\mathcal{BRNN}$ version with hidden-to-hidden connections.

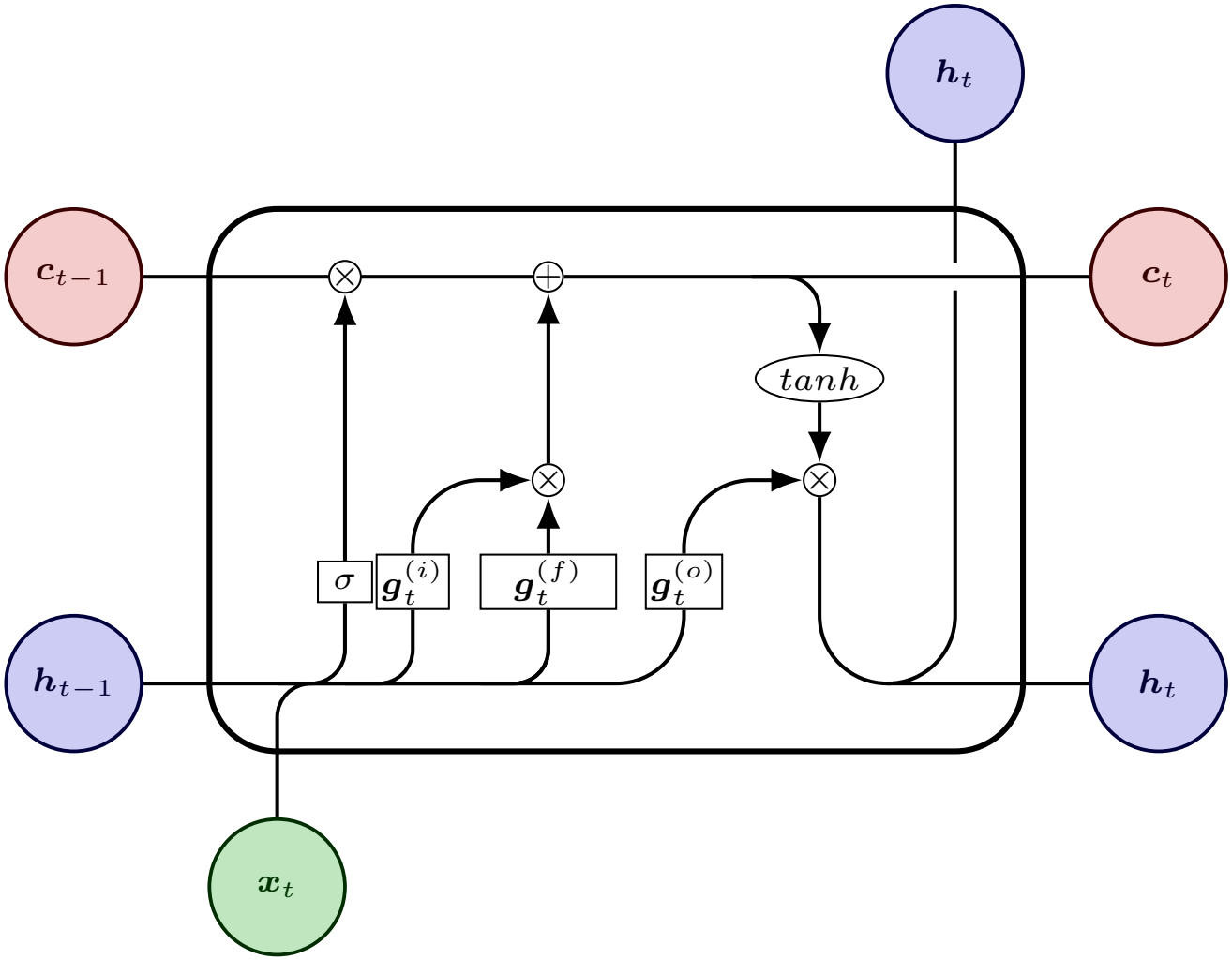


Figure – Sketch of the standard $\mathcal{LSTM}$ architecture.

Ideas

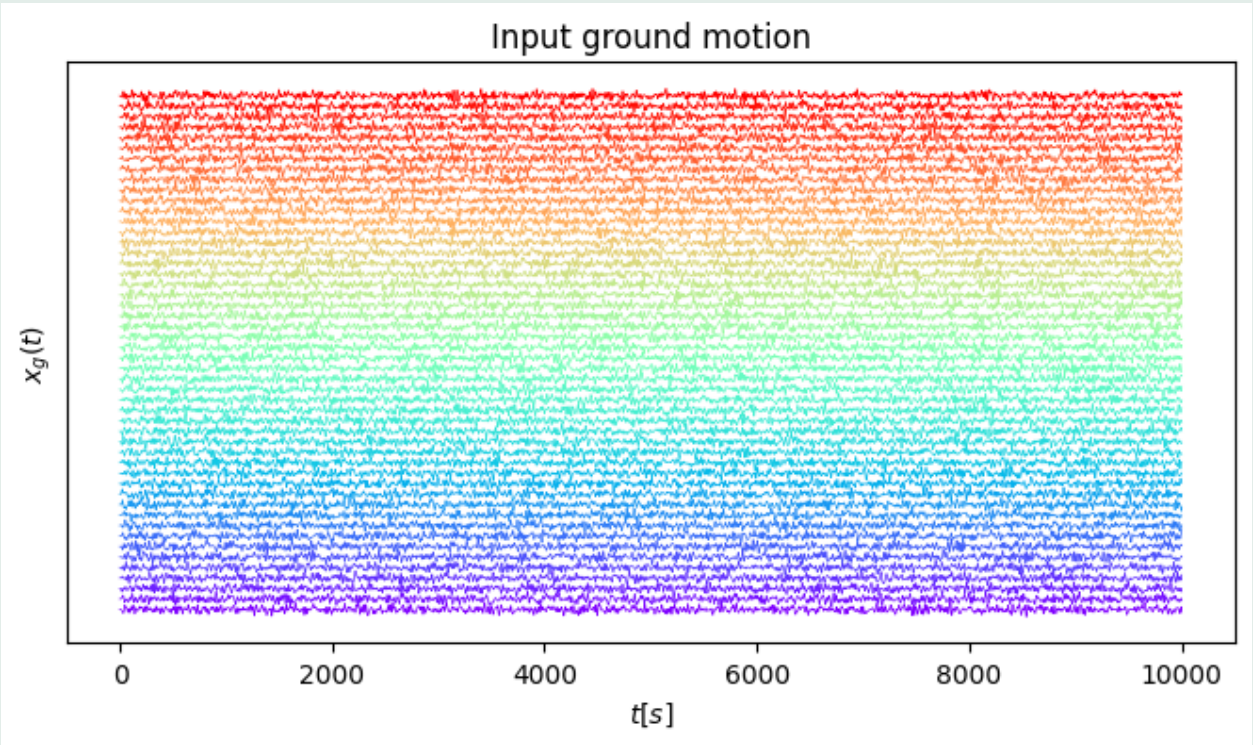
- Add "memory cells" c_t with "gates"
- Introduce a forget gate $g^{(f)}$ for past hidden states

$$
\begin{cases}
z_t^{(i)} = g^{(i)}\left(W^{(ii)}\, x\,[t] + W^{(ih)}\, h_{t-1} + W^{(ic)}\, c_{t-1} + b^{(i)}\right) \\
z_t^{(f)} = g^{(f)}\left(W^{(fi)}\, x\,[t] + W^{(fh)}\, h_{t-1} + W^{(fc)}\, c_{t-1} + b^{(f)}\right) \\
c_t = \sum_j z_{t,j}^{(i)} \cdot tanh\left(\left\langle w_j^{(ci)}, x\,[t]\right\rangle + \left\langle w_j^{(ch)}, h_{t-1}\right\rangle + b_j^{(c)}\right) e_j + \sum_j z_{t,j}^{(f)} \cdot c_{t-1,j} e_j \\
z_t^{(o)} = g^{(o)}\left(W^{(oi)}\, x\,[t] + W^{(oh)}\, h_{t-1} + W^{(oc)}\, c_t + b^{(o)}\right) \\
h_t = \sum_j z_{t,j}^{(o)} \cdot tanh\left(z_{t,j}^{(c)}\right) e_j
\end{cases}
$$

Predict the non-linear response of a MDOF with $\mathcal{LSTM}$

$$\mathbb{M}\ddot{\boldsymbol{y}}(t) + \mathbf{C}\dot{\boldsymbol{y}}(t) + \boldsymbol{F}(\boldsymbol{y}(t), \dot{\boldsymbol{y}}(t)) = -\mathbb{M}\boldsymbol{\Gamma}\ddot{x}_G(t)$$

$$\dot{f}_i(t) = k_i \cdot (\dot{y}_{i+1}(t) - \dot{y}_i(t)) - \alpha_i \cdot |\dot{y}_{i+1}(t) - \dot{y}_i(t)| \cdot |f_i(t)|^{n_i - 1} \cdot f_i(t) - \beta_i \cdot (\dot{y}_{i+1}(t) - \dot{y}_i(t)) \cdot |f$$



Hands-on

Recap of basic probability

Probability measure

$\boldsymbol{X} : (\Omega, \mathcal{E}, \mathbb{P}) \rightarrow (\mathcal{X}, \Xi)$, with a σ-algebra $\mathcal{E} \subset \mathcal{B}(\mathbb{R})$ on Ω

$$\forall A \in \mathcal{E} \qquad P_X(A) = \mathbb{P}(\boldsymbol{X}^{-1}(A)) = \mathbb{P}(\boldsymbol{X} \in A)$$

Absolute continuity

P_X is σ-finite measure dominated by a σ-finite measure μ, i.e. iff $\forall \varepsilon > 0, \exists \delta(\varepsilon)$ such that $\forall A \in \Xi$ such that $\mu(A) < \delta(\varepsilon)$

$$P_X(A) < \varepsilon$$

Theorem

Radon-Nikodym theorem

For two σ-finite measures on a σ-algebra Ξ, namely μ and P_X such that P_X is absolutely continuous with the respect to μ, $\exists p_X \in L^1(\mu)$ such that

$$P_X(A) = \int_A p_X(\boldsymbol{x}) \cdot \mu(d\boldsymbol{x}), \qquad \forall A \in \Xi$$

p_x is called probability density of P_X and it corresponds to the Radon-Nikodym derivative $p_X = \frac{dP_X}{d\mu}$.

Theorem

Strong law of large numbers (Kolmogorov)
Given a set of random variables $(X_n)_{n=1}^{N}$ *independent identically distributed (i.i.d.) and Lebesgue integrable, with expected value* $\mu_X < +\infty$*, and a sample average* $\bar{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n$*, then :*

$$\mathbb{P}\left[\lim_{N\to+\infty} \bar{X}_n - \mu_X\right] = 1$$

which means that the sample average $\bar{X}_N$ *converges almost surely to the expected value* μ*.*

Theorem

Weak law of large numbers (Khintchine)

Given a sequence of random variables $(X_i)_{i=1}^N$ *independent identically distributed (i.i.d.) and Lebesgue integrable, with expected value* $\mu_X < +\infty$*, and a sample average* $\bar{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$*, then :*

$$\forall \varepsilon > 0, \qquad \lim_{N \to +\infty} \mathbb{P}\left[|\bar{X}_n - \mu_X| \leq \varepsilon\right] = 1$$

which means that the sample average $\bar{X}_N$ *converges in probability to* μ

Theorem

Given a sequence of random variables $(X_n)_{n=1}^N$ *that converges in probability to* X*, then any continuous function* $f : \mathcal{X} \to \mathbb{R}$ *converges in probability too.*

Consistent estimator

An estimator $\boldsymbol{\theta}_N = \boldsymbol{\theta}\,(X_1, \ldots, X_N)$ is said to be *consistent* if it converges in probability to its limit $\boldsymbol{\theta}$.

Remark

$\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ is a consistent estimator of μ_X. $\bar{X}_N^2$ converges in probability to μ_X^2 and the estimator $\bar{Y}_N = \frac{1}{N}\sum_{i=1}^{N} X_i^2$ converges in probability to $\mathbb{E}\left[X_i^2\right]$. Therefore, the variance estimator of each X_i s_N^2 that reads :

$$s_N^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}_N\right)^2$$

converges in probability to the variance of X_i, $\sigma_{X_i}^2 = \mathbb{V}\left[X_i\right]$

Theorem

Central limit theorem

Given a sequence of i.i.d. random variables $X = \{X_i\}_{i=1}^{N}$ *with expected value* $\mathbb{E}[X_i] = \mu$ *and variance* $\mathbb{V}[X_i] = \sigma^2 < +\infty$ *and a random variable* $\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$*, the random variable*

$$Z_N = \sqrt{N}\frac{\bar{X}_N - \mu}{\sigma}, \qquad \mathbb{E}[Z_N] = 0, \mathbb{V}[Z_N] = 1$$

$$p(Z_N) \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)$$

Statistical model

$\mathcal{H}_\Theta := \{P_\theta, \theta \in \Theta\}$, if $P \in \mathcal{H}_\Theta$, there exist θ^* such that $P_{\theta^*} = P$.
θ^* is unique if and only if $\mathcal{H}_\Theta$ is *identifiable*, i.e., $\theta \mapsto P_\theta$ is injective.

Likelihood maximization

Idea : find $\hat{\theta} \approx \theta^*$ maximizing the *likelihood*

$$\hat{\theta}(\mathcal{D}_X) = \arg\max_{\theta \in \Theta} p_\theta(\boldsymbol{x} | \boldsymbol{x} \in \mathcal{D}_X) = \arg\max_{\theta \in \Theta} = \prod_{i=1}^{N} p_\theta(\boldsymbol{x}_i)$$

- $\hat{\theta}(\mathcal{D}_X)$ consistent (i.e. it must converge in probability to θ^*)
- $\mathcal{D}_X$ i.i.d. and *exhaustive*, i.e., sufficient to characterize $P = P_{\theta^*}$

Regularity conditions

C1 Θ is an open set and

$$\forall \boldsymbol{x} \in \mathcal{X}, \forall (\boldsymbol{\theta}, \boldsymbol{\theta}') \in \Theta^2 \qquad p_{\boldsymbol{\theta}}(\boldsymbol{x}) > 0 \iff p_{\boldsymbol{\theta}'}(\boldsymbol{x}) > 0$$

$p_{\boldsymbol{\theta}} \in \mathcal{H}_\Theta$ have the same support $\mathrm{supp}(p_{\boldsymbol{\theta}})$, $\forall \theta \in \Theta$.

C2 $\forall \boldsymbol{\theta} \in \Theta$, $p_{\boldsymbol{\theta}}$ can be differentiated under the integral [2] :

$$\nabla_\theta \int_{\mathrm{supp}(p_{\boldsymbol{\theta}})} p_{\boldsymbol{\theta}} \cdot \mu(d\boldsymbol{x}) = \int_{\mathrm{supp}(p_{\boldsymbol{\theta}})} \nabla_\theta p_{\boldsymbol{\theta}} \cdot \mu(d\boldsymbol{x}) \qquad (2)$$

---

2. According to [Bil95], C2 holds if the gradient is locally dominated by an integrable function g, i.e., it exist a neighborhood $\square$ of θ and g such that $\int_{\mathrm{supp}(p_{\boldsymbol{\theta}})} g(x)\mu(dx) < +\infty$ such that almost everywhere on a neighborhood V of $\boldsymbol{x}$:

$$\left|\frac{\partial p_{\boldsymbol{\theta}}}{\partial \theta_k}\right| \leq g$$

Theorem

Maximum Likelihood

Given a sequence of i.i.d. random variables $\mathcal{D}_X = \left\{ \boldsymbol{x}_i \in \mathcal{X} \right\}_{i=1}^{N}$, *then :*

$$\forall (\boldsymbol{\theta}, \boldsymbol{\theta}^*) \in \Theta, \boldsymbol{\theta} \neq \boldsymbol{\theta}^*, \forall \varepsilon > 0 \qquad \lim_{N \to +\infty} \mathbb{P}\left[p_{\boldsymbol{\theta}} \left(\mathcal{D}_X \right) - p_{\boldsymbol{\theta}^*} \left(\mathcal{D}_X \right) < \varepsilon \right] = 1$$

- *if* $p \in \mathcal{H}_\Theta$, $\exists \boldsymbol{\theta}^* \in \Theta$ *such that* $p = p_{\boldsymbol{\theta}^*} = \arg\max_{\boldsymbol{\theta} \in \Theta} p_{\boldsymbol{\theta}}(\boldsymbol{X})$
- $\hat{\boldsymbol{\theta}}(\mathcal{D}_X) = \arg\max_{\boldsymbol{\theta} \in \Theta} p_{\boldsymbol{\theta}}(\boldsymbol{\theta}; \mathcal{D}_X) = \arg\max_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{N} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$

Jensen inequality

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\ln\left(\frac{p_{\boldsymbol{\theta}}\left(\boldsymbol{X}\right)}{p_{\boldsymbol{\theta}^*}\left(\boldsymbol{X}\right)}\right)\right] < 0$$

Theorem

Maximum Likelihood of bijection $g(\boldsymbol{\theta})$. *[Bil95]*

Given a sequence of i.i.d. observations $\mathcal{D}_X = \left\{\boldsymbol{x}_i \in \mathcal{X}\right\}_{i=1}^{N}$ *and* $\boldsymbol{\theta}$ *the of parameters that maximizes the likelihood* $p_{\boldsymbol{\theta}}$. *Then* $g\left(\hat{\boldsymbol{\theta}}\right)$ *is the maximum likelihood estimator of* $g(\boldsymbol{\theta})$ *(*g *bijection).*

Theorem

Maximum Log-Likelihood *Given a sequence of i.i.d. random variables* $\mathcal{D}_X = \left\{ \boldsymbol{x}_i \in \mathcal{X} \right\}_{i=1}^{N}$, *then :*

$$\forall \boldsymbol{\theta} \neq \boldsymbol{\theta}^*, \forall \varepsilon > 0 \qquad \lim_{N \to +\infty} \mathbb{P}\left[\frac{1}{N} \ln \left(\frac{p_{\boldsymbol{\theta}} \left(\mathcal{D}_X \right)}{p_{\boldsymbol{\theta}^*} \left(\mathcal{D}_X \right)} \right) < \varepsilon \right] = 1$$

- if $p \in \mathcal{H}_\Theta, \exists \boldsymbol{\theta}^* \in \Theta$ such that $p = p_{\boldsymbol{\theta}^*} = \arg\max_{\boldsymbol{\theta} \in \Theta} \ln p_{\boldsymbol{\theta}} \left(\boldsymbol{X} \right)$
- $\hat{\boldsymbol{\theta}} \left(\mathcal{D}_X \right) = \arg\max_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_X) = \arg\max_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{N} \ln p_{\boldsymbol{\theta}} \left(\boldsymbol{x}_i \right)$

Converging sequence to maximum log-likelihood

- Open neighborhood $O_\varepsilon\left(\boldsymbol{\theta}^*\right)$
- Sequence $\boldsymbol{\theta}_N := \boldsymbol{\theta}\left(\mathcal{D}_X\right)$, $\mathcal{D}_X = \left\{\boldsymbol{x}_i \in \mathcal{X}\right\}_{i=1}^N$

$$S_N := \left\{ \boldsymbol{x}_i \in \mathcal{D}_X \middle| \ln p_{\boldsymbol{\theta}^*}\left(\boldsymbol{x}_i\right) > \sup_{\boldsymbol{\theta} \in \partial O_\varepsilon\left(\boldsymbol{\theta}^*\right)} \ln p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) \right\}$$

$$\forall \varepsilon > 0 \quad \lim_{N \to +\infty} \mathbb{P}\left[\left|\mathbb{P}\left(S_N\right) - 1\right| < \varepsilon\right] = 1 \Longrightarrow \boldsymbol{\theta}^* \in \mathrm{Int}\left(\Theta\right)$$

Likelihood vs Log-Likelihood

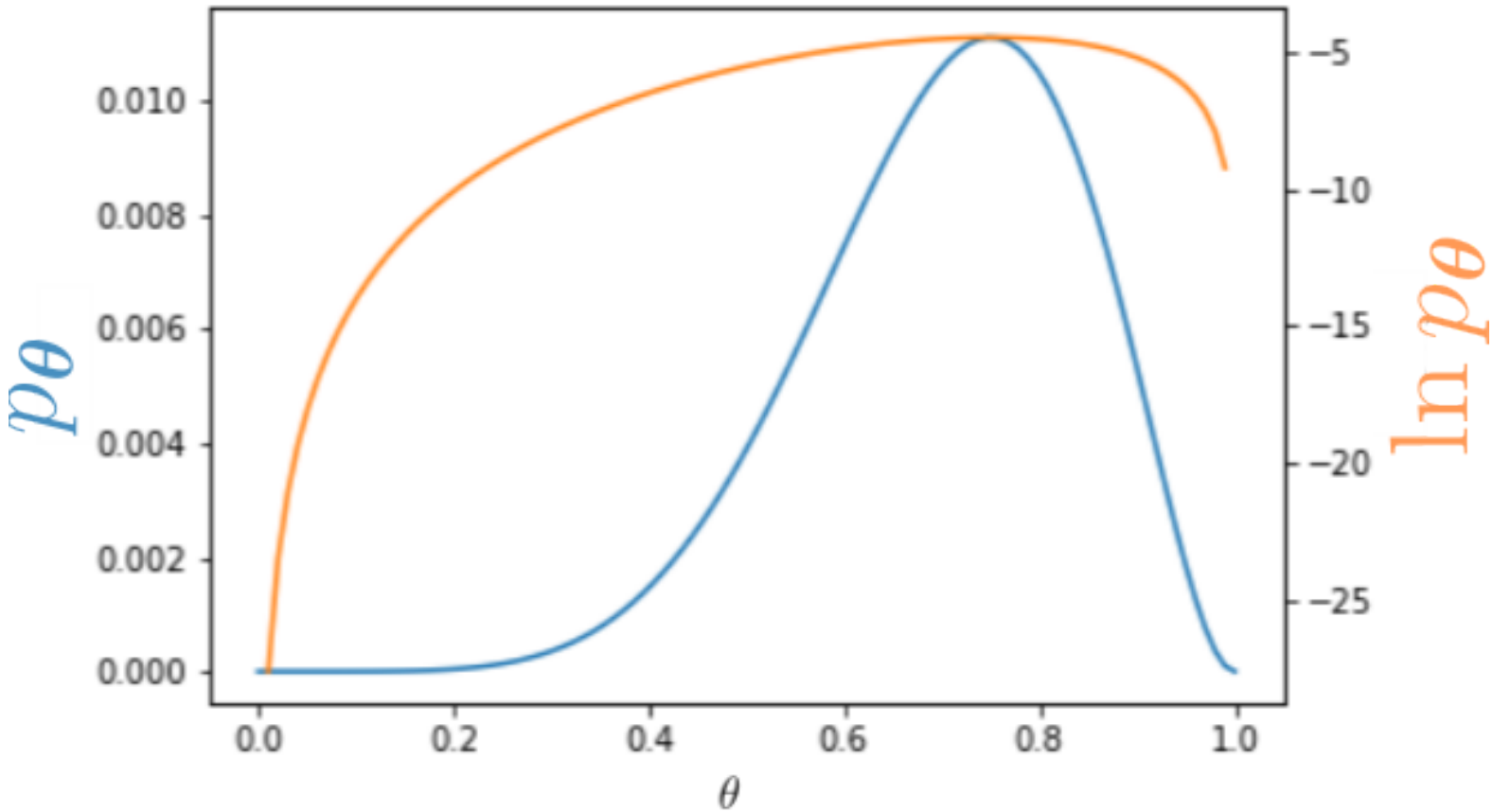


Figure – Bernoulli distribution $p_\theta(\mathcal{D}_X) = \prod_{i=1}^{N} \theta^{1_{x_i=1}} (1-\theta)^{1_{x_i=0}}$.

`stats.stackexchange.com/questions/486007/measuring-predictive-uncertainty-with-negative-log-likelihood-nll`

Convergence to "true" probability distribution

Converging sequence to maximum log-likelihood

- Open neighborhood $O_\varepsilon(\boldsymbol{\theta}^*)$, $\boldsymbol{\theta}_N := \boldsymbol{\theta}(\mathcal{D}_X)$
- Regularity C2

$$\nabla_\theta \int_{\mathrm{supp}(p_{\boldsymbol{\theta}})} p_{\boldsymbol{\theta}} \cdot \mu(d\boldsymbol{x}) = \int_{\mathrm{supp}(p_{\boldsymbol{\theta}})} \nabla_\theta p_{\boldsymbol{\theta}} \cdot \mu(d\boldsymbol{x})$$

- Rolle's theorem : $\exists \hat{\boldsymbol{\theta}}_N \in \bar{O}_\varepsilon(\boldsymbol{\theta}^*)$ such that $\nabla_{\boldsymbol{\theta}} \ln p_{\hat{\boldsymbol{\theta}}_N}(\boldsymbol{x}) = \mathbf{0}$
- Convergence

$$S_N \subset \tilde{S}_N := \left\{ \boldsymbol{x}_i \in \mathcal{D}_X | \exists \hat{\boldsymbol{\theta}}_N \in \bar{O}_\varepsilon(\boldsymbol{\theta}^*) \text{ such that } \nabla_{\boldsymbol{\theta}} \ln p_{\hat{\boldsymbol{\theta}}_N}(\boldsymbol{x}_i) = \mathbf{0} \right\}$$

$$\mathbb{P}(S_N) \leq \mathbb{P}(\tilde{S}_N) \leq 1 \qquad \forall \varepsilon > 0 \quad \lim_{N \to +\infty} \mathbb{P}\left[|\mathbb{P}(\tilde{S}_N) - 1| < \varepsilon \right] = 1$$

Recap

- $\exists(\hat{\boldsymbol{\theta}}_N)_{N\in\mathbb{N}^*} = \hat{\boldsymbol{\theta}}(\mathcal{D}_X) \xrightarrow[N\to\infty]{\mathbb{P}} \boldsymbol{\theta}^*$
- If C2 holds : $\boldsymbol{s}_{\hat{\boldsymbol{\theta}}_N}(\mathcal{D}_X) = \boldsymbol{\nabla}_{\boldsymbol{\theta}} p_{\hat{\boldsymbol{\theta}}_N} = \mathbf{0}$
- Under C1 and C2 holds :

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}} \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}^*}(\boldsymbol{X}) = \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}} \boldsymbol{s}_{\boldsymbol{\theta}^*}(\boldsymbol{X}) = \mathbf{0}$$

- $\ell\left(\boldsymbol{\theta}; \mathcal{D}_X\right) = \ln p_{\boldsymbol{\theta}}(\mathcal{D}_X)$

Fisher Information Matrix at $\boldsymbol{\theta}^*$: auto-covariance of the score

$$\mathbb{I}_F(\boldsymbol{\theta}^*; \boldsymbol{X}) = \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}^*}(\boldsymbol{X}) \otimes \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}^*}(\boldsymbol{X})\right] \geq 0$$

If $\boldsymbol{x}_i$ i.i.d. : $\mathbb{I}_F(\boldsymbol{\theta}^*; \mathcal{D}_X) = N\mathbb{I}_F(\boldsymbol{\theta}^*; X_1)$

Meaning of the FIM (under regularity conditions of p_{θ})

$$\mathbb{I}_F(\boldsymbol{\theta}^*; \mathcal{D}_X) = -\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\mathbf{H}_{\ell}(\boldsymbol{\theta}^*; \mathcal{D}_X)\right] \qquad (p_{\boldsymbol{\theta}} \in C^2(\Theta))$$

Theorem

Cramér-Rao bound *: statistical algorithm to estimate* $\boldsymbol{\theta}^*$*, that reads* $\boldsymbol{\theta} = \mathrm{A}\left(\boldsymbol{x}\right)$ *:*

$$\bar{\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\mathrm{A}\left(\boldsymbol{X}\right)\right]$$

and the variance of the estimator is bounded by below as follows :

$$\mathbb{V}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left(\mathrm{A}\left(\boldsymbol{X}\right)\right) \geq \frac{\|\boldsymbol{\nabla}_{\boldsymbol{\theta}^*}\bar{\boldsymbol{\theta}}\|^2}{N \cdot \mathbb{I}_F(\boldsymbol{\theta}^*; X_1)}$$

A *is unbiased if :* $\mathbb{V}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left(\mathrm{A}\left(\boldsymbol{X}\right)\right) = \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}^*}}\left[\left(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)^2\right] \geq \frac{1}{N \cdot \mathbb{I}_F(\boldsymbol{\theta}^*; X_1)}$

Theorem

MLE convergence in standard normal probability *[Bil95].*

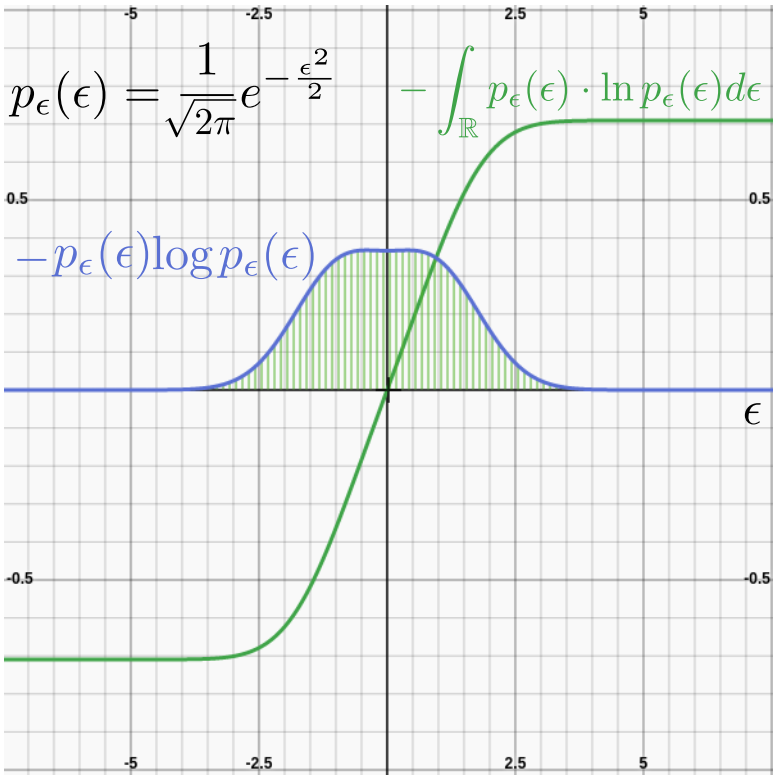
$$\forall \varepsilon > 0 \quad \lim_{N \to +\infty} \mathbb{P}\left[\left| p_{\sqrt{N}(\hat{\theta}_N - \theta^*)} - \mathcal{N}\left(\mathbf{0}, \mathbb{I}_F^{-1}\left(\theta^*; \mathcal{D}_X\right)\right) \right| < \varepsilon \right] = 1$$

The MLE is an asymptotically "optimum" estimator :

$$\forall \varepsilon > 0 \quad \lim_{N \to \infty} \mathbb{P}\left[\left| \frac{1}{N \cdot \mathbb{I}_F(\theta^*; X_1) \cdot \mathbb{V}_{\boldsymbol{x} \sim p_{\theta^*}}\left(\mathrm{A}\left(\boldsymbol{X}\right)\right)} - 1 \right| < \varepsilon \right] = 1$$

Shannon's differential entropy (Jaynes correction)

$$\mathbb{H}_d(p(\boldsymbol{X})) = -\int_{\mathcal{X}} p(\boldsymbol{x}) \cdot \ln\left(\frac{p(\boldsymbol{x})}{m(\boldsymbol{x})}\right) \cdot \mu(d\boldsymbol{x})$$



$$\mathbb{H}_d(\mathcal{N}_{(\boldsymbol{\mu},\boldsymbol{\Sigma})}(\boldsymbol{X})) = \frac{d_X}{2} \ln 2\pi + \frac{1}{2} \mathrm{Tr}\left(\boldsymbol{\Sigma}\right)$$

$(d_X = 1, \boldsymbol{\Sigma} = 1 \longrightarrow \mathbb{H}_d = 1.418938533204673)$

- The conditional entropy of two discrete random variables X and Y reads :

$$\mathbb{H}_d(p(\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}_i)) = -\mathbb{E}_{\boldsymbol{y}\sim p(\boldsymbol{y}|\boldsymbol{x})}\left[\ln(p(\boldsymbol{Y}|X=\boldsymbol{x}_i))\right]$$

and

$$\mathbb{H}_d(p(\boldsymbol{Y}|\boldsymbol{X})) = -\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\mathbb{H}_d(\boldsymbol{p}(Y|\boldsymbol{X}=\boldsymbol{x}))\right] = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p(\boldsymbol{x},\boldsymbol{y}_i)}\left[\ln p(\boldsymbol{Y}|\boldsymbol{X})\right]$$

- The joint entropy of two discrete random variables X and Y reads :

$$\mathbb{H}_d(p(\boldsymbol{X},\boldsymbol{Y})) = -\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p(\boldsymbol{x},\boldsymbol{y})}\left[\ln(p(\boldsymbol{X},\boldsymbol{Y}))\right]$$

$$\mathbb{H}_d(p(\boldsymbol{X},\boldsymbol{Y})) = \mathbb{H}_d(p(\boldsymbol{Y}|\boldsymbol{X})) + \mathbb{H}_d(p(\boldsymbol{X})) = \mathbb{H}_d(p(\boldsymbol{X}|\boldsymbol{Y})) + \mathbb{H}_d(p(\boldsymbol{Y}))$$

- Kullback-Leibler divergence :

$$0 \leq \mathbb{D}_{KL}(p\|q) = \mathbb{E}_{x\sim p}\left[\ln \frac{p}{q}\right] < +\infty \quad \mathbb{D}_{KL}(p\|q) \geq \mathbb{E}_{x\sim p}\left[\frac{p-q}{p}\right]$$

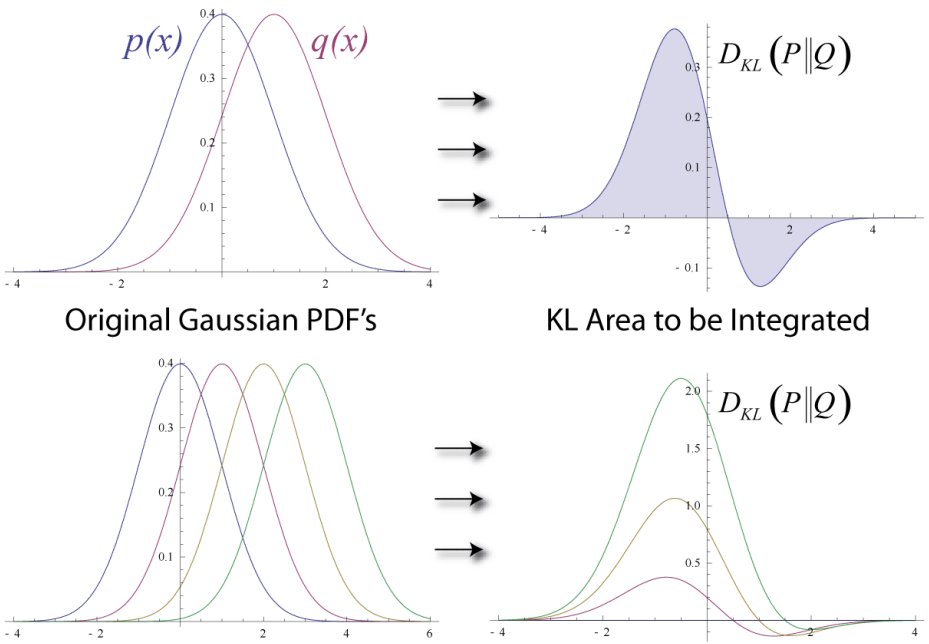


Figure – $\mathbb{D}_{KL}(p\|q) = 0$ implies that $p = q$*a.e.* if and only if $\text{supp}(p) \cap \text{supp}(q) \neq \emptyset$

- Mutual information :

$$\mathbb{I}(p(\boldsymbol{X}, \boldsymbol{Y})) = \mathbb{D}_{KL}\left(p(\boldsymbol{X}, \boldsymbol{Y}) \| p(\boldsymbol{X}) \cdot p(\boldsymbol{Y})\right) \geq 0$$

$$\mathbb{H}(p(\boldsymbol{Y})) - \mathbb{H}(p(\boldsymbol{Y}|\boldsymbol{X})) = \mathbb{I}(p(\boldsymbol{X}, \boldsymbol{Y})) = \mathbb{H}(p(\boldsymbol{X})) - \mathbb{H}(p(\boldsymbol{X}|\boldsymbol{Y}))$$

- $\sup_{\boldsymbol{Y}} \mathbb{I}(p(\boldsymbol{X}, \boldsymbol{Y})) = \mathbb{H}_d(p(\boldsymbol{X})) - \inf_{\boldsymbol{Y}} \mathbb{H}_d(p(\boldsymbol{Y}|\boldsymbol{X})) \to 0$, i.e., forcing the *disentanglement* between $\boldsymbol{X}$ and $\boldsymbol{Y}$ [Che+16].

Theorem

Gibbs-Boltzmann theorem *: available observations* $\left(\boldsymbol{x}_k, y_k\right)$

$$\begin{cases} \max\limits_{q\in\mathcal{H}} \mathbb{H}_d(q) = \max\limits_{q\in\mathcal{H}} \left(-\int_{\mathcal{X}} \ln q\left(\boldsymbol{x}\right) \cdot q\left(\boldsymbol{x}\right) d\boldsymbol{x} \right) & (3) \\ c_k\left(q\right) = \mu_{y_k} - \int_{\mathcal{X}} y_k\left(\boldsymbol{x}\right) \cdot q\left(\boldsymbol{x}\right) d\boldsymbol{x} = 0, \quad c_k : \mathbb{R}^n \to \mathbb{R}, 1 \leq k \leq K & (4) \end{cases}$$

if it exists it is defined as $p_{\boldsymbol{\theta}} \in \mathcal{H}$*, which reads :*

$$p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = \underset{q\in\mathcal{H}, \boldsymbol{c}:\boldsymbol{c}=\boldsymbol{0}}{\arg\max}\ \mathbb{H}_d(q) = \frac{e^{-\sum_{k=1}^{K} y_k(\boldsymbol{x})\cdot\theta_k}}{Z}$$

with $Z = \int_{\mathcal{X}} e^{-\sum_{k=1}^{K} \theta_k \cdot y_k(\boldsymbol{x})} \cdot \mu\left(d\boldsymbol{x}\right)$ *a normalization constant. Moreover,* $\mathbb{H}_d\left(p_{\boldsymbol{\theta}}\right) \geq \mathbb{H}_d\left(p\right)$.

$$p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = \frac{e^{-\sum_{k=1}^{K} y_k(\boldsymbol{x})\cdot\theta_k}}{Z}$$

Log-Likelihood

$$\ln p_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = -\ln Z - \left\langle \boldsymbol{y}\left(\boldsymbol{x}\right), \boldsymbol{\theta}\right\rangle$$

$$\boldsymbol{s}_{\boldsymbol{\theta}}\left(\boldsymbol{x}\right) = \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}}\left[\boldsymbol{y}\left(\boldsymbol{X}\right)\right] - \boldsymbol{y}\left(\boldsymbol{x}\right) = \boldsymbol{\mu}_{\boldsymbol{y}} - \boldsymbol{y}\left(\boldsymbol{x}\right)$$

$$\mathbf{H}_{\ell}(\boldsymbol{\theta}; \boldsymbol{X}) = -\mathbb{C}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}}\left(\boldsymbol{y}\left(\boldsymbol{X}\right)\right)$$

- Unbiased score : $\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\theta}}}\left[\boldsymbol{s}_{\boldsymbol{\theta}}\left(\boldsymbol{X}\right)\right] = \mathbf{0}$
- FIM is the covariance of the observations $\boldsymbol{y}\left(\boldsymbol{X}\right)$
- $\boldsymbol{\theta}_N$ convergence in $\mathcal{N}\left(\mathbf{0}, \mathbb{I}_F^{-1}\left(\boldsymbol{\theta}^*; \mathcal{D}_X\right)\right) \in \mathcal{H}_{\boldsymbol{\theta}}$

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \max_{p_{\boldsymbol{\theta}}\in\mathcal{H}_{\Theta}} \mathbb{H}_d\left(p_{\boldsymbol{\theta}}\left(Y|\boldsymbol{X}\right)\right)$$

Bibliography

Bibliography I

[Bil95] **Billingsley, Patrick**. *Measure and Probability*. John Wiley et Sons : New York, 1995.

[Che+16] **Chen, Xi et al.** "InfoGAN : Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets". en. In : arXiv :1606.03657 (juin 2016). arXiv :1606.03657 [cs, stat]. URL : http://arxiv.org/abs/1606.03657.

[ES16] **Eldan, Ronen et Shamir, Ohad**. "The power of depth for feedforward neural networks". In : *Workshop and Conference Proceedings*. T. 49. PMLR, 2016, p. 1-34.

[Heu+] **Heusel, Martin et al.** "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". en. In : ().

[Mai99] **Maiorov, V.E.** "On Best Approximation by Ridge Functions". en. In : *Journal of Approximation Theory* 99.1 (juill. 1999), p. 68-94. ISSN : 00219045. DOI : 10.1006/jath.1998.3304. URL : https://linkinghub.elsevier.com/retrieve/pii/S0021904598933044.

[MP43] **McCulloch, Warren S. et Pitts, Walter**. "A logical calculus of the ideas immanent in nervous activity". en. In : *The Bulletin of Mathematical Biophysics* 5.4 (déc. 1943), p. 115-133. ISSN : 0007-4985, 1522-9602. DOI : 10.1007/BF02478259. URL : http://link.springer.com/10.1007/BF02478259.

[Ros57] **Rosenblatt, Murray**. "Some Purely Deterministic Processes". In : *Journal of Mathematics and Mechanics* 6.6 (1957), p. 801-810. ISSN : 00959057, 19435274. URL : http://www.jstor.org/stable/24900623.

[Sal+18] **Salehinejad, Hojjat et al.** "Recent Advances in Recurrent Neural Networks". en. In : arXiv :1801.01078 (fév. 2018). arXiv :1801.01078 [cs]. URL : http://arxiv.org/abs/1801.01078.