Butlletí de la Societat Catalana de Matemàtiques
Vol. 36, núm. 1, 2021. Pàg. 5–67. DOI: 10.2436/20.2002.01.97

Aprenentatge algorísmic i xarxes neuronals profundes

Joan Bruna i Sebastià Xambó

Resum: En aquest article trobareu una descripció de la natura de l'aprenentatge algorísmic, així com de les seves modalitats més rellevants, i una presentació dels principals ingredients matemàtics que li serveixen de base, tant per a la definició i l'estudi de models com per a l'anàlisi dels algorismes. També hi trobareu una bibliografia extensa i unes recomanacions per aprofundir en el seu estudi.

Paraules clau: aprenentatge algorísmic, malefici dimensional, xarxes neuronals, gradient descendent estocàstic, dimensió VC, complexitat de Rademacher, doble descens, causalitat i explicabilitat, xarxes hipercomplexes.

Classificació MSC2010: 68T01, 68Q32, 68T05, 62F15, 68W25, 82C32, 62M45, 65K10, 90C26, 93E35, 30G35.

Índex

Introducció

En aquesta secció expliquem, en un llenguatge poc formal, algunes de les idees bàsiques de l'*aprenentatge algorísmic* (AA), o *aprenentatge automàtic*, locucions a les quals atribuïm, com a primera aproximació, el mateix significat que a l'anglesa *machine learning* (ML), és a dir, la recerca d'algorismes eficients que tornen, a partir d'un conjunt de dades (experiència), predictors acurats d'informacions associades a noves dades.

En termes usuals, l'AA forma part de l'àmbit de l'anomenada *intel·ligència artificial* (IA, o AI en les sigles angleses) i comprèn l'estudi de les *xarxes neuronals* (XN, o NN en les sigles angleses). Pel que fa a referències generals sobre aquestes qüestions, vegeu l'apèndix B (pàgina 48). Aquest apèndix conté un apartat per a cada secció de l'article, començant per §B.0, en què podeu trobar una llista de tractats sobre les qüestions que ens ocuparan d'ara endavant. En cadascun d'aquests apartats podeu trobar, ultra les citacions inserides a la secció corresponent d'aquest article, referències complementàries, comentaris bibliogràfics i, quan escau, breus definicions, a l'inici de l'apartat, dels conceptes que es poden considerar una mica especialitzats.

Hi ha diverses menes d'AA. En aquesta secció només considerarem el cas de l'AA *supervisat*. El propòsit d'aquesta modalitat és la creació i la investigació d'algorismes que tornin una funció $f\colon \mathcal{X} \to \mathcal{Y}$ amb capacitat acceptable de predir els valors d'una funció $f^*\colon \mathcal{X} \to \mathcal{Y}$, desconeguda per l'algorisme i a la qual anomenarem *expert* o *supervisor*, a partir d'una sèrie d'exemples, és a dir, d'un conjunt finit de parells

$$\mathcal{D} = \{(x^j, y^j) : x^j \in \mathcal{X},\ y^j = f^*(x^j) \in \mathcal{Y},\ j \in [n]\}, \quad [n] = 1, \dots, n.$$

Sovint ens referirem a un tal algorisme com l'*aprenent*, ja que la seva tasca és aconseguir una extrapolació dels exemples que predigui raonablement bé els valors produïts per l'expert. És important assenyalar que per al tractament d'incerteses sobre les y^j cal recórrer al model més general del qual donem notícia a l'apartat B.2.

Hi ha dues menes de problemes d'AA supervisat que tenen una rellevància cabdal: *classificació*, quan $\mathcal{Y}$ és finit (conjunt de *classes*), i d'*interpolació* o *regressió*, quan $\mathcal{Y} = \mathbb{R}$. Formalment, podem tractar els dos casos posant $f(x) \simeq f^*(x)$ per indicar $f(x) = f^*(x)$ en el cas de classificació i $f(x) \approx f^*(x)$ en el cas de regressió, on $\approx$ denota igualtat aproximada segons un criteri prefixat. Amb aquesta convenció, podem mesurar la bondat de f amb la *taxa d'aprenentatge*, és a dir, la proporció a de casos en què $f(x^j) \simeq f^*(x^j) = y^i$ ($j = 1, \dots, n$), i amb la *fidelitat* [accuracy],[1] que també en podem dir *taxa de generalització* o *capacitat predictora*, és a dir, la proporció g de casos en què $f(x) \simeq f^*(x)$ ($x \in \mathcal{X}$). Així, $1 - a$ és la taxa d'*errors d'aprenentatge* i $1 - g$, la taxa d'*errors de generalització*. En la pràctica, g se substitueix per la proporció

1 Atès que molts termes són propostes de traducció al català de paraules o locucions angleses, sovint aquestes seran consignades entre claudàtors i en lletres sans-serif just després d'aquelles.

de casos en què $f(\bar{x}^j) \simeq \bar{y}^j$, on $\overline{\mathcal{D}} = \{(\bar{x}^j, \bar{y}^j = f^*(\bar{x}^j)) : \bar{x}^j \in \mathcal{X},\ j \in [\bar{n}]\}$ és un conjunt d'exemples (*exemples de test*) generat independentment de $\mathcal{D}$, i un dels nostres propòsits és analitzar les condicions de validesa d'aquest procediment.

Adonem-nos que assolir una taxa d'aprenentatge $a = 1$ equival a una memorització perfecta dels exemples $\mathcal{D}$, cosa que comporta, en els models més coneguts, un comportament pobre davant d'exemples nous, que en la pràctica vol dir davant dels exemples $\overline{\mathcal{D}}$. Per tant, hem d'esperar que el millor ús dels exemples $\mathcal{D}$ sovint només es pugui produir amb un valor adient de la taxa d'aprenentatge. Més enllà d'aquest valor, apareix el *sobreaprenentatge* [overfitting] i, per sota, el *subaprenentatge* [underfitting]. Trobar aquest valor òptim de la taxa d'aprenentatge és un dels mèrits que ha de tenir un AA. Aquests conceptes s'il·lustren i es comenten a la figura 1.

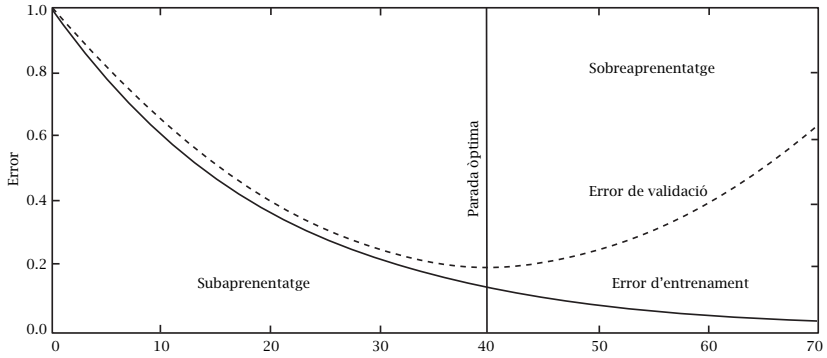

FIGURA 1: Estem suposant que l'aprenentatge, com ara d'una regressió lineal, té lloc en un bucle, els cicles del qual s'anomenen *èpoques* (eix d'abscisses), i que en cada cicle la taxa d'aprenentatge sobre les dades augmenta. Al principi també disminueix l'error de validació, que és com estimem la capacitat generalitzadora, però arriba un moment que aquest error torna a créixer. Aquest moment és el de parada òptima. Abans, encara s'està en règim de subaprenentatge i després de sobreaprenentatge, en què els cicles augmenten la memorització de les dades d'entrenament, fins i tot els aspectes irrellevants (com ara la presència d'algun tipus de «soroll»), en detriment de la capacitat de generalització. Aquest esquema serà modificat a §5 per incloure algunes subtileses del problema de generalització descobertes en els darrers anys (figura 9).

És un fet que es coneixen tècniques per crear algorismes amb bones taxes d'aprenentatge i generalització, almenys per a certs problemes. Actualment aquest fet és notícia gairebé cada dia i en diversos dominis, i és especialment notori, per la cobertura mediàtica, l'èxit d'algorismes campions en jocs i competicions de tota mena. Entre aquests, darrerament han cridat l'atenció els casos del go i del pòquer i, més recentment, el cas d'algorismes que guanyen en jocs de consola. Però encara potser són més espectaculars els resultats en els àm-

bits de les imatges (predictors biomèdics, reconeixement de persones, vehicles autònoms) i de les llengües (parla o escriptura), sobretot per les implicacions que tenen per a les dinàmiques socials i econòmiques arreu.

Exemple: Regressió lineal regularitzada Suposem que les dades x són vectors i que els valors y són números reals. Sigui w un vector (desconegut) de *pesos* (un pes per a cadascuna de les n components dels vectors x), i posem $f^* = f_w$, on

$$f_w(x) = w \cdot x = w_1 x_1 + \cdots + w_n x_n$$

(una *suma ponderada* dels valors de x). Per tal d'optimitzar la taxa d'aprenentatge, podem prendre com a vector w el que ens proporciona el mètode de *mínims quadrats*:

$$\operatorname{argmin}_w \sum_{i=1}^{n} (w \cdot x^i - y^i)^2. \tag{1}$$

Fixem-nos que el mínim és nul si, i només si, $w \cdot x^i = y^i$ per a tot i, que equival a dir que la taxa d'aprenentatge és 1. Per tenir a ratlla el sobreaprenentatge, i millorar així la capacitat predictora, es pot recórrer a una versió *regularitzada* de (1), la qual cosa vol dir escollir w d'acord amb la relació

$$\operatorname{argmin}_w \sum_i (w \cdot x^i - y^i)^2 + \lambda ||w||_2^2, \tag{2}$$

on λ és un número real positiu que en la pràctica és determinat per l'AA. Remarquem que un λ petit afavoreix el sobreaprenentatge, mentre que amb un λ gran fa que $|w|$ hagi de ser petit i, per tant, w té menys possibilitats d'assolir una bona taxa d'aprenentatge. La solució de les expressions (1) i (2) es pot obtenir per algorismes estàndard d'optimització (vegeu §4).

Exemples de classificació Potser un dels més familiars és el problema de filtrar missatges de correus electrònics segons que siguin *brossa* [spam] o no. Un altre exemple és el reconeixement de caràcters manuscrits, en què la capacitat predictora té la flexibilitat d'adaptar-se a la infinitat de variacions possibles. D'especial interès per als matemàtics poden ser els programes que transformen la imatge d'un text matemàtic en una versió LaTeX d'aquest (vegeu `https://mathpix.com/`), o aplicacions per a mòbil que donen informació sobre objectes usant la seva fotografia, o que transformen la parla en un idioma en un text escrit i que a més poden traduir-lo, tant la veu com l'escrit, a un altre idioma.

Regressió logística Diem que la probabilitat $p = p(x)$, $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, segueix una *regressió logística* si el log de la *possibilitat* [odds] de x, $p/(1-p)$, segueix una regressió lineal:

$$\log \frac{p}{1-p} = w_1 x_1 + \cdots + w_d x_d = w \cdot x.$$

Trobats els pesos w d'aquesta regressió lineal (normalment, usant taules de resultats previs en què els vectors de característiques apareixen amb les seves freqüències), llavors obtenim

$$p = \frac{1}{1 + e^{-w \cdot x}},$$

una relació que pot servir per fer prediccions sobre la probabilitat de qualsevol nou vector de característiques. La funció $\sigma(t) = 1/(1+e^{-t})$ té forma sigmoide i es coneix també amb el nom de *funció logística* (figura 2). De vegades convé usar la funció $2k\sigma(t) - k = k(1-e^{-t})(1-e^{-t})^{-1}$, que té forma sigmoide però variant entre els extrems $\pm k$, i que podem denotar σ_k. Per exemple, $\sigma_{1/2} = \sigma - 1/2$ varia entre $-1/2$ i $1/2$, mentre que σ_1 varia entre -1 i 1.

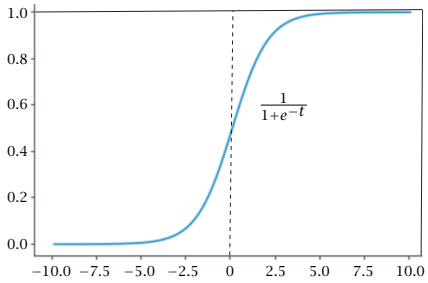

FIGURA 2: Gràfic de la funció logística.

Organització del contingut A la primera secció s'apleguen algunes idees de caire relativament elemental, amb molta tradició en els estudis de matemàtiques i estadística, que es poden veure com a arrels dels models matemàtics de l'aprenentatge i que en tot cas segueixen tenint un paper en les teories més actuals.

El model bàsic en què basem la resta de l'article s'exposa a §2. El repte principal de l'AA davant de les ingents quantitats de dades de què es disposa és l'anomenat *malefici dimensional* [curse of dimensionality], que considerem breument en el primer apartat, 2.1. Els ingredients del model bàsic i la seva significació es descriuen a l'apartat 2.2, que acaba amb el que anomenem *teorema fonamental* de l'AA. En els dos darrers apartats es descriuen models matemàtics de les neurones i de les xarxes neuronals i de la seva funció com a aproximadors universals.

L'estat de l'art dels fonaments teòrics de l'AA es desgrana a les tres seccions següents: §3 (Aproximació), §4 (Optimització) i §5 (Generalització). Atesa la importància dels temes, ens remetem a la introducció de cadascuna per analitzar-ne l'abast amb més detall.

A la darrera secció, §6, fem referència a alguns altres models prometedors. També esmentem diversos problemes oberts. A l'apèndix A recollim algunes nocions de probabilitat emprades anteriorment. Havent ja explicat el propòsit de l'apèndix B, només ens queda dir que l'article acaba amb la secció de referències, que conté la bibliografia citada.

Finalment, cal indicar que ometem pràcticament totes les demostracions dels resultats que hi incloem, però que procurem compensar-ho amb referències adients, sigui en el context de cada enunciat o en l'apartat corresponent de l'apèndix B.

1 Preludis

Encara que el tema principal d'aquest article sigui l'AA basat en xarxes neuronals, i els resultats en què es basa, ens ha semblat adient incloure aquesta secció preliminar per donar compte d'algunes nocions bàsiques que han tingut, i segueixen tenint, un paper important en l'evolució de les teories de l'aprenentatge i el desenvolupament del que s'anomena *ciència de les dades* [data science]. La motivació principal de la secció és fer de coixí entre coneixements més o menys familiars i les seccions posteriors, de natura més tècnica. Per a referències, vegeu §B.1.

1.1 La via bayesiana

La ben coneguda *regla de Bayes-Laplace*, sovint anomenada simplement *fórmula de Bayes*, és de fet un dels resultats pioners i fonamentals de la teoria de l'aprenentatge.

Donats esdeveniments X, Y, tenim que

$$P(X,Y) = \begin{cases} P(X)\cdot P(Y|X), \\ P(Y)\cdot P(X|Y), \end{cases}$$

i aquestes relacions són equivalents a

$$\frac{P(X|Y)}{P(X)} = \frac{P(Y|X)}{P(Y)} = \frac{P(X,Y)}{P(X)\cdot P(Y)}.$$

Denotem amb $L(X,Y)$ aquest valor, que és simètric en X i Y. Llavors podem escriure

$$P(Y|X) = P(Y)L(X,Y),$$

que és la *regla de Bayes-Laplace*. Veiem que $L(X,Y)$ és el factor pel qual hem de multiplicar la probabilitat $P(Y)$ de Y, prèvia a l'observació de X [prior], per obtenir la probabilitat $P(Y|X)$ de Y posterior a l'observació de X. En altres termes, $L(X,Y)$ indica el grau d'aprenentatge sobre Y aportat pel fet d'haver experimentat X.

Classificació bayesiana: la regla MAP Si un esdeveniment X ha estat observat, i aquest fet es pot atribuir a una de les hipòtesis $Y_1,\dots,Y_r$, la regla MAP (màxim *a posteriori*) selecciona la Y_j que maximitza $P(Y_j|X)$. Per la regla de Bayes-Laplace, això equival a escollir la Y_j que maximitza $P(X|Y_j)P(Y_j)$. En el cas especial en què les Y_j són equiprobables, es tracta de maximitzar $P(X|Y_j)$.

Un cas particular és el d'una classificació binària dels objectes de $\mathcal{X}$. Suposem $\mathcal{X} = \mathcal{X}_0 \sqcup \mathcal{X}_1$, però que l'observació d'un $x \in \mathcal{X}$ no determina amb certesa, generalment per la presència de soroll o altres causes, a quina de les dues classes pertany. El model estadístic d'aquesta situació és que l'atribució d'una classe $y \in \{0,1\}$ a una observació x es regeix per una probabilitat conjunta $P(x,y) = P(x|y)P(y)$. És a dir, suposem que les probabilitats $P(0)$ i $P(1)$ són conegudes, així com les probabilitats $P(x|y)$ que expressen la incertesa sobre la classe y de x. Quantificada així la incertesa, la regla de decisió òptima és l'anomenat *classificador de Bayes*, $f^B\colon \mathcal{X} \to \{0,1\}$, definit per la màxima probabilitat posterior $P(y|x) = P(x|y)P(y)/P(x)$, és a dir:

$$f^B(x) = \begin{cases} 0 & \text{si } P(0|x) \geqslant P(1|x), \\ 1 & \text{si } P(0|x) < P(1|x). \end{cases}$$

Aquest classificador és òptim per a la distribució P i la seva proporció d'errors és $R^* = \int \min\{P(0|x), P(1|x)\}P(x)\,dx$ en mitjana. Sense més informació sobre P, només es pot assegurar que $0 \leqslant R^* \leqslant 1/2$.

La perspectiva bayesiana de l'AA Partint de les dades $\mathcal{D}$, d'un possible model $\mathcal{M}$ per explicar-les i d'un conjunt de paràmetres o pesos W del model, la regla de Bayes-Laplace ens dona

$$P(W|\mathcal{D},\mathcal{M}) = \frac{P(\mathcal{D}|W,\mathcal{M})P(W|\mathcal{M})}{P(\mathcal{D}|\mathcal{M})},$$

on $P(W|\mathcal{D},\mathcal{M})$ és la probabilitat de W posterior a haver observat les dades $\mathcal{D}$ suposant el model $\mathcal{M}$. A la dreta de la igualtat tenim $P(\mathcal{D}|W,\mathcal{M})$, la versemblança de les dades $\mathcal{D}$ segons el valor dels pesos W relatius al model $\mathcal{M}$; $P(W|\mathcal{M})$, la probabilitat *a priori* dels paràmetres W respecte del model $\mathcal{M}$; i $P(\mathcal{D}|\mathcal{M})$, la versemblança marginal o evidència del model, és a dir, $\int P(\mathcal{D}|W,\mathcal{M})P(W,\mathcal{M})\,dW$. La probabilitat posterior esdevé la probabilitat *a priori* per processar la informació adquirida en considerar noves dades.

Un model $\mathcal{M}$ ja après es pot usar per predir la probabilitat de noves dades $\mathcal{D}'$:

$$P(\mathcal{D}'|\mathcal{D},\mathcal{M}) = \int P(\mathcal{D}'|W,\mathcal{D},\mathcal{M})P(W|\mathcal{D},\mathcal{M})\,dW.$$

El paradigma també permet contrastar diferents models $\mathcal{M}$ en relació a les dades $\mathcal{D}$:

$$P(\mathcal{M}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M})P(\mathcal{M})}{P(\mathcal{D})}, \quad P(\mathcal{D}|\mathcal{M}) = \int P(\mathcal{D}|W,\mathcal{M})P(W|\mathcal{M})\,dW.$$

1.2 Anàlisi de components principals (PCA)

Sigui X una matriu real $m \times n$. Podem pensar que X és el resultat d'observar m característiques de n objectes, de manera que la fila X^j conté les n observacions $(x_1^j, \dots, x_n^j)$ de la característica j-èsima ($j \in [m]$). Equivalentment, la columna $X_k = (x_k^1, \dots, x_k^m)^T$ ($k \in [n]$) conté els valors de les m característiques de l'objecte k-èsim. En tot cas, $X = [X_1, \dots, X_n] = [X^1, \dots, X^m]$.

Considerem la mitjana de X^j, $\mu_j = E[X^j]$, i la covariància de X^j i X^k, $\sigma_{jk} = \mathrm{Cov}(X^j, X^k) = E[X^j \cdot X^k] - \mu^j \mu^k$. La matriu simètrica $\Sigma = \mathrm{Cov}(X) = (\sigma_{jk})$, de tipus $m \times m$, és semidefinida positiva i diem que és la matriu de covariància de X. Adonem-nos que $\sigma_{jj} = \mathrm{Var}(X^j)$ (la variància de X^j), i que $\mathrm{Var}(X^j) = \sigma_j^2$, on $\sigma_j \geqslant 0$ és la desviació estàndard de X^j.

Donat un m-vector unitari u, resulta que $\mathrm{Var}(u^T X) = u^T \Sigma u$, i que *aquest valor és màxim precisament quan u és un vector propi u_1 de Σ amb valor propi màxim*. En aquestes condicions, es diu que $u = u_1$ és la component principal de X. Notem que $u^T X$ és el vector format amb les projeccions de les columnes de X sobre u, de manera que aquestes projeccions recullen la variabilitat màxima de X en una direcció.

La segona component principal de X és el vector propi unitari u_2 corresponent al segon valor propi (en ordre no creixent) de Σ. Aquest vector maximitza $\mathrm{Var}(u^T X) = u^T \Sigma u$ per als vectors unitaris u perpendiculars a u_1. Prosseguint aquest procés, s'obté una base ortonormal $u_1, \ldots, u_m$ de $\mathbb{R}^m$ tal que $\mathrm{Var}(u_r^T X) = u_r^T \Sigma u_r$ és màxima per a vectors unitaris perpendiculars a $u_1, \ldots, u_{r-1}$. Si posem U_r per denotar la matriu formada pels vectors $u_1, \ldots, u_r$, la matriu $U_r^T X$ és de tipus $r \times n$ i incorpora la variabilitat de X atribuïble als primers r valors propis (ordenats en sentit no creixent) de Σ.

La tècnica de les components principals és un primer exemple de reducció dimensional, un concepte al qual retornarem més endavant, i es pot entendre com una forma d'aprenentatge no supervisat. També cal dir que $U_r^T X$ es pot emprar com una forma de preprocessament aplicable a les dades X abans de sotmetre-les a un procediment d'aprenentatge supervisat. El valor de r s'escull de manera que $u_{r+1}, \ldots, u_m$ tinguin un paper negligible en l'explicació de la variabilitat de X.

1.3 Descomposició en valors singulars (SVD)

Sigui X una matriu real $m \times n$, que considerem com a dades a l'estil del que hem vist a §1.2, i sigui r el seu rang. Les matrius simètriques XX^T i $X^T X$, de tipus $m \times m$ i $n \times n$ respectivament, tenen rang r, són semidefinides positives i tenen els mateixos valors propis no nuls $\lambda_1^2, \ldots, \lambda_r^2$, on podem suposar que $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r > 0$. A més, si posem U i V per denotar les matrius ortonormals de vectors propis de XX^T i $X^T X$, llavors $X = U \Lambda V^T$, essent $\Lambda_{jj} = \lambda_j$, per a $j = 1, \ldots, r$, els únics valors no nuls de Λ. Amb això, $XX^T = U(\Lambda\Lambda^T)U^T$ i $X^T X = V(\Lambda^T \Lambda)V^T$, on els primers r valors de les diagonals de $\Lambda\Lambda^T$ i $\Lambda^T\Lambda$ són $\lambda_1^2, \ldots, \lambda_r^2$ i tots els altres són nuls en ambdues matrius (de tipus $m \times m$ i $n \times n$, respectivament). Ara, atès que $U\Lambda = (\lambda_1 u_1, \ldots, \lambda_r u_r)$, obtenim la celebrada *descomposició en valors singulars de X*:

$$X = \lambda_1 u_1 v_1^T + \cdots + \lambda_r u_r v_r^T.$$

De fet, resulta que per a $k = 1, \ldots, r$, la matriu

$$M_k = \lambda_1 u_1 v_1^T + \cdots + \lambda_k u_k v_k^T$$

és l'aproximació òptima de X amb matrius de rang k (teorema d'Eckart-Young). Aquest resultat és també, doncs, una forma de reducció dimensional. El valor de k s'escull de manera que $X - M_k$ sigui negligible en el context on s'utilitzi.

Una aplicació important de la descomposició singular és que la solució a per mínims quadrats del sistema lineal $Xa = b$ és $a = X^\dagger b$, on $X^\dagger$ és la pseudoinversa de Moore-Penrose de X, és a dir, $X^\dagger = V\Lambda^\dagger U^T$, amb $\Lambda^\dagger_{jj} = \lambda_j^{-1}$ per a $j = 1, \dots, r$ i totes les altres entrades nul·les.

1.4 Aprenentatge per mètodes no paramètrics

Si bé el focus principal d'aquest article és l'aprenentatge supervisat, té interès presentar breument la idea d'aprenentatge no supervisat. Ho fem descrivint dos algorismes: k-mitjana [k-Means], un procediment d'agrupació de dades, i un de classificació, k-NN [nearest neighbors], en què les dades disponibles estan classificades per un expert i la decisió s'aconsegueix mirant les k dades disponibles més properes a la dada per classificar.

***k*-mitjana** En termes generals, aquest algorisme divideix un conjunt de vectors n-dimensionals no etiquetats $\mathcal{D}$ en k grups minimitzant una certa funció de cost, però l'essència del seu funcionament queda ben reflectida en els tres passos següents: (1) Escollim k vectors diferents $z_1, \dots, z_k$ (en principi, poden ser del mateix conjunt de dades). (2) Assignem cada vector x^j de $\mathcal{D}$ al z_i més proper, és a dir, el primer que satisfà $d(x^j, z_i) = \min(d(x^j, z_1), \dots, d(x^j, z_k))$, de manera que en resulta una classificació inicial de $\mathcal{D}$ en k grups. (3) Actualitzem cada z_i prenent el baricentre del grup de z_i. (4) Iterem fins que els z_i siguin estables (segons una tolerància prefixada).

***k*-NN** Suposem que tenim un conjunt de dades $\mathcal{D} = \{(x^j, y^j)\}$ ($j \in [n]$), on les x^j són vectors N-dimensionals i les y^j, elements d'un conjunt finit Y. Sigui k un número enter positiu. Donat un vector x arbitrari de l'espai dels x^j, l'algorisme k-NN el classifica posant-li una etiqueta de la manera següent: (1) busca els k vectors $x^{j_1}, \dots, x^{j_k}$ de manera que les distàncies $d(x, x^{j_1}), \dots, d(x, x^{j_k})$ siguin les més petites possibles, i (2) assigna a x la moda del conjunt $\{y^{j_1}, \dots, y^{j_k}\}$. Per trobar el conjunt $\{x^{j_1}, \dots, x^{j_k}\}$, basta fer una llista dels parells $(j, d^j = d(x, x^j))$, ordenar les d^j de manera no decreixent i retenir els primers k parells $(j_1, d^{j_1}), \dots, (j_k, d^{j_k})$. També cal dir que si en el pas (2) hi ha empat en el nombre de dos o més grups, cal desempatar triant el que doni distància mínima. En el cas d'una classificació binària, és aconsellable escollir k senar.

L'algorisme k-NN es pot modificar sense dificultat de manera que k creixi amb n. El teorema següent proporciona condicions suficients perquè aquest algorisme modificat sigui un classificador universalment consistent.

TEOREMA 1 ([187, 207]). *Sigui f_n el classificador binari construït amb una mostra de n punts. Si $n \to \infty$ i $k \to \infty$ de manera que $k/n \to 0$* [e. g., si $k = \log n$], *llavors $R(f_n) \to R^*$ per a qualsevol distribució de probabilitat P, on $R(f_n)$ i R^* són la proporció d'errors de f_n i del classificador de Bayes, respectivament.*

Tanmateix, aquest teorema no té les conseqüències pràctiques que podria semblar a primera vista, ja que oculta un cas de malefici dimensional, en el sentit que per aconseguir $R(f_n) - R^* \leqslant \varepsilon$ cal, en el pitjor dels casos, un nombre d'exemples $n = O(\varepsilon^{-d})$ [176]. Veurem més detalls d'aquest malefici dimensional a §2.1.

2 Aprenentatge inductiu en gran dimensió

El problema de l'aprenentatge inductiu, tal com l'hem considerat a les seccions precedents, es redueix essencialment a un problema d'interpolació de dades. És per tant un problema clàssic, amb una llarga història en estadística i tractament del senyal. Què fa que la matemàtica de l'AA sigui especial, doncs?

2.1 El malefici dimensional

La clau és la gran dimensió dels espais de dades. Senyals com les imatges, els àudios i els vídeos viuen en espais de milions de dimensions, i els mètodes clàssics d'interpolació resulten inservibles en aquest règim, ja que, per obtenir solucions consistents sota hipòtesis febles de regularitat local (com ara Lipschitz), és inevitable que el nombre d'exemples n hagi de créixer de forma exponencial en la dimensió —un fenomen conegut col·loquialment com el *malefici dimensional* [curse of dimensionality]. En efecte, la divisió del cub unitat d'un espai de dimensió d en cubs de costat ε conté ε^{-d} cubs, és a dir, una quantitat que creix exponencialment amb la dimensió. Una observació relacionada és que el volum de l'esfera de radi 1 en dimensió d decreix molt ràpidament a partir de $d = 5$ (vegeu la figura 3), de manera que les propietats familiars vàlides en el pla, en l'espai o en dimensions baixes no serveixen en dimensions grans. Ras i curt, el malefici dimensional rau en el fet que la interpolació de punts en dimensions grans és intrínsecament més difícil que en dimensions baixes.

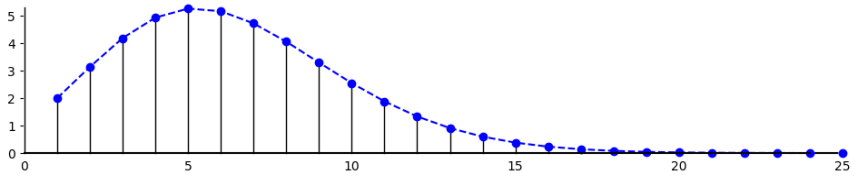

FIGURA 3: Volum v_n de l'esfera S^n de radi 1 per a $n = 1, 2, \ldots, 25$. Tenim que $v_1 = 2$, $v_2 = \pi$ i, per a $n > 2$, $v_n = \frac{2\pi}{n} v_{n-2}$. Així v_n creix quan $n < 2\pi$ i decreix monòtonament per a $n > 2\pi$. De fet, v_n convergeix molt ràpidament a 0 quan $n \to \infty$. Per exemple, $v_{100} \simeq 2.368 \times 10^{-40}$.

Per afrontar el malefici dimensional cal, doncs, construir teories matemàtiques adaptades al tractament d'objectes de gran dimensió. A diferència dels exemples de regressió lineal i logística de les seccions anteriors, primer és necessari introduir espais d'aproximació no lineals. Això planteja nous reptes

algorísmics per poder definir estimadors òptims en aquests espais. Entre les dificultats principals, cal esmentar la presència inevitable de funcions no convexes i la necessitat d'usar eines avançades de probabilitat (com ara les tècniques de concentració de la mesura) que garanteixin l'aprenentatge.

L'objectiu d'aquesta secció és presentar primer, §2.2, el model bàsic d'AA supervisat que usarem a partir d'ara com a teló de fons i escenari per a totes les consideracions que seguiran. A continuació, §2.3, s'introdueixen els models de neurona i de xarxes neuronals, cosa que en particular ens permet disposar d'una gran abundància d'espais d'hipòtesis (un dels elements clau del model bàsic), i garantir la propietat d'aproximació universal, §2.4.

L'estudi més detallat del paper del model bàsic per superar els obstacles del malefici dimensional l'hem sistematitzat a les seccions següents, §3, §4 i §5, dedicades als problemes d'aproximació, optimització i generalització, respectivament.

2.2 Model bàsic

Tot seguit descrivim els ingredients que concorren en el model matemàtic bàsic de l'aprenentatge inductiu. Com a primer exemple, podeu tenir en compte la regressió lineal explicada a la introducció.

Domini de les dades: $\mathcal{X}$, l'espai (o conjunt) del qual s'extreuen les dades. La teoria de la informació permet, en molts casos, que aquestes dades es puguin representar com a vectors d'un espai vectorial real. Per exemple, per a les imatges de N píxels, $\mathcal{X}$ és (una regió de) $\mathbb{R}^N$ o de $\mathbb{R}^{3N}$ segons que les imatges siguin monocromes o de color RGB [Red-Green-Blue]. La dimensió d'aquests espais, per a les imatges que ordinàriament tenen interès, és molt gran. En general, doncs, hem d'estar preparats per haver de treballar amb espais $\mathcal{X}$ de dimensió molt gran.

Generació de dades: La generació de dades es regeix per la selecció aleatòria d'elements de $\mathcal{X}$ d'acord amb una distribució de probabilitat P sobre $\mathcal{X}$. Aquesta distribució no és coneguda per l'AA i en general està molt lluny de la distribució uniforme. Per exemple, les imatges que usualment ens trobem presenten regularitats que les distingeixen d'aquelles formades per píxels seleccionats a l'atzar segons la distribució uniforme i de manera independent. I una observació similar val per al cas dels sons, com ara la música.

Espai d'hipòtesis: És un espai $\mathcal{F}$ de funcions $f\colon \mathcal{X} \to \mathbb{R}$ que considerem P-mesurables. La selecció d'un tal espai es coneix com a *biaix inductiu*, ja que expressa les suposicions a priori sobre la forma esperada de la solució destil·lada per l'algorisme.

L'espai $\mathcal{F}$ sovint s'especifica com un espai de funcions parametritzades. En el cas de la regressió lineal, l'espai de funcions era el dels polinomis multivariats de grau 1. A l'apartat següent veurem que les neurones i les xarxes neuronals es poden veure com a factories de funcions parametritzades no lineals, la qual cosa és el tret més visible de la seva utilitat per a l'AA.

Complexitat: La mesura de la *complexitat* de les hipòtesis és una funció $\gamma\colon \mathcal{F} \to \mathbb{R}^+$. Exemple: $\gamma(f) = ||f||$ (la norma de f) quan $\mathcal{F}$ és un espai de Banach. Per a tot $\delta \in \mathbb{R}^+$, posem $\mathcal{F}_\delta = \{f \in \mathcal{F} : \gamma(f) \leqslant \delta\}$. En el cas de la norma, és la bola (tancada) de $\mathcal{F}$ de radi δ, que té l'avantatge de ser un conjunt convex. Aquesta noció permet facultar l'AA amb un criteri per graduar l'esforç de cerca segons una complexitat creixent. En el cas de les xarxes neuronals, aquesta complexitat es controla mitjançant una *penalització* implícita a través de l'algorisme d'optimització (§4 i §5.4) o d'una *regularització* explícita, com ara una penalització sobre els pesos de la xarxa.

Expert o supervisor: Una funció donada $f^*\colon \mathcal{X} \to \mathbb{R}$. Per a cada dada x, l'expert produeix un *exemple*: el parell (x, y), $y = f^*(x)$. L'AA no coneix l'expert (raó per la qual aquest també s'anomena *oracle*), però el seu propòsit és imitar-lo tan fidelment com sigui possible seguint els procediments que es descriuen més avall. En cas que l'expert sigui un element de $\mathcal{F}$ (cosa que en general no passa), diem que és *realitzable*.

Dades d'entrenament: Un conjunt d'exemples

$$\mathcal{D} = \{(x^i, y^i = f^*(x^i)) : x^i \in \mathcal{X})\}_{i\in[n]}$$

produïts per l'expert, on els x^i es generen segons la distribució P de manera independent, $x^i \sim P$ en símbols.

Defecte: El *defecte* [loss] de $f \in \mathcal{F}$ és un indicador, denotat $L(f)$, de la separació entre f i f^*. Un exemple usual és l'esperança de $|f(x) - f^*(x)|^2$ relativa a P, $\mathbb{E}_P|f(x) - f^*(x)|^2$. Més generalment, podem emprar $L(f) = \mathbb{E}_P\, \ell(f(x), f^*(x))$, on $\ell\colon \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ és una funció no negativa amb $\ell(y, y') = 0$ si, i només si, $y = y'$ (diem que ℓ és un *defecte puntual*). En alguns dominis d'aplicació, del defecte se'n diu *risc* i, encara en d'altres, *error*.

Objectiu de l'AA Si bé l'objectiu primari és aproximar l'expert f^*, en termes del model bàsic es tracta de trobar un estimador $\hat{f} \in \mathcal{F}$, usant només $\mathcal{D}$, tal que el seu defecte $L(\hat{f})$ sigui una bona aproximació del *mínim defecte* assolible amb funcions de $\mathcal{F}$, és a dir, de $L_\mathcal{F} = \min_{h\in\mathcal{F}} L(h)$. Aquest estimador es construeix a partir d'una *minimització del defecte empíric* [empirical risk minimization, ERM], on el *defecte* o *risc empíric* [empirical risk] és el funcional $\hat{L}_\mathcal{D}\colon \mathcal{F} \to \mathbb{R}^+$ definit per la fórmula

$$\hat{L}_\mathcal{D}(h) = \frac{1}{n}\sum_{i=1}^{n} |h(x^i) - y^i|^2.$$

Es tracta d'un estimador no esbiaixat de $L(h)$ mitjançant les dades $\mathcal{D}$. Naturalment, l'expressió $|h(x^i) - y^i|^2$ s'ha de substituir per $\ell(h(x^i), y^i)$ si L es defineix en termes d'un defecte puntual ℓ.

Com que hi ha una discrepància (aleatòria) entre el funcional que ens interessa minimitzar (L, desconegut) i el funcional de què disposem ($\hat{L}_{\mathcal{D}}$), és necessari introduir algun tipus de regularització per tal de poder controlar aquestes fluctuacions. Per exemple, podem considerar el defecte empíric δ-*restringit*,

$$\hat{L}_{\mathcal{D},\delta} = \min_{h\in\mathcal{F}_\delta} \hat{L}_{\mathcal{D}}(h), \tag{3}$$

o el defecte empíric λ-*regularitzat* (cf. la discussió de l'equació (2)), o λ-*penalitzat*,

$$\min_{h\in\mathcal{F}} (\hat{L}_{\mathcal{D}}(h) + \lambda\gamma(h)), \tag{4}$$

on $\gamma(h)$ és la complexitat de h introduïda a la pàgina 17.

Com veurem posteriorment, la minimització del risc empíric s'aconsegueix mitjançat mètodes coneguts genèricament com a *algorismes de gradient descendent*.

Descomposició de l'error Donat un estimador $\hat{f}$ obtingut a partir de l'ERM dins d'un espai d'hipòtesis $\mathcal{F}$, l'objectiu de la teoria de l'aprenentatge estadístic és fitar $L(\hat{f}) - L_{\mathcal{F}}$, una quantitat que alguns autors anomenen *penediment* [regret], ja que expressa la discrepància que emetria un oracle que conegués $L(\hat{f})$ i $L_{\mathcal{F}}$. Aquest penediment es pot descompondre de la manera següent ([26, 10]):

$$L(\hat{f}) - L_{\mathcal{F}} = L(\hat{f}) - L_{\mathcal{F}_\delta} + L_{\mathcal{F}_\delta} - L_{\mathcal{F}}.$$

La significació del sumand $L_{\mathcal{F}_\delta} - L_{\mathcal{F}}$ es descriu més avall amb el nom *error d'aproximació* i denotat ε_{apr}. Per altra banda, podem escriure

$$L(\hat{f}) - L_{\mathcal{F}_\delta} = L(\hat{f}) - \hat{L}_{\mathcal{D}}(\hat{f}) + \hat{L}_{\mathcal{D}}(\hat{f}) - \hat{L}_{\mathcal{D},\delta} + \hat{L}_{\mathcal{D},\delta} - L_{\mathcal{F}_\delta}.$$

La diferència $\varepsilon_{\text{opt}} = \hat{L}_{\mathcal{D}}(\hat{f}) - \hat{L}_{\mathcal{D},\delta}$ s'analitza més avall com a *error d'optimització*. Ens queda per considerar la suma $L(\hat{f}) - \hat{L}_{\mathcal{D}}(\hat{f}) + \hat{L}_{\mathcal{D},\delta} - L_{\mathcal{F}_\delta}$, de la qual donarem una fita superior. D'una banda, és clar que $L(\hat{f}) - \hat{L}_{\mathcal{D}}(\hat{f}) \leqslant \varepsilon_{\text{est}}$, definit com $\sup_{h\in\mathcal{F}_\delta} |L(h) - \hat{L}_{\mathcal{D}}(h)|$ i explicat més avall com a *error estadístic* o *de fluctuació*, i de l'altra, també tenim $\hat{L}_{\mathcal{D},\delta} - L_{\mathcal{F}_\delta} \leqslant \varepsilon_{\text{est}}$, ja que si $h \in \mathcal{F}_\delta$ compleix $L_{\mathcal{F}_\delta} = L(h)$, llavors $\hat{L}_{\mathcal{D},\delta} - L_{\mathcal{F}_\delta} \leqslant \hat{L}_{\mathcal{D}}(h) - L(h) \leqslant |L(h) - \hat{L}_{\mathcal{D}}(h)| \leqslant \varepsilon_{\text{est}}$. Podem resumir aquestes consideracions en l'enunciat següent:

TEOREMA 2 (FITACIÓ DEL PENEDIMENT). $L(\hat{f}) - L_{\mathcal{F}} \leqslant \varepsilon_{\text{apr}} + \varepsilon_{\text{opt}} + 2\varepsilon_{\text{est}}$.

Error d'aproximació: L'hem definit com la diferència $\varepsilon_{\text{apr}} = L_{\mathcal{F}_\delta} - L_{\mathcal{F}}$. No depèn de les dades $\mathcal{D}$, és no negatiu i disminueix en créixer δ. Mesura l'aproximació del mínim defecte $L_{\mathcal{F}}$ que es pot aconseguir amb funcions de $\mathcal{F}_\delta$. Per a la discussió sobre el disseny d'espais $\mathcal{F}$ que permetin aproximar f^*, vegeu §3.

Error d'optimització: Per $h \in \mathcal{F}_\delta$, l'hem definit com a $\varepsilon_{\text{opt}} = \hat{L}_{\mathcal{D}}(h) - \hat{L}_{\mathcal{D},\delta}$, on $\hat{L}_{\mathcal{D},\delta}$ és el mínim dels defectes empírics de funcions de $\mathcal{F}_\delta$. En la pràctica, es fixa una tolerància $\varepsilon > 0$ i s'aplica un algorisme de minimització del risc empíric per aconseguir $\hat{f} \in \mathcal{F}_\delta$ tal que el seu error d'optimització sigui $\leqslant \varepsilon$, és a dir,

$$\hat{L}_{\mathcal{D}}(\hat{f}) - \hat{L}_{\mathcal{D},\delta} \leqslant \varepsilon. \tag{5}$$

El problema principal per assolir (5) és computacional, degut al paisatge no convex del risc empíric, i l'estudiem amb detall a §4.

Error estadístic: Per a una $h \in \mathcal{F}_\delta$, $|L(h) - \hat{L}_{\mathcal{D}}(h)|$ és l'error que es comet en substituir el defecte $L(h)$ pel defecte empíric $\hat{L}_{\mathcal{D}}(h)$. En el pitjor dels casos, aquest error és $\varepsilon_{\text{est}} = \sup_{h\in\mathcal{F}_\delta} |L(h) - \hat{L}_{\mathcal{D}}(h)|$, i d'aquesta quantitat en diem *error estadístic* o *error de fluctuació*. Analitzem el paper que té en l'anàlisi de la capacitat de generalització a §5.

El teorema 2 ens diu que podem garantir un bon aprenentatge (és a dir, $L(\hat{f}) - L_{\mathcal{F}}$ petit i, per tant, que $\hat{f}$ sigui una bona aproximació de f^* amb una funció de $\mathcal{F}$), si podem assegurar que els tres termes d'error són petits. Per aconseguir això, cal una bona elecció de $\mathcal{F}$, un bon algorisme d'optimització per tal de poder assolir la desigualtat (5) i eines per controlar les fluctuacions que produeixen l'error estadístic. És a dir, ens plantegem: (1) Com podem trobar espais d'hipòtesis $\mathcal{F}$ amb bones propietats d'aproximació en dimensions grans? (2) De quins algorismes es disposa per aconseguir minimitzar $\hat{L}_{\mathcal{D}}(h)$? (3) Com es poden controlar les fluctuacions estadístiques? A l'estudi d'aquestes qüestions dediquem les seccions §3, §4 i §5, respectivament.

2.3 Neurones i xarxes neuronals

En aquest apartat introduïm la neurona com a eina bàsica per construir espais funcionals amb bones propietats d'aproximació i generalització en gran dimensió.

En AA, una *neurona* (vegeu la figura 4*a*)) és una funció de la forma

$$x \mapsto \chi(\theta, x) = a\sigma(x \cdot w + w_0), \tag{6}$$

on $a, w_0 \in \mathbb{R}$, $w \in \mathbb{R}^d$, $\theta = (a, w, w_0)$, i on σ és una funció de forma *sigmoide*, com ara la *funció logística* $\sigma(t) = (1 + e^{-t})^{-1}$ (figura 2). Diem que θ són els *paràmetres* de la neurona. Més concretament, w és el vector de *pesos* i w_0, el *biaix*. Aquest model computa, per a cada valor dels paràmetres θ, una funció la gràfica de la qual té forma de *talús* [ridge] (vegeu la figura 4*b*)), ja que és constant sobre cada hiperplà perpendicular al vector w i en la direcció de w té forma sigmoide.

Afegint la constant $x_0 = 1$ com un input extra, i prenent $\bar{x} = (1, x)$ i $\bar{w} = (w_0, w)$, tenim que $x \cdot w + w_0 = \bar{x} \cdot \bar{w}$, de manera que si $a = 1$ i σ és la funció logística, la neurona computa la regressió logística de $\bar{x}$ amb pesos $\bar{w}$.

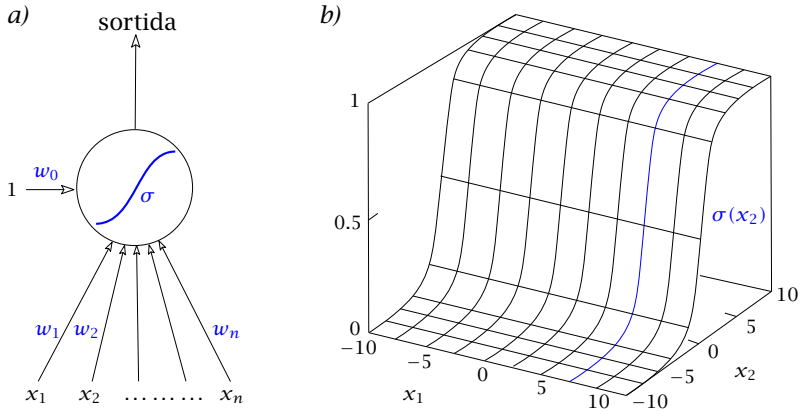

FIGURA 4: *a*) Esquema d'una neurona. *b*) La gràfica de la funció calculada per una neurona té forma de talús. Funcionalment, podem considerar que els pesos w_j i els paràmetres de σ, col·lectivament θ, formen part de la neurona, i que aquesta dona com a valor de sortida $f_\theta(x_1,\dots,x_n)$. Gràficament, representarem aquesta funcionalitat amb un cercle.

Si la neurona representada a la figura 4*a*) es compon amb la funció $s \mapsto \pm 1$ segons que $s = f_\theta(x) \geqslant \frac{1}{2}$ o $< \frac{1}{2}$, essencialment tenim l'anomenat *perceptró de Rosenblatt* (1958), que històricament té la importància de ser el primer sistema que podia aprendre una classificació binària a partir d'exemples per un procés d'ajustament dels paràmetres θ.

Una *xarxa neuronal* (XN) es pot pensar com una *composició de neurones* d'acord amb un graf de connexions que es coneix com l'*arquitectura* de la xarxa. Aquí ens fixarem primordialment en el cas de grafs dirigits, i deixarem per a més endavant alguns comentaris sobre xarxes no dirigides, com ara les de Hopfield o les anomenades *màquines de Boltzmann*. Tampoc considerarem xarxes amb retroalimentació, és a dir, que continguin camins tancats.

L'arquitectura estàndard d'una XN *directa* [feed-forward] és un graf dirigit estructurat en *capes* [layers] L_j, com a la figura 5, i la seva signatura funcional es pot condensar en l'esquema següent:

$$\mathcal{N}: \quad \text{Entrada} = L_0 \to L_1 \to L_2 \to \cdots \to L_m \to L_{m+1} = \text{Sortida}. \tag{7}$$

Convencionalment, una xarxa es considera *soma* [shallow] si $m = 1$ i *profunda* [deep] si $m > 1$. Les capes $L_1,\dots,L_m$ es consideren *ocultes* [hidden], mentre que les capes d'entrada i sortida són *visibles*. L'*amplada* d'una capa és el nombre de neurones que conté.

Cada capa consta d'un cert nombre de neurones i només hi ha connexions de les neurones d'una capa envers neurones de la capa següent. Cada capa rep un senyal x de la capa anterior i produeix un senyal de sortida x' que envia a la capa següent, cosa que podem representar com una relació funcional:

$x' = f_j(x)$. La capa L_0 recull el senyal d'entrada (un so o una imatge, per exemple), paper que presenta una certa analogia amb el dels òrgans sensorials dels éssers vius. La sortida, representada per la capa L_{m+1}, és el resultat d'aplicar progressivament [feedforward] les funcions $f_1, f_2, \ldots, f_m$ (és a dir, la composició $f_m \circ f_{m-1} \circ \cdots \circ f_1$) a l'entrada. Seguint l'analogia biològica, la sortida pot ser el senyal enviat a diversos sistemes de l'ésser viu, com ara l'aparell locomotor, o als òrgans de fonació dels humans, entre molts d'altres.

La funció $f_j \colon x \mapsto x'$ depèn del conjunt de paràmetres (o *pesos*) associats a les connexions neuronals que arriben a L_j, de manera que f_j és una funció parametritzada que podem denotar f_{θ_j} i així l'acció progressiva de la xarxa és la funció parametritzada $f_\theta = f_{\theta_m} \circ \cdots \circ f_{\theta_1}$ ($\theta = (\theta_1, \ldots, \theta_m)$). Atès que les funcions d'activació de les neurones són no lineals, també ho és aquesta funció. El nombre total de paràmetres d'una xarxa neuronal és en general molt gran, especialment, les profundes. Fins on pugui valer l'analogia biològica, els paràmetres de la xarxa fan el paper de les connexions sinàptiques, de les quals s'estima que n'hi ha, en el cas dels cervells humans, desenes de bilions.

El caràcter d'una capa L_j de la XN (7) depèn de la manera en què s'usen els paràmetres θ_j en la definició de f_{θ_j}. Un cas important és el de les capes *convolutives*, un qualificatiu genèric per designar que en la definició de $f_{\theta_j}(x)$ s'usa alguna mena de *convolució*, o de *correlació creuada*, entre θ_j i x, i en aquest cas els paràmetres θ_j fan el paper de *filtres*. Una XN es considera que és *convolutiva* (XNC) si conté almenys una capa convolutiva.

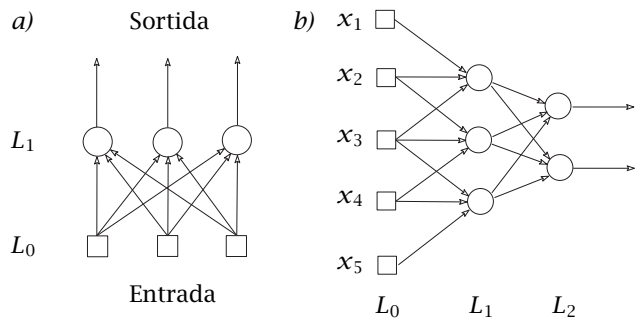

FIGURA 5: *a*) Xarxa sense neurones ocultes i totalment connectada. *b*) Xarxa amb una capa oculta de tres neurones, L_1, totalment connectada a les neurones de sortida, L_2. La capa d'entrada, L_0, només està parcialment connectada a la capa L_1. Com que cada neurona oculta rep les sortides de tres neurones de l'entrada, els pesos de les tres neurones ocultes podrien ser els mateixos, cas en el qual es parla de *pesos compartits*. Aquesta idea és a la base de la noció de xarxes convolutives que veurem en una secció posterior. Si els tres pesos són w_1, w_2, w_3, l'estat de les neurones ocultes és $y_j = \sum_{i=1}^{3} w_i x_{i+j-1}$, $j = 1, 2, 3$, és a dir, el vector d'aquests estats és la convolució del vector d'entrades x amb el vector de pesos w.

2.4 Aproximadors universals

Les xarxes neuronals estàndard tenen la capacitat d'aproximar qualsevol funció mesurable fins a qualsevol precisió desitjada. De fet, això es pot aconseguir amb xarxes somes, però amb un nombre de neurones ocultes que augmenta a mesura que es demana més precisió.

TEOREMA 3 ([49, 91, 15, 156]). *Sigui $\sigma\colon \mathbb{R} \to \mathbb{R}$ una funció contínua, acotada, monòtona creixent i no constant* (e. g. una sigmoide o tanh)*. Llavors l'espai $\mathcal{F}$ de funcions de la forma*

$$\sum_{j=1}^{N} a_j \sigma(w_j \cdot x + b_j), \tag{8}$$

on $N \in \mathbb{Z}^+$, $w_j \in \mathbb{R}^n$, $a_j, b_j \in \mathbb{R}$, és dens (convergència uniforme sobre compactes) en l'espai de funcions contínues de $\mathbb{R}^n$. En particular, resulta que tota funció contínua de X es pot aproximar (uniformement sobre compactes) tant com vulguem mitjançant una xarxa neuronal soma.

Vegeu [156, proposició 6.4] per a una derivació en el marc del teorema de Stone-Weirstrass. Aquest resultat dona confiança al fet que les xarxes neuronals poden aproximar qualsevol funció que tingui interès pràctic, però per si mateix no és un resultat quantitatiu, ja que no incorpora, per exemple, cap control sobre el nombre de neurones de la capa oculta necessari per assolir un cert error d'aproximació. Recordem de l'apartat 2.2 que l'aproximació és només una peça del puzle, i que un bon algorisme d'aprenentatge necessita a més una bona taxa de generalització i d'optimització —propietats que, en la pràctica, requereixen arquitectures més profundes.

3 Aproximació

Un problema fonamental és garantir que l'aprenentatge sigui factible fins i tot quan $n \ll \varepsilon^{-d}$, una situació en la qual els mètodes clàssics no són aplicables (apartat 3.1), però on les xarxes neuronals donen certes esperances. Per encarar aquesta qüestió, ens fixarem primer en el paper de la profunditat (apartat 3.2) i després en la geometria espacial induïda per la física del problema (apartat 3.3).

3.1 El malefici dimensional de l'aproximació

L'anàlisi de funcions $f\colon \Omega \to \mathbb{R}$ definides en un domini Ω de baixa dimensió és una disciplina madura i amb llarga història en les matemàtiques. La regularitat d'una funció en una, dues o tres dimensions pot caracteritzar-se precisament per expressar l'estructura d'un problema específic; per exemple, les solucions d'una equació d'ones, o d'una imatge $f(u)$, $u \in [0,1]^2$, usant eines d'anàlisi harmònica (com ara la transformada de Fourier o les ondetes), o variants geomètriques (com ara les crestetes [ridgelets]). Aquest control permet resoldre tota mena de problemes inversos (com ara la compressió d'imatges, ressonàncies

magnètiques, etc.) amb garanties estadístiques i computacionals. En aquestes aplicacions, podem dir que els models matemàtics estan «al dia» respecte als algorismes utilitzats en la pràctica.

La situació en dimensions grans és totalment diferent: als espais funcionals basats en nocions globals de regularitat, com ara els espais de Sobolev, els manca adaptabilitat quan la dimensió creix: un resultat clàssic en estadística no paramètrica ([197]) estableix que estimar una funció f de la classe de Sobolev en dimensió d amb error de mínims quadrats ε requereix $O(\varepsilon^{-2-d/s})$ mostres, on s és el nombre de derivades de f. Dit d'una altra manera: només les funcions extremament regulars, amb milions de derivades finites, són estimables en el règim de dimensions grans! Nocions locals de regularitat, com ara funcions simplement Lipschitz, també pateixen malefici dimensional, ja que són necessàries ε^{-d} mostres per obtenir errors minimax d'ordre ε [206].

Cal, doncs, treballar amb una teoria d'aproximació alternativa per respondre al repte de les dimensions grans. En aquest context, els espais funcionals associats a les xarxes neuronals, començant per les funcions somes descrites per l'equació (8), han estat estudiats per grans analistes (Pinkus, Matusek o, fins i tot, Bourgain) com a generalitzacions de les representacions integrals en Fourier:

$$f(x) = \int g(\theta)\sigma(x \cdot \theta)\, d\theta, \tag{9}$$

on g és la «transformada» i σ, una activació no polinòmica genèrica que generalitza l'exponencial complexa $\sigma(x \cdot \theta) = e^{-i(x\cdot\theta)}$ de la transformada de Fourier. Un resultat important de [10] és que aquests espais funcionals permeten *adaptar-se* a les estructures de baixa dimensió sense malefici: si $f(x) = \tilde{f}(Ux)$, on $U \in \mathbb{R}^{k\times d}$ amb $k \ll d$ és un operador de projecció desconegut, la regularitat de f s'expressa en termes de $y = Ux \in \mathbb{R}^k$, que té dimensió k potencialment molt més baixa que la dimensió d de l'espai ambient. L'essència d'aquests espais és, un cop més, l'*esparsitat* [sparsity] que es pot imposar a g en la representació (9), en consonància amb resultats fonamentals dels anys 1990 i 2000 [55] que van transformar el tractament de senyal emprant la condició d'esparsitat.

3.2 De xarxes somes a xarxes convolutives profundes

Els arguments matemàtics d'aproximació precedents s'han de contrastar amb els arguments experimentals, que mostren una diferència enorme entre les xarxes somes genèriques i les xarxes convolutives modernes, amb centenars de capes. Es pot dir que l'anàlisi harmònica de les xarxes profundes està encara en la seva infància. De fet, els resultats més profunds sobre això provenen de la comunitat d'informàtica teòrica, que ha estudiat des dels anys setanta qüestions d'aproximació de circuits lògics basant-se en l'anàlisi harmònica de l'hipercub [118].

En el context de xarxes neuronals, alguns autors han estudiat el rol de la profunditat des del punt de vista de l'aproximació. Citem en particular [58], on

els autors presenten una classe de funcions radials i oscil·lants en dimensió d que requereixen un nombre exponencial de neurones per tal que puguin ser aproximades per xarxes somes, fet que dona un contrapunt negatiu al resultat d'aproximació universal del teorema 3, però ensems mostren que es poden aproximar satisfactòriament amb xarxes d'una sola capa oculta.

A més de la profunditat, un altre aspecte crucial de les xarxes neuronals modernes és que incorporen coneixements de la física (en l'estructura espacial d'imatges, vídeos i sons, per exemple), com descrivim tot seguit.

3.3 El rol de la geometria espacial

Fins ara hem prosseguit pensant que $\mathcal{X}$ és $\mathbb{R}^d$, $d \gg 1$, però en les aplicacions més importants (com ara la visió artificial, el llenguatge artificial —text i parla— o la física) disposem de més estructura. En molts casos, els elements de $\mathcal{X}$ són funcions $x\colon \Omega \to \mathbb{R}^k$, on Ω és una regió d'un espai de dimensió baixa i k, un número enter positiu petit. En el cas d'una imatge, Ω pot ser un rectangle del pla i k, el nombre de canals de la imatge, és a dir, $k = 1$ per a imatges monocromes o $k = 3$ per a imatges de color. En el cas de la parla, Ω pot ser $\mathbb{R}$, de manera que $x(t)$ és la intensitat del so en l'instant t, amb $k = 2$ si el so és estèreo. En el cas de la física de N partícules, $\Omega = (\mathbb{R}^3 \times \mathbb{R}^3)^N$, on la partícula i-èsima està determinada pel parell $(p_i, q_i) \in \mathbb{R}^3 \times \mathbb{R}^3$ format pel seu moment lineal p_i i la seva posició q_i. En tots aquests casos, l'alta dimensió de les observacions x camufla l'estructura subjacent de dimensió baixa, i la pregunta clau és com podem usar aquesta estructura (que anomenarem *estructura funcional* de $\mathcal{X}$) per al tractament del problema esmentat.

Simetries globals En AA, les simetries apareixen com a transformacions de les entrades que no afecten la funció que volem aprendre. És una situació que presenta una analogia amb moltes de les matemàtiques i la física. Per exemple, la relació entre el grup de Galois d'un polinomi i l'estructura de les seves arrels, o entre les simetries d'un sistema físic i les lleis de conservació que el regeixen (teoremes de Noether).

Una simetria de $\mathcal{F}$ és una transformació invertible $T\colon \mathcal{X} \to \mathcal{X}$ tal que $f(x) = f(Tx)$ per a qualsevol $f \in \mathcal{F}$. D'aquestes simetries, que formen un grup G amb l'operació de composició, les més significatives per als nostres propòsits són les induïdes a partir de simetries de l'estructura funcional. Més concretament, si $\tau\colon \Omega \to \Omega$ és una transformació, llavors podem considerar $T_\tau\colon \mathcal{X} \to \mathcal{X}$, $x \mapsto x_\tau$, definida per la relació $x_\tau(u) = x(\tau(u))$. Per exemple, en el cas d'una translació $\tau = t_v$, $t_v(u) = u + v$, ens queda $x_{t_v}(u) = x(u + v)$.

Tanmateix, la invariància global no és una premissa prou forta per endegar l'estimació en dimensions altes. Els grups de transformacions són típicament de dimensió baixa; en processament del senyal, sovint són subgrups del grup afí $\mathrm{Aff}(\Omega)$, de dimensió 6 en el cas de les imatges. Es pot definir una premissa molt més forta especificant com es comporta f respecte de pertorbacions geomètriques de Ω properes a elements d'aquests grups de simetries globals.

Deformacions locals i separació d'escales Considerem ara un camp vectorial $\tau\colon \Omega \to \mathbb{R}^d$, prou regular, que produeix una deformació local $\varphi(u) := u + \tau(u)$, la qual actua sobre $L^2(\Omega)$ per composició, $x_\tau(u) := x(\varphi(u))$. Com a exemples tenim translacions locals, canvis del punt d'observació, rotacions i transposicions en freqüència [33], que han estat utilitzats àmpliament com a models de variabilitat d'imatge en visió per computador [96, 62, 70]. Observem que si $\|\tau\| < 1$, llavors φ és un difeomorfisme.

La major part de les tasques en visió per computador són estables per deformacions, en el sentit que la predicció feta per f no canvia gaire si la imatge d'entrada s'ha deformat lleugerament [33, 122]. En tasques que són invariants per translació aquesta premissa pot expressar-se

$$|f(x_\tau) - f(x)| \approx \|\tau\|, \tag{10}$$

per a tot x i tot τ, on $\|\tau\|$ mesura la distància de φ al grup de translacions; per exemple $\|\tau\| := \sup_u \|\nabla\tau(u)\|$ [33]. L'estabilitat per deformacions és una premissa forta, perquè l'espai de deformacions locals té dimensió gran, en contraposició amb el grup de simetries globals, com s'il·lustra a la figura 6. Addicionalment, aquesta estabilitat per deformacions es pot expressar en termes d'una anàlisi multiresolució [MRA] [120] de la forma següent. Una imatge contínua x definida a Ω pot interpretar-se com una combinació lineal $x = x_J + \sum_j \tilde{x}_j$ de senyals $\tilde{x}_j$, $j \in (-\infty, J)$, on x_J captura les estructures a gran escala 2^J, i $\tilde{x}_j$, els *detalls* a escales 2^j més petites, com s'il·lustra a la figura 7. L'anàlisi multiresolució construeix operadors lineals basats en ondetes [wavelets] per obtenir aquesta descomposició $x \mapsto \{x_J, \tilde{x}_j\}$. Similarment, les funcions f d'interès en visió admeten una factorització aproximada en termes d'una anàlisi multiresolució: denotant amb P_j un projector (no necessàriament lineal) cap a l'espai de senyals a escala com a mínim j, l'estabilitat geomètrica de f s'expressa com

$$f(x) \approx f_j(P_j(x)),$$

on f_j és una nova funció definida sobre imatges a una resolució més baixa, limitada a les escales més grans que j. En altres paraules, les interaccions entre píxels es poden «separar» segons les escales, tractant primer les interaccions locals mitjançant l'operador P_J. Aplicant aquesta factorització iteradament a f_j, $f_{j+1}, \ldots$, obtenim una separació de l'aprenentatge a cada escala. Aquest principi de separació d'escales apareix en diversos àmbits de la física matemàtica, per exemple, en el *mètode multipolar ràpid* [fast multipole method] [73]. Aquest principi d'estabilitat per deformacions locals ha estat explotat en visió per computador en diversos models, incloent-hi les xarxes convolutives, i analitzat matemàticament amb la *transformada dispersiva* [scattering transform] [33, 122].

FIGURA 6: Les deformacions d'una imatge x per un camp τ regular no alteren la informació semàntica x_τ. En canvi, deformacions grans (imatge dreta) sí que ho poden fer.

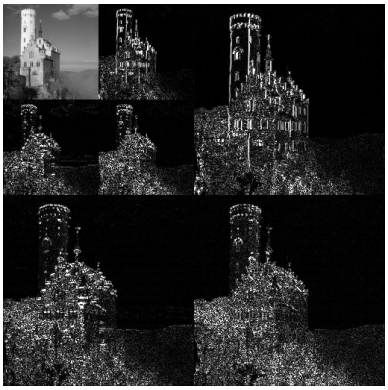

FIGURA 7: Descomposicio multiresolució. (Font: Wikipedia.)

3.4 Representacions amb estabilitat geomètrica

Motivats per la premissa d'estabilitat geomètrica, estem interessats a construir representacions de senyals que hi siguin compatibles. Suposem, per concretar-ho, que l'estimació $\hat{f}$ de f, la funció objectiu, té la forma

$$\hat{f}(x) := \Phi(x) \cdot w, \tag{11}$$

on $\Phi\colon L^2(\Omega) \to \mathbb{R}^K$ designa la representació del senyal i $w \in \mathbb{R}^K$, els coeficients de classificació o regressió. En una XNC, Φ seria l'operador que projecta el senyal d'entrada al darrer nivell de la seva representació dispersant, i w s'associa al darrer senyal de sortida de la xarxa. La representació dispersant s'usa en anàlisi harmònica per a extreure informació de senyals amb certes garanties d'invariància i d'estabilitat per deformacions. La idea central és aplicar de forma iterada una transformada en ondetes amb una transformació no lineal (*rectificació*), que permet capturar interaccions entre diferents escales de la

transformada (d'aquí prové el nom *dispersant*). La relació lineal (11) entre $\Phi(x)$ i $\hat{f}(x)$ implica que l'estabilitat geomètrica en la representació és suficient per garantir un predictor també geomètricament estable. En efecte, si suposem

$$\forall\, x, \tau,\ \|\Phi(x) - \Phi(x_\tau)\| \lesssim \|x\|\|\tau\|, \tag{12}$$

aleshores, per Cauchy-Schwarz

$$|\hat{f}(x) - \hat{f}(x_\tau)| \leqslant \|w\|\|\Phi(x) - \Phi(x_\tau)\| \lesssim \|w\|\|x\|\|\tau\|.$$

Això motiva l'estudi de representacions del senyal que compleixen (12), i que al mateix temps Φ capti prou informació que asseguri que $\|\Phi(x) - \Phi(x')\|$ és gran si $|f(x) - f(x')|$ també ho és. En aquest context, un repte notable de cara a assolir ambdues propietats simultàniament passa per transformar les components en alta freqüència de x de forma estable.

En tasques de reconeixement, a més d'estabilitat geomètrica, és natural demanar estabilitat respecte a la mètrica de $L^2(\Omega)$:

$$\forall\, x, x' \in L^2(\Omega),\ \|\Phi(x) - \Phi(x')\| \lesssim \|x - x'\|. \tag{13}$$

Aquesta propietat d'estabilitat assegura que la presència de soroll additiu en el senyal d'entrada no canviarà dràsticament les característiques de la representació.

Les dues estabilitats desitjades, (12) i (13), també poden ser interpretades en termes de robustesa enfront dels anomenats *exemples adversos* [191]. En efecte, el context genèric dels exemples adversos consisteix a produir petites pertorbacions x' d'un senyal donat x (mesurat en normes convenients) de manera que $|(\Phi(x) - \Phi(x')) \cdot w|$ sigui gran. L'estabilitat de les representacions significa que aquests exemples adversos no poden obtenir-se amb petites pertorbacions additives o geomètriques.

En resum, les propietats d'estabilitat (12) i (13) es poden prendre com a axiomes per definir espais de funcions adaptats. La transformada dispersiva ([33, 121]) defineix un espai de funcions lineals mitjançant l'equació (11), que es pot generalitzar gràcies a la teoria dels nuclis reproductors [24] amb els mateixos principis d'estabilitat. Mencionem breument el fet que aquests principis d'estabilitat geomètrica es poden generalitzar a dominis no euclidians, gràcies a les eines intrínseques en anàlisi harmònica, en varietats riemanianes i en grafs, a més de la teoria de representació de grups, cosa que dona lloc a l'anomenat *aprenentatge geomètric profund* [geometric deep learning]; vegeu [31] per a un resum detallat d'aquest tipus de generalitzacions. Tot i que aquesta visió axiomàtica permet assolir bons resultats numèrics amb mètodes matemàticament rigorosos, destaquem per concloure que defineix espais d'aproximació (lineals) molt més petits que els definits per xarxes neuronals profundes, i que entendre aquest contrast és un dels problemes oberts més importants de la matemàtica de l'AA.

4 Optimització

Més enllà de les qüestions obertes descrites a la secció anterior, en el trencaclosques de l'AA l'aspecte matemàtic possiblement més espinós és computacional:

com podem definir algorismes amb garanties d'optimització del risc empíric en el cas en què la classe de funcions sigui no lineal? Amb aquesta finalitat, revisem el mètode del gradient descendent a §4.1, i tot seguit expliquem primer la connexió amb nuclis reproductors, §4.2, i després, amb els sistemes dinàmics de transport de mesura a §4.3.

4.1 Mètode del gradient descendent

Recordem, del model bàsic, que l'objectiu de l'aprenentatge supervisat consisteix a minimitzar el risc empíric en una classe d'hipòtesis amb capacitat limitada. En la forma penalitzada, això es tradueix en (cf. equació (4))

$$\min_{h\in\mathcal{F}} E(h), \quad E(h) = L_{\mathcal{D}}(h) + \lambda\gamma(h). \tag{14}$$

Aquí estem formulant el problema d'aprenentatge de forma *implícita*, buscant directament funcions de la classe $\mathcal{F}$, per exemple, funcions resultants d'una xarxa neuronal amb una arquitectura predeterminada. El problema (14) és conceptualment simple, però computacionalment complicat, ja que en la pràctica l'espai $\mathcal{F}$ sovint té una geometria no euclidiana. Si pensem $\mathcal{F}$ com un subespai de l'espai de *totes* les funcions $f\colon \mathcal{X} \to \mathbb{R}$, podem interpretar (14) geomètricament com la minimització d'un cost convex en un domini arbitrari.

Tot i que és possible definir mètodes iteratius per optimitzar funcions en aquest nivell de generalitat, en el nostre cas simplificarem el problema introduint una *parametrització* de l'espai d'hipòtesis $\phi\colon \Theta \to \mathcal{F}$, on Θ és un domini euclidià, i on suposem que ϕ és diferenciable i exhaustiva. En el cas d'una xarxa neuronal, $\boldsymbol{\theta} \in \Theta$ representa els paràmetres que cal ajustar, i $\phi(\boldsymbol{\theta})$ és la funció implementada per la xarxa corresponent als paràmetres $\boldsymbol{\theta}$. Aquesta parametrització ens permet definir un domini d'optimització euclidià, al preu que ara la funció per optimitzar és generalment no convexa:

$$\min_{\boldsymbol{\theta}\in\Theta} \tilde{E}(\boldsymbol{\theta}), \quad \tilde{E}(\boldsymbol{\theta}) := E(\phi(\boldsymbol{\theta})). \tag{15}$$

En aquestes condicions, una estratègia clàssica per atacar (15) és el mètode del *gradient descendent* (Cauchy, 1847). Començant per un punt $h_0 = \phi(\boldsymbol{\theta}_0) \in \mathcal{F}$ arbitrari, considerem la iteració

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta\nabla\tilde{E}(\boldsymbol{\theta}_k), \tag{16}$$

on $\nabla\tilde{E}$ és la primera variació o gradient de $\tilde{E}$, i $\eta > 0$ és un paràmetre ajustable corresponent al pas d'optimització. Podem interpretar (16) geomètricament a partir de l'aproximació lineal de $\tilde{E}$ en un entorn de $\boldsymbol{\theta}_k$. Suposant que $\tilde{E}$ és β-regular [β-smooth] i $\eta < \beta^{-1}$, tenim la majoració

$$\tilde{E}(\boldsymbol{\theta}) \leqslant \tilde{E}(\boldsymbol{\theta}_k) + \nabla\tilde{E}(\boldsymbol{\theta}_k)\cdot(\boldsymbol{\theta} - \boldsymbol{\theta}_k) + \eta^{-1}\|\boldsymbol{\theta} - \boldsymbol{\theta}_k\|^2,$$

la qual ens duu a reescriure (16) en forma variacional:

$$\boldsymbol{\theta}_{k+1} = \operatorname{argmin}_{\boldsymbol{\theta}}\ \tilde{E}(\boldsymbol{\theta}_k) + \nabla\tilde{E}(\boldsymbol{\theta}_k)\cdot(\boldsymbol{\theta} - \boldsymbol{\theta}_k) + \eta^{-1}\|\boldsymbol{\theta} - \boldsymbol{\theta}_k\|^2. \tag{17}$$

L'algorisme, doncs, explota la regularitat local de $\tilde{E}$ per definir un model quadràtic majorant la funció, i trobar un punt en l'entorn euclidià del paràmetre actual on es pugui reduir l'error. El mètode del gradient descendent admet diverses generalitzacions en les quals no entrarem en aquest article, per exemple, explotant aproximacions de segon ordre o generalitzacions en mètriques no euclidianes. Cal esmentar que en el cas convex, el mètode del gradient descendent gaudeix d'una anàlisi matemàtica precisa, que permet garantir que el mètode, en la seva versió accelerada [138], és òptim respecte a un model de complexitat oracle que utilitza combinacions lineals de gradients passats [137], en el sentit que la taxa de convergència $\Theta(1/t^2)$ no pot millorar-se. Per a més detalls sobre el mètode, referim el lector a l'excel·lent monografia [34].

En el nostre context, el resultat crucial és que el mètode del gradient descendent permet trobar mínims locals de $\tilde{E}$ sense malefici dimensional: per a un error $\varepsilon > 0$ donat, es requereixen $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ iteracions de gradient descendent (apropiadament modificat en presència de soroll) per trobar un mínim local aproximat a menys de ε [97], on $\widetilde{\mathcal{O}}(f)$ significa $\mathcal{O}(f)$ ignorant factors logarítmics.

Malgrat aquest comportament robust davant el malefici dimensional, el mètode del gradient descendent, tal com l'hem descrit a (16), és problemàtic en aplicacions de gran escala, ja que cada actualització dels paràmetres reclama el gradient del risc empíric $\tilde{E}$, que depèn de totes les dades. Això ha motivat una recerca fructífera en mètodes de gradient estocàstic, començant pel treball seminal de Robbins i Monro als anys cinquanta, [163]. Sense entrar en detalls tècnics, el mètode del gradient estocàstic substitueix l'oracle del gradient [gradient oracle] per un oracle estocàstic: donat el punt actual $\boldsymbol{\theta}$, l'algorisme té accés a un vector aleatori $\boldsymbol{\theta}'$ tal que l'esperança condicional $\mathbb{E}(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \nabla\tilde{E}(\boldsymbol{\theta})$; dit d'una altra manera, tenim accés a un estimador sense biaix del gradient de la funció. En el cas del risc empíric, aquest estimador s'obté avaluant el risc en un sol punt, $\nabla R_i(\boldsymbol{\theta})$, on $R_i(\boldsymbol{\theta}) = |\phi(\boldsymbol{\theta}, x_i) - y_i|^2 + \lambda\gamma(\phi(\boldsymbol{\theta}))$, on $\phi(\boldsymbol{\theta}, x_i) = \phi(\boldsymbol{\theta})(x_i)$, o en un subconjunt petit de punts (mini-batch). En aquestes condicions, és possible establir garanties sobre la complexitat d'iteració similars a les del cas determinista, de l'ordre de $\widetilde{\mathcal{O}}(d\varepsilon^{-4})$ iteracions estocàstiques per trobar un mínim local aproximat a menys de ε [97].

Mencionem en particular que, en el cas de les XN, una variant popular del mètode del gradient descendent considera passos de gradient *adaptatius*, on l'hiperparàmetre η se substitueix per un pas adaptat a cada coordenada de $\boldsymbol{\theta}$. Entre les diverses alternatives, la més popular és *Adam* [100], on el pas de gradient es normalitza en cada coordenada a partir d'una estimació de la norma i es combina amb un component d'inèrcia [momentum].

En resum, en el règim on manipulem espais d'hipòtesis amb graus de llibertat enormes, sense cap altra estructura que la regularitat local de la parametrització, el mètode del gradient descendent estocàstic es perfila com l'eina canònica per aprendre, amb una complexitat d'iteració redimida del malefici dimensional. Tanmateix, aquesta anàlisi només garanteix que és relativament fàcil de trobar un mínim local, no un mínim global. És possible certificar la convergència global del risc empíric en el context de xarxes neuronals? A l'apartat següent veurem que, en alguns casos, la termodinàmica pot donar respostes positives.

4.2 Xarxes neuronals ultraparametritzades i nuclis tangents

El mètode del gradient descendent (16) es pot interpretar com una discretització d'Euler d'una equació diferencial ordinària

$$\dot{\boldsymbol{\theta}} = -\nabla \tilde{E}(\boldsymbol{\theta}),$$

amb condició inicial aleatòria $\boldsymbol{\theta}(0)$ donada per una certa distribució de probabilitat definida sobre l'espai de paràmetres Θ.

Aquesta equació, anomenada *flux del gradient*, defineix una dinàmica contínua $\boldsymbol{\theta}(t) \in \Theta$, $t \geqslant 0$, en l'espai de paràmetres, que al seu torn genera una dinàmica en l'espai funcional $\mathcal{F}$ definida per $h(t) = \phi(\boldsymbol{\theta}(t))$; vegeu figura 8. La regla de la cadena implica immediatament un flux de gradient equivalent a $\mathcal{F}$, donat per

$$\dot{h} = -\mathcal{K}(t) \cdot \nabla E(h(t)), \tag{18}$$

on $\mathcal{K}(t) := D\phi(\boldsymbol{\theta}(t))^\top D\phi(\boldsymbol{\theta}(t))$ és un canvi de mètrica anomenat *nucli tangent*, que projecta el gradient funcional $\nabla E(h(t))$ a l'espai tangent de la varietat $\mathcal{F}$, i on $D\phi(\boldsymbol{\theta})$ és la derivada de ϕ en el punt $\boldsymbol{\theta}$. La dificultat per analitzar (18) matemàticament rau en la variació temporal del nucli tangent $\mathcal{K}(t)$.

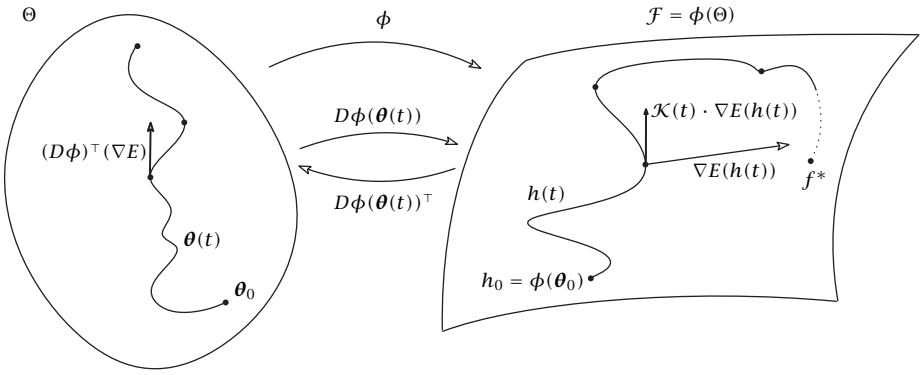

FIGURA 8: Algorisme del gradient descendent: relacions geomètriques i analítiques entre l'espai de paràmetres Θ i l'espai d'hipòtesis $\mathcal{F}$. En general, f^* està fora de $\mathcal{F}$.

De fet, la variació temporal de $\mathcal{K}(t)$ es pot entendre geomètricament com la curvatura de l'espai de funcions $\mathcal{F}$ en un entorn del punt $h(t)$. En el cas que $\phi(\boldsymbol{\theta})$ sigui una xarxa neuronal com a l'equació (7), un resultat important de [95, 57, 56, 45] és que aquesta curvatura desapareix progressivament quan la xarxa es torna més ampla, sota una certa normalització dels pesos i, per tant, $\mathcal{K}(t) \to \mathcal{K}(0)$ uniformement en temps finit [95]. Per exemple, en el cas d'una xarxa soma, si considerem

$$\phi(\boldsymbol{\theta}, \cdot) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \chi(\theta_j, \cdot),$$

on $\chi(\theta,\cdot)$ és la *neurona* definida a (6), i els paràmetres θ_j es consideren independents i idènticament distribuïts segons una distribució μ_0, el nucli tangent esdevé

$$\begin{aligned}\mathcal{K}(t)[x,x'] &= \frac{1}{m}\sum_{j=1}^{m}\nabla_\theta\chi(\theta_j(t),x)\nabla_\theta\chi(\theta_j(t),x')\\ &\to \mathbb{E}_{\theta\sim\mu_0}[\nabla_\theta\chi(\theta,x)\nabla_\theta\chi(\theta,x')]\\ &:= \overline{\mathcal{K}}(x,x') \quad (m\to\infty).\end{aligned}$$

En condicions força generals [160], el nucli tangent a una amplada finita m es concentra uniformement cap a $\overline{\mathcal{K}}$, amb fluctuacions de l'ordre $\sim \frac{1}{\sqrt{m}}$. En aquest règim asimptòtic $m\to\infty$, i considerant per simplificar el cas no regularitzat $E(h)=\frac{1}{2}\|h-f^*\|^2$, la dinàmica d'entrenament se simplifica a

$$\dot f = -\overline{\mathcal{K}}\nabla E(f(t)) = -\overline{\mathcal{K}}(f(t)-f^*),$$

que correspon a la dinàmica lineal associada a un model de regressió lineal en un espai de Hilbert generat pel nucli *reproductor* $\overline{\mathcal{K}}$. Aquests espais funcionals (RKHS en anglès) són generalitzacions en dimensió infinita d'espais euclidians, la qual cosa permet un estudi aprofundit i precís de les qüestions d'aproximació, generalització i optimització. En contrapartida, aquests espais sofreixen del malefici dimensional, com s'ha explicat a l'apartat 3.1.

Amb aquesta parametrització, les xarxes neuronals amples es comporten doncs com a models lineals, caracteritzats per un nucli tangent que no es mou de la inicialització. Si bé aquest fenomen permet comprendre el perquè del bon comportament del mètode del gradient descendent, deixa un regust amarg, ja que no explica matemàticament els avantatges dels models no lineals definits per les xarxes, ni permet evitar el malefici dimensional. Com veurem tot seguit, és possible capturar els aspectes no lineals de les xarxes amb un model convex —canviant la normalització.

4.3 Límits termodinàmics i xarxes somes com a sistemes de partícules

Fixem-nos ara en el cas de les xarxes somes (una sola capa oculta) i considerem una nova normalització (en $1/m$ enlloc d'$1/\sqrt{m}$) dels pesos de la forma

$$\phi(\boldsymbol{\theta},x) = \frac{1}{m}\sum_{j=1}^{m}\chi(\theta_j,x), \tag{19}$$

on $\boldsymbol{\theta}=(\theta_1,\dots,\theta_m)$ i la neurona $\chi(\theta,\cdot)\colon \mathcal{X}\to\mathbb{R}$ és una funció parametritzada per $\theta\in\Omega\subseteq\mathbb{R}^m$. Per exemple, $\chi(\theta,x)=a\sigma(x\cdot b+c)$, amb $\theta=(a,b,c)\in\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}$. Gràcies a l'estructura particularment simple de les xarxes somes, podem pensar la funció ϕ com una mitjana de m funcions simples $\chi(\theta_j)$, parametritzades per m *partícules* $\theta_1,\dots,\theta_m\in\Omega$. La visió paramètrica d'una

xarxa de neurones en termes dels seus pesos (o partícules) correspon en el llenguatge de mecànica de fluids a un model lagrangià del sistema, que contrasta amb el model eulerià, que expressa el sistema en termes de la *densitat de partícules* μ, definida com una mesura de probabilitat sobre l'espai Ω. Efectivament, podem reescriure (19) com

$$\phi(\mu, x) = \int_\Omega \chi(\theta, x)\mu(d\theta), \tag{20}$$

prenent $\mu = \mu^{(m)}$ com la mesura empírica associada a les m partícules:

$$\mu^{(m)}(d\theta) = \frac{1}{m}\sum_{j=1}^{m} \delta_{\theta_j}(d\theta).$$

La perspectiva euleriana permet abstreure'ns de l'aspecte discret de les xarxes somes, en el sentit que l'essència de l'entrenament no és comprendre la trajectòria específica de cada neurona, sinó la trajectòria de la funció que defineixen conjuntament quan en considerem un nombre m suficientment gran. Matemàticament, això correspon a l'exemple canònic de concentració de la mesura, la llei dels grans nombres, que diu que si cada neurona θ_j s'inicialitza independentment a partir d'una llei μ_0, aleshores la mesura empírica $\mu^{(m)}$ convergeix feblement a μ_0 amb taxa $1/\sqrt{m}$:

$$\mathbb{E}\left[\int g(\theta)(\mu^{(m)}(d\theta) - \mu_0(d\theta))\right] \lesssim \frac{1}{\sqrt{m}}, \tag{21}$$

on l'esperança és respecte de la mesura empírica, i g és una funció de test Lipschitz. Dit altrament, una xarxa soma $\phi(\boldsymbol{\theta}, \cdot)$, inicialitzada amb m partícules independents idènticament distribuïdes $\theta_j \sim \mu_0$, s'aproxima a la funció de *camp mitjà* [mean field] $\phi(\mu_0, x) = \mathbb{E}_{\theta\sim\mu_0}\chi(\theta, x)$, amb unes fluctuacions típiques de l'ordre $1/\sqrt{m}$, uniformement en x sobre compactes. Això no és més que l'estimador clàssic de Monte Carlo. Veurem, seguidament, com aquesta perspectiva de mostreig també permet incorporar l'aspecte dinàmic d'entrenament.

La densitat permet parametritzar funcions $\mathcal{F}$ de forma lineal, com podem comprovar a (20). En conseqüència, l'alternativa euleriana de (15) és ara

$$\min_{\mu\in\mathcal{P}(\Omega)} \mathcal{E}[\mu] := E(\phi(\mu, \cdot)), \tag{22}$$

on $\mathcal{P}(\Omega)$ denota l'espai de mesures de probabilitat sobre Ω. Gràcies a la convexitat original de E i a la linealitat de (20), aquest problema és convex respecte de μ, contràriament a (15).

Malgrat aquesta bona notícia, l'optimització del problema (22) no és trivial, ja que la geometria de $\mathcal{P}(\Omega)$ que dona lloc a aquesta convexitat (la geometria de

mixtures, en què el punt mitjà entre dues mesures μ i ν és la mixtura $\frac{1}{2}(\mu+\nu)$) no és compatible amb el mètode de gradient descendent que es pot implementar a la pràctica. Concretament, l'algorisme de gradient descendent, actuant sobre els paràmetres de cada partícula, defineix un operador proximal anàleg al de (17), però amb una mètrica de *transport*:

$$\mu_{k+1} = \arg\min_{\mu}\{\mathcal{E}[\mu_k] + \delta\mathcal{E}[\mu_k]\cdot(\mu-\mu_k) + \eta^{-1}W_2(\mu,\mu_k)\}, \tag{23}$$

on $W_2(\mu,\mu')$ és la distància de 2-Wasserstein entre dues mesures a $\mathcal{P}(\Omega)$, i $\delta\mathcal{E}$ és la primera variació de $\mathcal{E}$ respecte a μ. Similarment al cas euclidià, aquest pas proximal admet un límit en temps continu, però en aquest cas, en lloc d'una equació diferencial ordinària, obtenim una equació en derivades parcials que és, en terminologia física, l'equació de continuïtat de massa en un transport de mesura ([126, 168, 44, 180]):

$$\partial_t\mu_t = \operatorname{div}(\nabla\delta\mathcal{E}(\mu_t)\cdot\mu_t). \tag{24}$$

Aquest formalisme permet obtenir garanties d'optimització en xarxes somes arbitràriament amples [44, 167], malgrat el fet que el funcional no és *convex per desplaçaments*, juntament amb cotes de generalització sense malefici dimensional [10]. Tanmateix, aquestes garanties d'optimització són vàlides en el límit termodinàmic, també dit de *camp mitjà* [mean field], on l'evolució de la mesura comença a μ_0 i no en la seva versió empírica $\mu^{(m)}$. De la mateixa manera que tenim una llei dels grans nombres (i també un teorema de límit central), a la inicialització (vegeu equació (21)) les trajectòries que comencen a μ_0 i $\mu^{(m)}$, diguem-ne μ_t i $\mu_t^{(m)}$, es poden acotar, sota determinades condicions, de manera que conservin la taxa de convergència feble en $1/\sqrt{m}$ uniformement en temps [43].

En resum, la visió euleriana de les xarxes somes permet identificar una estructura matemàtica robusta, a més d'estudiar el problema d'optimització amb eines de transport de la mesura i mecànica estadística. Malgrat que les garanties d'aprenentatge són encara qualitatives (el nombre de neurones necessàries per garantir optimalitat en el cas general segueix sent exponencial en la dimensió ambient), les eines permeten ser optimistes sobre una teoria matemàtica de xarxes somes. Les extensions a xarxes més profundes s'han començat a estudiar [154, 60], així com en problemes amb simetries [223], o problemes d'optimització competitiva en jocs de suma zero [54].

5 Generalització

Un AA té una bona capacitat de generalització si la hipòtesi $h \in \mathcal{F}$ que escull difereix poc de l'expert que ha produït els exemples, és a dir, si el defecte $L(h)$ és petit. Com es pot garantir aquesta condició si l'AA només coneix els exemples $\mathcal{D}$ i, d'una manera o altra, l'espai $\mathcal{F}$? El problema pot semblar impossible si tenim en compte que $\mathcal{D}$ sempre és finit, que el malefici dimensional sempre sotja totes les cantonades i que $\mathcal{F}$, per regla general, és infinit.

Si té solució, quina forma podem esperar que tingui un enunciat sobre la capacitat de generalització? Atès que es tracta de trobar una fita de $L(h)$ en funció del que coneix l'AA, podem esperar una expressió d'aquesta fita en què intervingui el risc empíric de h, $\hat{L}_{\mathcal{D}}(h)$, la mida de $\mathcal{D}$, n, i alguna mesura de la cabuda o riquesa de $\mathcal{F}$, que en abstracte podem denotar $\mathsf{Cab}(\mathcal{F})$. A més, l'enunciat ha de ser necessàriament probabilista, la qual cosa vol dir que la fita serà vàlida amb una certa probabilitat, que, com és costum, expressarem com una confiança $\geqslant 1-\delta$ per a un valor δ petit prefixat. En resum, per aquesta via, que és la usual en aprenentatge estadístic, tindrem expressions de la forma $L(h) \leqslant_\delta \hat{L}_{\mathcal{D}}(h) + \mathsf{Fun}(\mathsf{Cab}(\mathcal{F}), n, \delta)$, on $\leqslant_\delta$ vol dir amb probabilitat almenys $1-\delta$ relativament a les mostres $\mathcal{D}$ i on Fun és una certa funció dels arguments $\mathsf{Cab}(\mathcal{F})$, n i δ. El resultat que segueix és un exemple per il·lustrar aquest procediment.

TEOREMA 4 ([176, COROL·LARI 4.6]). *Suposem que $\mathcal{F}$ és un conjunt finit d'hipòtesis binàries. Llavors, per a qualsevol $\delta > 0$,*

$$L(h) \leqslant_\delta \hat{L}_{\mathcal{D}}(h) + \sqrt{\frac{\ln|\mathcal{F}| + \ln\frac{2}{\delta}}{2n}}.$$

Veiem que podem interpretar que $\mathsf{Cab}(\mathcal{F}) = |\mathcal{F}|$. El resultat ens diu que la generalització augmenta si la minimització del risc empíric ens assegura que $\hat{L}_{\mathcal{D}}(h)$ és petit i que n és prou gran perquè el segon sumand de l'expressió també sigui petit.

L'alternativa és expressar la relació $L(h) \leqslant_\delta \mathsf{Fun}$ com una fita inferior de n (*complexitat de la mostra*). Per exemple, el teorema anterior es pot expressar dient, amb les mateixes hipòtesis, que $|L(h) - \hat{L}_{\mathcal{D}}(h)| \leqslant_\delta \varepsilon$ si

$$n \geqslant \frac{1}{2\varepsilon^2}\left(\ln|\mathcal{F}| + \ln\frac{2}{\delta}\right). \tag{25}$$

Cal remarcar que (25) i, per tant, el teorema 4, es poden millorar en el *cas realitzable*, és a dir, quan $f^* \in \mathcal{F}$. En efecte, en aquestes condicions es pot establir una fita lineal en ε^{-1} ([176, corol·lari 2.3]):

$$n \geqslant \frac{1}{\varepsilon}\left(\ln|\mathcal{F}| + \ln\frac{1}{\delta}\right). \tag{26}$$

Tanmateix, els espais d'hipòtesis són habitualment infinits i l'estudi de la generalització que es pot aconseguir és més sofisticat, ja que cal introduir nocions apropiades de $\mathsf{Cab}(\mathcal{F})$ i establir fites superiors de $L(h)$ o inferiors de n que assegurin la generalització amb confiança $1-\delta$. En aquesta secció introduïm diverses nocions de *cabuda* d'un espai d'hipòtesis $\mathcal{F}$. La de Vapnik-Chervonenkis (VC), introduïda com un *índex* a [203, §1], val per a hipòtesis binàries i no depèn de la densitat de probabilitat P sobre $\mathcal{X}$, un fet que s'expressa dient que és una noció *agnòstica*. N'hi ha bones exposicions en molts textos, com ara [4], [108] o [176], i li dediquem l'apartat 5.1. El segon apartat, 5.2, el destinem a

l'anomenada *dimensió de Pollard* (Pol), introduïda a [158], que és similar a la VC, però sense la restricció a valors binaris, i el tercer apartat, 5.3, el dediquem a l'anomenada *dimensió de Rademacher* (RAD). Aquest concepte (exposat en detall a [176, capítol 26]; vegeu també les notes [12]) no és agnòstic i això fa, com veurem, que usualment permeti assolir fitacions millors de les quantitats que ens interessen. Finalment, a l'apartat 5.4, anotem diverses comparacions entre les tres nocions de cabuda que acabem d'esmentar.

5.1 La dimensió VC

Mirem primer el cas d'un espai $\mathcal{F}$ d'hipòtesis binàries $h\colon \mathcal{X} \to \{0,1\}$, que podem identificar a subconjunts de $\mathcal{X}$: $h \leftrightarrow h^{-1}(1) = 1_{h(x)=1}$. Diem que $\mathcal{F}$ *disgrega* [shatters] un conjunt finit $Z \subset \mathcal{X}$ si tota funció binària de Z és la restricció d'una $h \in \mathcal{F}$. La *dimensió* (o *cabuda*, o *capacitat*) de Vapnik-Chervonenkis de $\mathcal{F}$, denotada $\mathrm{VC}(\mathcal{F})$, es defineix com el màxim cardinal d'un subconjunt finit $Z \subset \mathcal{X}$ disgregat per $\mathcal{F}$, si aquest màxim existeix, o ∞ altrament. Veiem, doncs, que VC és un concepte purament combinatori. En termes del càlcul pràctic, en el cas de cabuda k finita normalment es pot exhibir un Z de cardinal k disgregat per $\mathcal{F}$ i provar tot seguit que cap subconjunt de cardinal $k+1$ ho pot ser. En el cas de capacitat infinita, s'ha de mostrar que per a tot $k \in \mathbb{N}$ hi ha un Z de cardinal k disgregat per $\mathcal{F}$.

EXEMPLES. (1) Sigui $\mathcal{X} = \mathbb{R}$ i $\mathcal{F}$ el conjunt de semirectes positives $[a, \infty)$, $a \in \mathbb{R}$. Llavors $\mathrm{VC}(\mathcal{F}) = 1$. En efecte, donat $Z = \{z\}$, $z \in \mathcal{X}$, hi ha semirectes que el contenen i d'altres que no, de manera que Z és disgregat per $\mathcal{F}$. Per contra, si $Z = \{z, z'\} \subset \mathbb{R}$, $z < z'$, una semirecta que contingui z també conté z' i, per tant, Z no pot ser disgregat per $\mathcal{F}$.

(2) Encara a $\mathbb{R}$, considerem el conjunt $\mathcal{F}$ d'intervals tancats. Llavors $\mathrm{VC}(\mathcal{F}) = 2$, ja que si $Z = \{z, z'\}$ com en l'exemple anterior, hi ha intervals tancats disjunts de Z, d'altres que el contenen, i encara d'altres que contenen z i no z', i viceversa. En canvi, si $Z = \{z, z', z''\}$, $z < z' < z''$, no hi ha cap interval tancat que contingui z i z'' i no contingui z', de manera que cap Z de cardinal 3 pot ser disgregat per $\mathcal{F}$. Adonem-nos que en l'exemple (1) tindríem també capacitat 2 si a més de les semirectes positives consideréssim alhora les negatives.

(3) Sigui $\mathcal{F}$ l'espai dels semiplans del pla $\mathcal{X} = \mathbb{R}^2$. Un conjunt Z de tres punts no alineats de $\mathcal{X}$ és disgregat per $\mathcal{F}$, ja que per a tot subconjunt Z' de Z hi ha semiplans h tals que $h \cap Z = Z'$. Per altra banda, cap conjunt de quatre punts distints de $\mathcal{X}$ pot ser disgregat per $\mathcal{F}$. En efecte, podem suposar que no n'hi ha tres d'alineats, ja que un semiplà que conté els dos extrems d'un segment conté qualsevol altre punt d'aquest segment. També podem suposar que cap dels punts és interior al triangle format pels altres tres, ja que llavors aquest punt automàticament pertany a qualsevol semiplà que contingui el triangle. I si els quatre punts formen un quadrilàter convex, no hi ha cap semiplà que contingui els extrems d'una diagonal i exclogui els extrems de l'altra diagonal. En resum, cap conjunt de quatre punts pot ser disgregat per $\mathcal{F}$ i, així, $\mathrm{VC}(\mathcal{F}) = 3$.

(4) L'exemple anterior és vàlid en qualsevol dimensió $d \geqslant 2$: la capacitat del conjunt $\mathcal{F}$ de semiespais de $\mathbb{R}^d$ és $d+1$. La part més senzilla és trobar un conjunt Z de $d+1$ punts disgregats per $\mathcal{F}$. Sigui $Z = \{z_0, z_1, \ldots, z_d\}$, on z_0 és l'origen i z_j el punt unitat sobre el j-èsim eix de coordenades. Sigui $\{0, 1, \ldots, d\} = A \sqcup B$ una partició arbitrària i posem $w_j = 1$ si $j \in A$ i $w_j = -1$ si $j \in B$. Llavors el semiespai definit per l'hiperplà d'equació $h(x) = w_0/2 + w_1x_1 + \cdots + w_sx_d = 0$ conté (exclou) els punts z_j per a $j \in A$ ($j \in B$), ja que $h(z_0) = w_0/2$ i $h(z_j) = w_0/2 + w_j$ per a $j > 0$, d'on es desprèn que $h(z_j)$ té el mateix signe que w_j per a tot j. La part que queda (que cap conjunt de punts Z de cardinal $d+2$ pot ser disgregat per $\mathcal{F}$) és una conseqüència del fet que existeix (lema de Radon) una partició $Z = Z' \sqcup Z''$ tal que $[Z'] \cap [Z''] \neq \emptyset$, on posem $[Z']$ i $[Z'']$ per denotar les envolupants convexes de Z' i Z'': si hi hagués un semiespai que separés Z' i Z'', aquest semiespai també separaria $[Z']$ i $[Z'']$.

(5) La capacitat de la família de polígons convexos de r o menys costats és $2r+1$. En efecte, sigui Z un conjunt de $2r+1$ punts sobre una circumferència i femne una partició $Z = Z' \sqcup Z''$. Si $|Z'| \leqslant r$, el polígon convex amb vèrtexs Z' està contingut a la circumferència i, per tant, exclou els punts de Z''. Si en canvi $|Z''| \leqslant r$, llavors podem construir, a partir de les tangents a la circumferència en els punts de Z'', un polígon convex de $|Z''|$ vèrtexs que inclou Z' i exclou Z'': basta desplaçar les tangents esmentades un infinitèsim envers el centre. Aquests arguments es poden modificar fàcilment per establir que els polígons convexos d'exactament r costats també tenen capacitat $2r+1$. Finalment, és clar que els polígons convexos, sense condicions sobre el nombre de costats, tenen capacitat ∞.

El paper de VC en el problema de la generalització queda palès en el teorema següent.

TEOREMA 5. *Considerem $k = \mathrm{VC}(\mathcal{F})$ i suposem $k < \infty$. Llavors tenim:*

(1) $L(h) \leqslant_\delta \hat{L}_{\mathcal{D}}(h) + O\left(\sqrt{\dfrac{k + \ln(1/\delta)}{n}}\right)$.

Alternativament, $n = O\left(\frac{1}{\varepsilon^2}(k + \ln(1/\delta))\right)$ assegura $|L(h) - \hat{L}_{\mathcal{D}}(h)| \leqslant_\delta \varepsilon$.

(2) *Si $h \in \mathcal{F}$ satisfà $L_{\mathcal{D}}(h) = 0$, llavors $n \geqslant \frac{8}{\varepsilon}(k\ln(16/\varepsilon) + \ln(2/\delta))$ garanteix $L(h) \leqslant_\delta \varepsilon$.* Remarca: l'expressió de la fita inferior de n és $O\left(\frac{k}{\varepsilon}\ln(1/\varepsilon) + \frac{1}{\varepsilon}\ln(1/\delta)\right)$.

En termes del cost, $L(h) \leqslant_\delta O\left(\frac{1}{n}(k\ln(n/k) + \ln(1/\delta))\right)$ si $\hat{L}_{\mathcal{D}}(h) = 0$.

Per acabar, anotem què passa quan VC és ∞ (cf. [176, teorema 6.6].

TEOREMA 6. *Per a una classe $\mathcal{F}$ d'hipòtesis binàries tal que $\mathrm{VC}(\mathcal{F}) = \infty$ existeixen oracles que cap algorisme pot aprendre.*

5.2 La dimensió de Pollard

Suposem que $\mathcal{Y} = [0,K] \subset \mathbb{R}$, $K > 0$, de manera que $\mathcal{F}$ és un conjunt de funcions $h\colon \mathcal{X} \to [0,K]$. Sigui $X = \{x_1,\dots,x_n\}$ un subconjunt de $\mathcal{X}$. Diem que $\mathcal{F}$ *disgrega* X *amb testimonis* $t_1,\dots,t_n \in \mathbb{R}$ si per a tot subconjunt X' de X existeix una funció $h \in \mathcal{F}$ tal que $h(x_j) \leqslant t_j$ o $h(x_j) > t_j$ segons que $x_j \in X'$ o $x_j \notin X'$.

La *dimensió de Pollard* de $\mathcal{F}$, que denotarem $\mathsf{Pol}(\mathcal{F})$, és el màxim cardinal d'un subconjunt X de $\mathcal{X}$ que $\mathcal{F}$ pot disgregar. En el cas binari, $\mathcal{Y} = \{0,1\}$, clarament tenim $\mathsf{Pol}(\mathcal{F}) = \mathrm{VC}(\mathcal{F})$. Un altre exemple és la igualtat $\mathsf{Pol}(\mathcal{F}) = \dim(\mathcal{F})$ si $\mathcal{F}$ és un espai vectorial de funcions reals de dimensió finita.

Amb les notacions anteriors, posem $p = \mathsf{Pol}(\mathcal{F})$ i sigui $\mathcal{D} \sim P^n$, és a dir, un subconjunt de n elements de $\mathcal{X}$ extrets d'acord amb la distribució P de manera independent. Llavors tenim:

TEOREMA 7 (POLLARD, 1984). *Per a tota* $h \in \mathcal{F}$,

$$|\hat{L}_{\mathcal{D}}(h) - \mathbb{E}_{x\sim P}[h(x)]| \leqslant_\delta K\sqrt{\frac{2p}{n}\ln\frac{en}{p}} + K\sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}.$$

Alternativament, $|\hat{L}_{\mathcal{D}}(h) - \mathbb{E}_{x\sim P}[h(x)]| \leqslant_\delta \varepsilon$ *per a* $n \geqslant \frac{8K^2}{\varepsilon^2}\left(p\ln\frac{8K^2}{\varepsilon^2} + \frac{1}{4}\ln\frac{1}{\delta}\right)$.

5.3 La complexitat de Rademacher

Aquesta noció es pot definir per a una família $\widetilde{\mathcal{F}}$ de funcions $\tilde{h}\colon \mathcal{Z} \to [a,b] \subset \mathbb{R}$, en què $(\mathcal{Z},P)$ és un espai de probabilitat, i el seu propòsit és calibrar la capacitat de $\widetilde{\mathcal{F}}$ d'encaixar un cert soroll aleatori. En l'aplicació al model bàsic, tindrem $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$, amb $P(z) = P(x,y) = P(y|x)P(x) = P_x(y)P(x)$, i $\widetilde{\mathcal{F}}$ serà la família de funcions $\tilde{h}\colon \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$, $h \in \mathcal{F}$, definides per la relació (el defecte puntual ℓ s'ha introduït a l'apartat «Defecte», pàgina 17):

$$\tilde{h}(x,y) = \ell(h(x),y). \tag{27}$$

En realitat hem de considerar dues nocions de complexitat de Rademacher: $\mathrm{RAD}_{\boldsymbol{z}}(\widetilde{\mathcal{F}})$, on $\boldsymbol{z} = \{z_1,\dots,z_n\}$ és una mostra donada ($z_j \sim P$), i $\mathrm{RAD}_n(\widetilde{\mathcal{F}})$. La segona és la *complexitat de Rademacher* (per a mostres de longitud n) i la primera, la *complexitat de Rademacher empírica* (relativa a una mostra $\boldsymbol{z}$). Per definir-les, ens cal introduir les *variables de Rademacher* $\boldsymbol{\sigma} = \sigma_1,\dots,\sigma_n$ a fi de representar un soroll binari aleatori. Són variables aleatòries independents, una per a cada dada de la mostra, a valors $\{-1,1\}$ equiprobables.

La correlació d'una mostra d'aquest soroll amb els valors $\tilde{h}(\boldsymbol{z}) = \tilde{h}(z_1),\dots,\tilde{h}(z_n)$ l'expressa el producte escalar $\boldsymbol{\sigma}\cdot\tilde{h}(\boldsymbol{z})$. Per tant, $\sup_{\tilde{h}\in\widetilde{\mathcal{F}}}\boldsymbol{\sigma}\cdot\tilde{h}(\boldsymbol{z})/n$ expressa la millor correlació que es pot obtenir, en mitjana sobre la mostra, amb funcions de $\widetilde{\mathcal{F}}$. Amb tot això, ja podem definir les dues complexitats de Rademacher:

$$\mathrm{RAD}_{\boldsymbol{z}}(\widetilde{\mathcal{F}}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\tilde{h}\in\widetilde{\mathcal{F}}}\frac{\boldsymbol{\sigma}\cdot\tilde{h}(\boldsymbol{z})}{n}\right] \quad \text{i} \quad \mathrm{RAD}_n(\widetilde{\mathcal{F}}) = \mathbb{E}_{\boldsymbol{z}\sim P^n}[\mathrm{RAD}_{\boldsymbol{z}}(\widetilde{\mathcal{F}})]. \tag{28}$$

TEOREMA 8. *Per a tot $\delta > 0$ i tota $\tilde{h} \in \widetilde{\mathcal{F}}$, es compleixen les desigualtats*

$$\mathbb{E}_{z\sim P}[\tilde{h}(z)] \leqslant_\delta \hat{L}_{\boldsymbol{z}}(\tilde{h}) + 2\mathrm{RAD}_{\boldsymbol{z}}(\widetilde{\mathcal{F}}) + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}, \tag{29}$$

$$\mathbb{E}_{z\sim P}[\tilde{h}(z)] \leqslant_\delta \hat{L}_{\boldsymbol{z}}(\tilde{h}) + 2\mathrm{RAD}_n(\widetilde{\mathcal{F}}) + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}. \tag{30}$$

Si particularitzem els resultats anteriors al cas del model bàsic, via la introducció de l'espai $\widetilde{\mathcal{F}}$ associat a $\mathcal{F}$ explicada al principi d'aquest apartat, obtenim:

TEOREMA 9 (FITACIÓ DE RADEMACHER). *Sigui $\mathcal{F}$ un espai d'hipòtesis i $\mathcal{D}$ un conjunt de dades empíriques. Llavors per a tota $h \in \mathcal{F}$ es compleix la desigualtat*

$$L(h) \leqslant_\delta \hat{L}_{\mathcal{D}}(h) + 2\mathrm{RAD}_n(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{2n}}. \tag{31}$$

Si abstraiem el paper de $\tilde{h}(\boldsymbol{z}) \in \mathbb{R}^n$ mirant-lo només com un vector $\boldsymbol{a} \in \mathbb{R}^n$, podem definir $\mathrm{RAD}(A) = \mathbb{E}_{\boldsymbol{\sigma}}[\sup_{\boldsymbol{a}\in A} \frac{\boldsymbol{\sigma}.\boldsymbol{a}}{n}]$ per a tot subconjunt A de $\mathbb{R}^n$, una noció que facilita l'establiment de propietats de RAD.

TEOREMA 10 (FITA DE RAD). *Sigui $A \subset \mathbb{R}^n$ finit i $L = \sup_{\boldsymbol{a}\in A} \|\boldsymbol{a} - \bar{\boldsymbol{a}}\|$, on $\bar{\boldsymbol{a}}$ és el baricentre de A. Llavors*

$$\mathrm{RAD}(A) \leqslant \frac{L\sqrt{2\ln|A|}}{n}. \tag{32}$$

Si $\mathcal{F}$ és una classe d'hipòtesis binàries tal que $\mathrm{VC}(\mathcal{F}) = k$, i $\mathcal{D} \sim P^n$, llavors

$$\mathrm{RAD}_{\mathcal{D}}(\mathcal{F}) \leqslant \sqrt{\frac{2k\ln n}{n}}. \tag{33}$$

Escau comentar aquí que trobar bones cotes de la complexitat de Rademacher en xarxes neuronals és actualment un front actiu de recerca.

5.4 Ramificacions

Fem primer una ullada a fitacions per altres vies; després, a fitacions no uniformes, i, finalment, a una qüestió que considerem especialment important: el fenomen anomenat de *doble descens* que es manifesta en l'entrenament

de xarxes neuronals profundes i que refuta la comprensió convencional del capmàs entre l'error d'entrenament i l'error de generalització (figura 1).

Fitacions per altres vies Es poden obtenir cotes de generalització per altres vies, com ara emprant nocions d'estabilitat algorísmica [27]. Aquestes nocions es basen en l'*anàlisi de sensibilitat*, que tracta de quantificar la repercussió en la sortida d'un sistema d'una variació en l'entrada amb el propòsit de poder dissenyar sistemes resilients enfront de sorolls pertorbadors de l'entrada. Per a l'estat de l'art sobre les cotes obtingudes amb aquests plantejaments, tenim l'article recent [28], que a més obté cotes inferiors de l'error de generalització i cotes superiors substancialment millors que les conegudes.

Fitacions no uniformes Les dimensions VC (agnòstica respecte de la distribució P que regeix la generació de dades) i RAD (que depèn de P) tenen un caràcter uniforme en la bola $\mathcal{F}_\delta$, una condició que comporta certes limitacions, [134], i que correspon al ritme *estadístic* $O(1/\sqrt{n})$ de generalització.

En situacions més favorables, és possible substituir aquesta anàlisi uniforme per una anàlisi més refinada basada en la complexitat local de l'espai d'hipòtesis a l'entorn de la solució de l'ERM regularitzat, culminant en un *ritme ràpid* [fast rate] $O(1/n)$ en el cas de certs espais de nuclis gaussians, [16]. Aquesta anàlisi refinada també es pot aplicar en models de regressió esparsa [208].

La paradoxa del doble descens Un altre aspecte «pessimista» de les cotes de generalització vistes fins ara és que són agnòstiques a l'algorisme d'optimització utilitzat. La importància de considerar també aquest aspecte s'il·lustra amb la paradoxa del «doble descens». Esquematitzada a la figura 9, és un fenomen que exemplifica el contrast entre la comprensió convencional de l'anomenat *compromís entre complexitat de l'espai d'hipòtesis i la taxa d'aprenentatge* [bias-variance trade-off] [176, capítol 5] i el que en els darrers anys ha revelat l'experimentació amb xarxes neuronals profundes [134, 23, 22]. El compromís en qüestió procura equilibrar el subaprenentatge i el sobreaprenentatge escollint un model que sigui prou expressiu per poder extreure l'estructura rellevant de les dades, però també prou simple per no arreplegar-ne detalls espuris. Segons aquesta percepció, les xarxes neuronals amb moltes neurones (o *ultraparametritzades*) estarien descartades, ja que l'elevat nombre de paràmetres de què depenen permet que puguin memoritzar totes les dades d'entrenament i, a causa d'això, la seva capacitat generalitzadora hauria de ser molt baixa o nul·la. Tanmateix, la pràctica mostra que efectivament es poden ajustar amb exactitud a les dades d'entrenament i ensems tenir una capacitat de generalització que, de fet, augmenta amb la complexitat de la xarxa profunda. Per sort, l'entrellat de la paradoxa s'ha explicat fàcilment a [126] mitjançant la teoria de matrius aleatòries, un article remarcable en què també es posa de manifest que l'algorisme d'optimització té un paper fonamental de regularització *implícita* de l'ERM, cosa que també explica (per a models tractables com ara els models lineals generalitzats) la bona generalització de les xarxes neuronals ultraparametritzades.

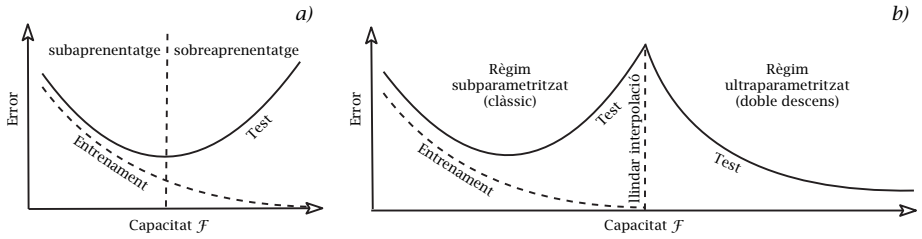

FIGURA 9: Adaptació de la figura 1 de [22], tenint en compte també els gràfics de [126]. *a*) Aquest gràfic és semblant al de la figura 1, però aquí a l'eix horitzontal hi ha la capacitat de $\mathcal{F}$. Representa una descripció que regeix quan el nombre de paràmetres de les funcions de $\mathcal{F}$ és inferior al nombre de dades disponibles en l'entrenament: en augmentar la capacitat, l'error d'entrenament disminueix monòtonament, però l'error de generalització, mesurat usant dades de test, primer disminueix i a partir d'un cert valor, que separa el subaprenentatge del sobreaprenentatge, augmenta. *b*) Representa el comportament extraordinari, i paradoxal a simple vista, que apareix quan el nombre de paràmetres supera el nombre de dades d'entrenament: l'error de generalització torna a disminuir a partir del pic que separa els dos règims, fet que possibilita un aprenentatge pràcticament perfecte de les dades d'entrenament i ensems una capacitat de generalització que augmenta amb el nombre de paràmetres.

Serveixi, com a cloenda de la secció, una citació de Donoho a l'article [55], ja esmentat, sobre les «benediccions de la dimensionalitat», que «són menys àmpliament percebudes, però inclouen el fenomen de la concentració de mesura (així anomenat en la geometria dels espais de Banach), la qual cosa significa que certes fluctuacions aleatòries estan molt ben controlades en altes dimensions, i també l'èxit dels mètodes asimptòtics, àmpliament usats en estadística matemàtica i en física estadística, les quals coses suggereixen que es poden fer afirmacions en escenaris de dimensions molt elevades que resulten massa complicades en dimensions moderades».

6 Altres models i problemes oberts

Hem anat veient que l'AA té relació amb diversos dominis, particularment de matemàtiques i estadística, però també d'algorísmia i complexitat, física matemàtica, o combinatòria. Això subratlla el seu caràcter transversal i multidisciplinari, però aquesta llista es pot estendre amb d'altres dominis que poden oferir oportunitats importants per a la recerca, sigui teòrica o aplicada. En el proper apartat, 6.1, s'indiquen breument algunes de les connexions que ens semblen més rellevants. Els dos apartats següents es dediquen a la consideració una mica més aprofundida de la relació amb el problema general d'incorporar explicabilitat als AA, apartat 6.2, i amb l'àrea de les XN algebrogeomètriques, apartat 6.3. En el darrer apartat, 6.4, s'indiquen algunes qüestions obertes o que haurien de merèixer més atenció.

6.1 Connexions

La tria d'aquest apartat fuig de la mera curiositat i en canvi es fixa en línies que aporten una comprensió més plena de l'AA i són rellevants per a indagacions futures.

Teoria de la informació Potser el tractat [119] és el que representa millor aquesta connexió, ja que la seva resposta a la pregunta de per què és beneficiós unificar la teoria de la informació i l'AA és que «són dues cares d'una mateixa moneda». I afegeix: «Els cervells són l'àpex dels sistemes de compressió i comunicació, mentre que els algorismes d'última generació per a la compressió de dades i per als codis correctors d'errors utilitzen les mateixes eines que l'aprenentatge automàtic».

Dels aspectes més concrets, destaquem les contribucions de Naftali Tishby i col·laboradors en el que anomenen «mètode/principi del coll d'ampolla de la informació» [information bottleneck method/principle]. Introduït a [193], és aplicat a l'AA profund a [194] i [177]. El treball més recent que hem trobat en aquest sentit, [157], és en col·laboració amb Z. Piran i R. Shwartz-Ziv, i el seu objecte és introduir el que anomenen un «coll d'ampolla dual».

AA amb dades generades matemàticament El treball [111] és paradigmàtic d'aquesta connexió i suggereix possibilitats insospitades fins ara. Una de les idees que hi trobem és generar aleatòriament una mostra gran d'expressions matemàtiques f, en un format apropiat, i formar la llista de parells (f', f), on f' és la derivada de f (relativament fàcil de calcular mitjançant algorismes ben coneguts). L'AA ha de predir les f a partir de les f'. Els autors expliquen el que han descobert així: «Les XN tenen fama de resoldre millor problemes estadístics o numèrics que efectuar càlculs o manipulacions simbòliques. En aquest article demostrem que els AA poden ser sorprenentment bons en tasques matemàtiques més elaborades, com ara la *integració simbòlica* i la *resolució d'equacions diferencials*. Proposem una sintaxi per representar problemes matemàtics i mètodes per generar grans conjunts de dades que es poden utilitzar per entrenar models de seqüència a seqüència. Aconseguim *resultats que superen els sistemes comercials d'àlgebra simbòlica com Matlab o Mathematica*». El més sorprenent és que l'AA no coneix cap teoria de la integració ni cap regla de les que aprenem en els primers cursos d'anàlisi.

Destaquem també uns treballs de Yang-Hui He i col·laboradors que obtenen resultats anàlegs en altres dominis: [81] (una monografia sobre qüestions plantejades per la teoria de cordes, i, especialment, de geometria algebraica i enumerativa), [82] (estructures algebraiques), [2] (teoria de nombres i la conjectura de Birch-Swinnerton-Dyer) i [83] (dirigit a esbrinar el comportament de l'AA en qüestions relatives a l'estudi de grafs finits, els autors escriuen que «les XN poden realitzar, amb alta eficiència i precisió, multitud de tasques que van des del reconeixement de la condició de planitud Ricci d'un graf, fins a la predicció de la bretxa espectral o la detecció de la presència de cicles hamiltonians»).

També encaixen perfectament en aquest punt els treballs [48] i [110]. Aquest darrer considera XN en grafs i ofereix una perspectiva de com es poden tractar amb AA.

AA en física Els físics també estan interessats en l'impacte dels AA en el seu treball. Alguns, com [214], cerquen un «físic artificial» que pugui aprendre sense supervisió. En una línia similar trobem [94].

En canvi [41] ofereix una extensa panoràmica de les connexions entre les ciències físiques i l'AA. En aquest sentit és oportú tenir en compte la iniciativa [101] del Departament d'Energia dels Estats Units de crear un versió quàntica d'Internet, la qual cosa veiem com un estímul més per seguir indagant en l'AA.

6.2 Causalitat i AA explicables

En aquesta connexió, al davant cal posar les idees de Judea Pearl, explicades detalladament en llibres com ara [145] (que en particular conté el remarcable algorisme de propagació de creences en grafs), [146] (art i ciència de causes i efectes) i [147] (els fruits madurs de la recerca de l'autor durant més de dues dècades, un tractat sobre els fonaments de la causalitat). Com a text complementari, aparegut després de la primera edició (any 2000) de [147], vegeu [181] (remarcable per la seva claredat i concisió) i [68] (un breu article d'anàlisi de l'àrea des del punt de vista de les ciències socials: «Tot es complica quan movem el focus del què al què passa si i al per què»).

Seguim encara amb J. Pearl i alguns treballs fonamentals apareguts en els darrers cinc anys: [148] (causes dels efectes i efectes de les causes), [149] (una apreciació de Trygve Haavelmo com un dels iniciadors dels càlculs causals), [199] («En contrast amb els enfocaments dels llibres de text com la maximització de l'esperança i el mètode del gradient, el nostre procediment és no iteratiu, produeix estimacions de paràmetres de forma tancada i elimina la necessitat d'inferència en una xarxa bayesiana») i [14] (sobre els problemes que comporta la fusió de dades en relació amb la causalitat). Finalment, dues peces força diferents però aclaridores: [150] (discuteix set obstacles a l'AA des del punt de vista de la causalitat) i [151] (una presentació relativament informal del seu pensament madur). Afegim el text [153] (introducció a la inferència causal i fonaments dels AA), que el gran nombre de citacions de treballs de Pearl confereix un caràcter de reconeixement a aquest investigador.

Explicabilitat D'aquest tema, estretament relacionat amb la causalitat i amb la interpretabilitat, seleccionem uns pocs treballs recents que el tracten des de diversos punts de vista i que ens semblen entrades recomanables per iniciar-se en el seu estudi: [59, 218, 8, 7, 66].

6.3 Xarxes neuronals algebrogeomètriques

Les quantitats x i w del model de neurona que hem considerat a l'apartat 2.3 són números reals. Però podem imaginar que siguin entitats d'una estructura

algebraica $\mathcal{A}$ suficient per garantir que l'expressió $x \cdot w = x_1 w_1 + \cdots + x_n w_n$, i una funció d'activació $\sigma\colon \mathcal{A} \to \mathcal{A}$, tinguin sentit. Per exemple, $\mathcal{A}$ pot ser una àlgebra real de dimensió finita, i σ, l'aplicació d'una sigmoide ordinària que actua component a component (respecte d'una base prefixada de l'àlgebra). Arribem així al concepte $\mathcal{A}$-*neurona* i, connectant neurones tal com hem fet a l'apartat esmentat, a la noció $\mathcal{A}$-*xarxa neuronal*, o $\mathcal{A}$-XN. Una altra generalització consisteix a substituir x i w per estructures de dades més generals, com ara $\mathcal{A}$-tensors [arrays], i el producte $x \cdot w$ per una operació $x \star w$ apropiada. Entre aquestes operacions, les més usades son certs productes bilineals, com ara la *convolució*, i també no lineals, com ara el max-pooling.

Així doncs, les neurones i xarxes neuronals usuals són $\mathbb{R}$-neurones i $\mathbb{R}$-xarxes neuronals. Més enllà dels números reals, entre els casos concrets més immediats d'àlgebres $\mathcal{A}$ podem esmentar $\mathbb{C}$ (números complexos), $\mathbb{H}$ (quaternions), $\mathbb{O}$ (octonions), una àlgebra de matrius $\mathbb{R}(n)$ o una àlgebra geomètrica $\mathcal{G} = \mathcal{G}_{r,s}$ de signatura (r, s). Per simplificar la terminologia, parlarem de xarxes reals, complexes, quaterniòniques (XNQ), octoniòniques, matricials i geomètriques (XNG), respectivament.

La raó d'usar àlgebres geomètriques és que el seu formalisme està òptimament adaptat a expressar els fets geomètrics de qualsevol *espai geomètric lineal*, és a dir, d'un espai vectorial real $E = E_{r,s}$ dotat d'una mètrica de signatura (r, s). La manera més directa d'introduir l'àlgebra geomètrica $\mathcal{G}_{r,s}$ d'aquest espai, i ensems més propera a les idees en què W. K. Clifford (1845–1879) es va basar per a la seva creació, és que l'àlgebra exterior $\Lambda E_{r,s}$ admet un producte bilineal **associatiu** (batejat com a *producte geomètric* pel mateix Clifford) tal que

$$xx' = x \cdot x' + x \wedge x', \quad x, x' \in E, \tag{34}$$

és a dir, que amalgama els dos productes (interior i exterior) introduïts per Grassmann (vegeu [216] per a una presentació detallada d'aquesta aproximació, que inclou també el tractament dels espais geomètrics no lineals, o [215, capítol 3] per a una presentació axiomàtica). Així podem pensar $\mathcal{G}_{r,s}$ com l'àlgebra exterior enriquida amb el producte geomètric (aquesta estructura també es coneix com a *àlgebra de Clifford*). És clar, doncs, que té dimensió 2^n, on $n = r + s = \dim E$. Adonem-nos que l'equació (34) mostra que la graduació lineal de $\mathcal{G}_{r,s}$, que de fet és una graduació respecte del producte exterior, no ho és respecte del producte geomètric. Però la descomposició $\mathcal{G} = \mathcal{G}^+ \oplus \mathcal{G}^-$ en components de graus parells ($\mathcal{G}^+$) i senars ($\mathcal{G}^-$) és una graduació mòdul 2 *també respecte del producte geomètric* (en darrera instància això es deriva de l'equació (34), en la qual el producte de dos vectors es resol en un escalar, que té grau 0, i un bivector, que té grau 2). En particular, $\mathcal{G}^+$ és una subàlgebra. Els isomorfismes $\mathcal{G}_{1,0} \simeq \mathbb{R} \oplus \mathbb{R}$, $\mathcal{G}_{0,1} \simeq \mathcal{G}^+_{2,0} \simeq \mathbb{C}$, $\mathcal{G}_{2,0} \simeq \mathbb{R}(2)$ o $\mathcal{G}_{0,2} \simeq \mathcal{G}^+_{3,0} \simeq \mathbb{H}$, fàcils d'establir directament, són de fet exemples d'un ordit general, com podeu trobar a [215]. D'aquests isomorfismes, els que connecten més estretament l'àlgebra amb la geometria són $\mathbf{C} = \mathcal{G}^+_{2,0} \simeq \mathbb{C}$ i $\mathbf{H} = \mathcal{G}^+_{3,0} \simeq \mathbb{H}$ en el cas del pla i espai euclidians, respectivament (de $\mathbf{C}$ i $\mathbf{H}$ direm que són els números

complexos i els quaternions *geomètrics*, puix que emergeixen directament de la geometria i no d'una definició *ad hoc* com la usual per introduir $\mathbb{C}$ i $\mathbb{H}$). Per a mostres de diverses aplicacions de G, vegeu [215, 112] i les seves bibliografies.

Tot seguit aportem algunes referències generals sobre diverses XN algebro-geomètriques, amb uns comentaris breus. Per a consideracions més detallades sobre aquestes i altres referències, vegeu [217].

Xarxes neuronals complexes Potser la idea més important d'aquestes xarxes és que poden explotar les propietats de la fase dels números complexos. A l'inici de l'estudi d'aquestes xarxes destaquen els treballs de Hirose i la seva escola, molt enfocats al tractament del senyal, amb reculls com [85] (2003) i tractats com [141] (2009) o [86] (2012; una segona edició d'un llibre del mateix títol i autor publicat el 2006), i la col·lecció de deu articles aplegats a [87] (2013), dels quals sobresurt el primer, pel mateix Hirose (l'editor del volum), amb el títol *Application fields and fundamental merits of complex-valued neural networks*. Al mateix cercle pertany el text [1], que il·lustra amb experiments gràfics molt convincents el valor de tenir en compte la fase.

Més recentment tenim [75] (2016), sobre xarxes complexes convolutives; [159] (2017), enfocat a la classificació d'imatges; [196] (2017), on l'èmfasi rau en les xarxes profundes, i [129], que aporta una valoració de les xarxes complexes en tasques de classificació de senyals reals. Finalment esmentem [198], en el qual és palesa la significació de les xarxes complexes des d'altres punts de vista, particularment el de l'AA profund, un «terme paraigua per a tècniques emergents que intenten generalitzar models neuronals profunds (estructurats) a dominis no euclidians com ara grafs i varietats».

Xarxes quaterniòniques L'interès d'aquestes xarxes prové de la relació que tenen els quaternions (geomètrics, si us plau) amb el grup de rotacions de l'espai euclidià ordinari, una relació especialment diàfana en termes de $\mathbf{H}$, ja que l'expressió $\underline{h}(x) = hx\bar{h}$, $h \in \mathbf{H}$ no nul, és un vector si x és un vector i $\underline{h}$ és una semblança de raó $|h|^2$ (una rotació si h és unitari). Una altra raó és que els quaternions tenen tres fases i que aquestes fases es poden usar per extreure informació valuosa dels senyals que s'han de processar.

La recerca en XNQ també ve de lluny, fins i tot abans que la de les xarxes complexes. Remetem a [29] i [88] per a informació històrica rellevant relativa al que s'anomena *anàlisi de Clifford*, sobretot en relació amb transformades de Fourier i d'ondetes en un context quaterniònic i en la seva generalització al context geomètric. En el tema més concret que ens ocupa, a l'origen hi ha Gerald Sommer i els seus col·laboradors: [38] (1998, generalització dels filtres de Gabor) i [36] (2000, generalització del perceptró multicapa). La memòria [42] presenta una teoria d'ondetes quaterniòniques «per a l'anàlisi i processament d'imatges» i [92] (2009) és una panoràmica de les propietats i aplicacions de les xarxes quaterniòniques fins aquell moment.

A la darrera dècada, la recerca en XNQ ha prosseguit tant en el front aplicat com en el teòric. L'article [93] s'ocupa del perceptró multicapa quaterniònic.

A [103] s'investiguen les XNQ de Hopfield i la seva invariància per rotacions. En els treballs [142] i [143], les XNQ s'apliquen a la comprensió del llenguatge parlat. Les XNQ profundes s'estudien a [67] i les convolutives, a [222]. Finalment [219] presenta una versió quaterniònica de les xarxes de càpsules adreçada a processar núvols de punts de l'espai ordinari, i a [130] s'introdueix una XNQ que ofereix invariància per contrast i sensibilitat als angles de rotació.

Xarxes geomètriques Curiosament, l'estudi de les XNG va començar fins i tot abans que el de les XNQ, com per exemple a [152], i G. Sommer en fou un impulsor a principis del mil·lenni amb treballs com ara [182], en què desenvolupa els fonaments teòrics que li serveixen per a problemàtiques com la visió artificial i la robòtica, [37], dedicat a una G-versió del perceptró de capes múltiples, [61] i [39], que desenvolupen la noció de *senyal monogènic*. Una culminació d'aquests esforços és la tesi de Sven Buchholz, [35], que s'ha de considerar, com indica el seu títol, una teoria de la computació neuronal amb àlgebres geomètriques. Com a mostra d'aplicacions, citem [162] (segmentació d'imatges), [20] (vectors de suport en el context geomètric), els volums [18] (computació geomètrica per a transformades d'ondetes, visió artificial, aprenentatge, control i acció) i [21] (computació geomètrica en enginyeria i informàtica), [63] (ús de l'àlgebra geomètrica per a la detecció de vores en imatges en color), [99] (tècniques d'optimització en estimació geomètrica), [155] (mètodes d'agrupació basats en l'àlgebra geomètrica conforme $\mathcal{G}_{4,1}$) i [209] (tractament d'imatges multiespectrals amb àlgebra geomètrica). Acabem amb [19], el primer volum del que ha de ser un tractat sistemàtic d'aquests desenvolupaments.

Altres xarxes algebrogeomètriques A l'article [213] (2020) es construeixen xarxes octoniòniques convolutives i s'apliquen a classificació d'imatges CIFAR-10 i CIFAR-100. Segons els autors, tenen millor convergència i precisió que altres xarxes aplicades a les mateixes tasques. Els octonions s'han aplicat també amb èxit a l'aprenentatge per diccionari [dictionary learning], com ara a [114] (2018), una aproximació que de fet es pot formular per a àlgebres més generals, que inclouen les geomètriques, com s'exposa a [115].

Examinem ara [89] (2020). En el resum es declara que:

> ... els nostres resultats demostren que hi ha àlgebres alternatives que proporcionen millors paràmetres i eficiència computacional en comparació amb $\mathbb{R}$. Considerem $\mathbb{C}$, $\mathbb{H}$, $\mathbb{R}(2)$, $\mathbb{C}(2)$, $\mathbb{R}(3)$, $\mathbb{R}(4)$, $\mathbb{R} \oplus \mathbb{R}$ (números dobles) i $\mathbb{R}^3$ amb el producte vectorial. A més, observem que la multiplicació d'aquestes àlgebres té una densitat de càlcul més elevada que la multiplicació real, una propietat útil en situacions amb una reutilització de paràmetres intrínsecament limitada, com ara inferència autoregressiva i xarxes neuronals esparses. Per tant, investiguem com es pot induir l'esparsitat a les xarxes considerades. Esperem que els nostres resultats pràctics en proves a gran escala fomentin una exploració posterior d'aquestes arquitectures no convencionals que interpel·len l'elecció per defecte d'utilitzar números reals per a pesos i activacions de xarxes neuronals.

Atès que potser no és el lloc per fer-ne una apreciació crítica, ens limitem a subratllar que ja hem comentat la natura geomètrica de $\mathbb{C}$, $\mathbb{H}$ i $\mathbb{R}(2)$. Pel que fa als altres casos, $\mathbb{R} \oplus \mathbb{R} \simeq \mathcal{G}_{1,0}$, $\mathbb{C}(2) \simeq \mathcal{G}_{1,3}$ (àlgebra geomètrica de l'espai-temps) i $\mathbb{R}(4) \simeq \mathcal{G}_{2,2}$. L'àlgebra $(\mathbb{R}^3, \times)$ és una àlgebra de Lie, però de fet la seva natura és geomètrica, ja que el producte vectorial de dos vectors és el dual de Hodge (dins l'àlgebra $\mathcal{G}_{3,0}$) del seu producte exterior. Per a totes aquestes qüestions podeu consultar [215]. Com a contrapunt, creiem que el valor de les xarxes geomètriques queda reforçat, tant per raons teòriques com pels aspectes computacionals.

6.4 Qüestions obertes

La teoria de l'AA és un camp de recerca activa, on conflueixen aspectes matemàtics particularment diversos, des de la teoria de la informació, per refinar nocions de complexitat adaptades a les xarxes ultraparametritzades; la física estadística, per analitzar límits computacionals i estadístics en gran dimensió, o l'anàlisi harmònica i l'optimització, elements crucials per descriure —i prescriure— algorismes en xarxes profundes.

En conseqüència, actualment hi ha un gran nombre de qüestions obertes en tots aquests fronts. Considerem-ne algunes de les més remarcables.

Models funcionals profunds Com s'ha esmentat a l'apartat 3.2, les xarxes neuronals profundes suposen un repte important en termes d'aproximació i d'optimització. Els mètodes de camp mitjà donen una resposta satisfactòria en el cas de les xarxes somes (apartat 4.3), i actualment un repte important és poder estendre'ls a models profunds, amb resultats preliminars que podeu trobar a [60].

Anàlisi quantitativa de xarxes somes Les xarxes somes, tot i ser la classe de xarxes neuronals més simple, són objecte de força qüestions obertes. En particular, resultats recents ([71, 53]) estudien cotes inferiors d'aprenentatge, és a dir, donen resultats negatius sobre la impossibilitat d'aprendre certes classes de funcions parametritzades per xarxes somes amb un nombre polinòmic de dades. En contrapartida, la teoria del camp mitjà pinta una situació més optimista (i més alineada amb el comportament empíric), però de moment en termes asimptòtics. Un problema d'importància màxima és, doncs, comprendre com es poden lligar aquestes dues perspectives amb una versió quantitativa dels resultats positius d'aprenentatge.

Variacions del mètode del gradient descendent Els resultats empírics de les xarxes profundes depenen críticament del mètode d'optimització. Tot i que el mètode del gradient descendent és la versió més bàsica i que ofereix una visió matemàtica més nítida [184, 57, 44, 56, 30, 71], és necessari incorporar variacions importants, com ara la versió estocàstica [169, 98, 46, 189, 97], o les tècniques de renormalització [batch/layer normalization] [125, 168].

Més enllà de l'aprenentatge supervisat El panorama de qüestions obertes és engrescador. Per esmentar un camp en particular, les tècniques d'autoaprenentatge [self-supervised learning] estan començant a donar fruits comparables, o fins i tot superiors, a les versions amb supervisió estàndard [74]. La teoria d'aprenentatge estadístic està tot just començant a considerar règims d'aprenentatge mes enllà del supervisat; per exemple, l'aprenentatge *per transferència* [transfer learning] [77], o l'esmentat autoaprenentatge [117, 195, 210] o la perspectiva de causalitat [6].

A Probabilitat

En aquest apèndix recollim alguns conceptes i resultats de teoria de probabilitats emprats en diversos indrets de les seccions precedents.

A.1 Desigualtat de Hoeffding

Considerem $X_j \in [a,b]$, $j = 1,\dots,m$, variables aleatòries independents amb la mateixa mitjana μ, i definim $\mu_m = \frac{1}{m}\sum_j X_j$. Llavors

$$P(|\mu - \mu_m| > \varepsilon) \leqslant 2e^{-2m\varepsilon^2/(b-a)^2}.$$

Com a exemple, podem considerar el cas $X_i \in [0,2\pi]$ amb distribució uniforme, de manera que $\mu = \pi$, $b-a = 2\pi$, i l'exponent queda igual a $-\frac{1}{2\pi^2}m\varepsilon^2$, és a dir, $P(|\pi - \mu_m| > \varepsilon) \leqslant 2e^{-\frac{1}{2\pi^2}m\varepsilon^2}$.

A.2 Probabilitat recíproca

Suposem que una variable aleatòria X satisfà $P(X > \varepsilon) \leqslant f(\varepsilon)$ per a tot $\varepsilon > 0$, on $f(\varepsilon) > 0$ és una funció donada. Si $\delta' > 0$ és del domini de f, i $f(\delta') = \delta$, llavors $P(X \leqslant \delta') = 1 - P(X > \delta') \geqslant 1 - f(\delta') = 1 - \delta$. És a dir, assegurem la relació $X \leqslant \delta'$ amb probabilitat almenys $1-\delta$.

La determinació de δ' com a funció de δ és una qüestió que cal tractar en cada cas concret. Per exemple, en la desigualtat de Hoeffding per a les $X_i \in [0,2\pi]$ uniformes, $\delta' = \pi\sqrt{\frac{2}{m}\ln\frac{2}{\delta}}$. Així doncs, $|\pi - \mu_m| \leqslant \pi\sqrt{\frac{2}{m}\ln\frac{2}{\delta}}$ amb probabilitat d'almenys $1-\delta$. Si considerem $Y_j = \cos X_j \in [-1,1]$, que tenen mitjana 0, el que trobem és $|\frac{1}{m}\sum_j Y_j| \leqslant \sqrt{\frac{2}{m}\ln\frac{2}{\delta}}$ amb probabilitat d'almenys $1-\delta$.

A.3 Desigualtat de McDiarmid

Sigui $f\colon \mathcal{Z}^n \to \mathbb{R}$ una funció per a la qual existeix una constant $c > 0$ tal que

$$|f(z_1,\dots,z_j,\dots,z_n) - f(z_1,\dots,z'_j,\dots,z_n| \leqslant c \tag{35}$$

per a tot $j = 1, \ldots, n$ i qualssevol $z_1, \ldots, z_n, z_j' \in \mathcal{Z}$. Si $\boldsymbol{z} = z_1, \ldots, z_n \in \mathcal{Z}^n$ és una seqüència de variables aleatòries independents, llavors es compleixen les desigualtats

$$P(f(\boldsymbol{z}) - \mathbb{E}[f(\boldsymbol{z}]) \geqslant \varepsilon) \leqslant \exp\left(\frac{-2\varepsilon^2}{mc^2}\right), \tag{36}$$

$$P(f(\boldsymbol{z}) - \mathbb{E}[f(\boldsymbol{z})]) \leqslant -\varepsilon) \leqslant \exp\left(\frac{-2\varepsilon^2}{mc^2}\right). \tag{37}$$

Utilitzant ara la probabilitat recíproca, podem assegurar que

$$|f(\boldsymbol{z}) - \mathbb{E}[f(\boldsymbol{z})]| \leqslant c\sqrt{\frac{m}{2}\ln\frac{2}{\delta}} \tag{38}$$

amb probabilitat no inferior a $1 - \delta$.

Aquesta desigualtat, dita de «diferències fitades», és un ingredient principal en la demostració de les desigualtats de Rademacher.

A.4 Lema de Gibbs i la divergència KL

Suposem que $p = p_1, \ldots, p_n$ i $q = q_1, \ldots, q_n$ són distribucions de probabilitat. Llavors

$$-\sum_j p_j \log_2(p_j) \leqslant -\sum_j p_j \log_2(q_j),$$

amb igualtat si, i només si, $q = p$.

D'això resulta que $\sum_j p_j \log_2(p_j/q_j) \geqslant 0$, amb igualtat només si $q = p$. Aquesta expressió se sol conèixer com a *divergència de Kullback-Leibler* de p i q, i es denota $\mathrm{KL}(p, q)$. Convé remarcar que en general $\mathrm{KL}(q, p) \neq \mathrm{KL}(p, q)$.

La divergència KL és una eina important en la teoria de les xarxes bayesianes per comparar les distribucions de probabilitat en temps consecutius. Per altra banda, vist que $H(p) = -\sum_j p_j \log_2(p_j)$ és l'*entropia* de la distribució p, és a dir, la informació mitjana proporcionada en una prova de p, és natural que la divergència KL sigui també significativa en la teoria de la informació.

B Notes bibliogràfiques

La bibliografia sobre AA, per no dir sobre IA, és vastíssima i creix diàriament. És, doncs, impossible donar-ne compte en un article panoràmic com aquest. Per tant, ens limitem a consignar la bibliografia que ens sembla més rellevant per a cada secció, sovint amb algun comentari o citació, i ensems a suggerir diversos materials que poden servir per aprofundir en els temes tractats o en d'altres de natura complementària.

B.0 Notes a la Introducció

Tractats d'AA Generals: [80, 176, 72, 161, 3]. Aproximació bayesiana: [192, 144, 172]. Aplicacions: [173, 11]. Aspectes computacionals: [76, 135, 139].

Articles clàssics [123] (primer model matemàtic de les neurones biològiques), [165] (introducció del perceptró), [211] (ADALINE, neurona lineal adaptativa), [166] (neurodinàmica), [171] (algorisme de retropropagació), [127] (tractat sobre el perceptró, introducció a la geometria computacional, història de les idees d'aprenentatge i XN fins a 1988), [65] i [116] (reconeixement de patrons visuals amb XN), [102] (recull d'articles sobre la percepció com a inferència bayesiana), [133] (enciclopèdia de ciències cognitives).

Les moltes cares de la IA És manifest que la IA, via el seu desenvolupament accelerat en tots els sectors, ja afecta tota la humanitat, i la direcció i el ritme de la seva evolució expansiva no sembla fàcil d'endevinar. Tanmateix, com s'indica a [66], «la societat ignora en gran manera les capacitats, els requisits i les pràctiques estàndard de la IA» i ensems «pren consciència dels perills que comporta la ignorància i, amb raó, demana solucions». Les reflexions, la bibliografia i les propostes d'aquest article són valuoses per a tothom concernit per aquestes incerteses i per la incidència que tenen en l'esfera ètica.

Entre els documents que analitzen les múltiples facetes de l'IA, actual i futura, ens semblen molt rellevants l'article [52] i l'informe [188]. Escrits d'aquesta mena, i com el citat en el paràgraf anterior, són imprescindibles per formar-se una idea de conjunt, amb molt poc aparell tècnic, de la IA tal com s'entén actualment.

B.1 Notes a la secció 1 (Preludis)

El tractat [132] ofereix una presentació sistemàtica a l'AA des d'una perspectiva probabilista. Per a una breu apreciació general sobre aquesta perspectiva, vegeu l'article [69].

El text [124] explica l'evolució històrica de la fórmula de Bayes-Laplace des de l'origen fins a la publicació del llibre. Documenta moltes aplicacions, com ara el fet que fou Alan Turing, en els seus treballs per a desxifrar l'Enigma, qui va promoure l'ús sistemàtic de la teoria de Bayes-Laplace en la forma en què s'ha conegut posteriorment. Un bon complement és [186], escrit per un dels protagonistes de la cerca de la bomba atòmica perduda l'any 1966 en el mar de la costa de Palomares a causa d'un accident de la força aèria americana i també del submarí nuclear *Scorpion* perdut per la marina americana a l'Atlàntic el maig de 1968.

Amb el subtítol «Why so many predictions fail-but some don't», l'autor de [179], potser l'harúspex més prestigiós en la predicció dels resultats electorals, ens diu que «El teorema de Bayes [...] implica que *hem de pensar d'una altra manera sobre les nostres idees —i com posar-les a prova*. Ens hem d'acostumar més estretament amb la *probabilitat* i la *incertesa*. Hem de pensar amb més cura sobre les suposicions i creences amb què confrontem un problema» (pàgina 15).

El lema del creador de la teoria de patrons moderna, Ulf Grenander (1923–2016), era *usar la teoria de patrons per crear estructures matemàtiques derivades tant del món natural com del món creat pels humans*. En aquesta teoria, els mètodes bayesians hi tenen un paper fonamental, com queda reflectit a [131]. En tot cas, la teoria de patrons és molt més general que el problema del seu reconeixement, fins ara el més estudiat, com ara a [25].

Anàlisi de components principals i descomposició en vectors singulars: [80, capítol 8], [113, capítol 7], [189], [208, capítol 8]. Aquests mètodes estan a l'origen de les tècniques de «reducció dimensional», un aspecte important de l'AA, [9]. Un tema relacionat és l'anomenat *anàlisi discriminant*, que podeu trobar en la majoria de tractats generals, com ara [109, capítol 8].

Finalment, [164] és una bona referència per a k-NN i k-mitjana, que comprèn també el tractament computacional. Coincidim amb l'autor en la noció que l'aprenentatge no supervisat és una àrea de recerca molt activa a la qual encara cal dedicar esforços per obtenir resultats generals satisfactoris.

B.2 Notes a la secció 2 (Aprenentatge inductiu en gran dimensió)

Sobre el malefici dimensional, Donoho en va donar una descripció precisa a [55]: «la intractabilitat aparent de la cerca sistemàtica, o de l'aproximació acurada d'una funció, o de la seva integració, en un espai d'alta dimensió».

De l'aprenentatge inductiu, també anomenat *aprenentatge estadístic*, [201] conté una exposició de l'estructura d'aquesta teoria per part d'un dels iniciadors de la seva forma moderna. Per a un resum, vegeu [202]. Esmentem també [108], un text relativament elemental, els extensos tractats [78] i [183] i la monografia [200].

En la presentació de l'aprenentatge inductiu hem recorregut a la noció d'expert, com una funció $f^*\colon \mathcal{X} \to \mathcal{Y}$. En la pràctica, un expert també està sotmès a un cert grau d'incertesa. El tractament d'aquesta situació es pot fer substituint f^* per una distribució de probabilitat $P(x, y)$ sobre $\mathcal{X} \times \mathcal{Y}$, que podem mirar, via la relació $P(x, y) = P(x)P(y|x)$, com una distribució de probabilitat $P_x(y) = P(y|x)$ sobre $\mathcal{Y}$ per a cada $x \in \mathcal{X}$. L'adaptació del model bàsic a aquesta situació més general no ofereix dificultat i es pot trobar exposada en molts textos, com ara [78, 212, 200].

Una altra idea que ha resultat molt productiva és la de produir bons predictors a partir de predictors dèbils (és a dir, que encerten només una mica millor que un predictor aleatori). La generació de predictors dèbils és relativament fàcil, i és un fet remarcable que hi hagi un algorisme senzill i eficient per construir un predictor fort a partir dels febles. Es coneixen com a *algorismes de potenciació* [boosting] i el més popular, AdaBoost (potenciació adaptativa), fou introduït a [64], un treball en el qual s'estableix una fita a l'error empíric de l'algorisme, es fa una anàlisi en termes de la dimensió VC i s'indiquen aplicacions a problemes tant de classificació múltiple (com ara reconeixement de cares) com de regressió. Podeu trobar-ne una exposició excel·lent a [176, capítol 10].

Pel que fa a les neurones i les xarxes neuronals, hem adoptat l'esquema de [204]. N'hi ha versions més detallades en molts textos, com ara [5, 80, 140]. Esmentem també l'article panoràmic [174].

Finalment, [49] compta com el primer article en què es va establir la capacitat de les xarxes neuronals per aproximar qualsevol funció contínua uniformement sobre compactes. Vegeu també [91] i [90]. L'article [220] ofereix una apreciació actualitzada d'aquesta problemàtica.

B.3 Notes a la secció 3 (Aproximació)

Espais de Sobolev Una funció $f\colon \Omega \to \mathbb{R}$ és de la classe de Sobolev $\mathcal{H}^{s,p}(\Omega)$ si f i les seves derivades fins a ordre s són integrables a $L^p(\Omega)$; vegeu [170] per a més detalls.

Notacions asimptòtiques Donades funcions $f, g\colon \mathbb{R} \to \mathbb{R}_+$, diem que $g(t)$ és $O(f(t))$ si existeix una constant $c > 0$ tal que $g(t) \leqslant cf(t)$ per a $t \gg 0$. Similarment, $g(t)$ és $\Omega(f(t))$ si existeix una constant $c' > 0$ tal que $g(t) \geqslant c'f(t)$ per a $t \gg 0$. Finalment, es diu que $g(t)$ és $\Theta(f(t))$ si és $O(f(t))$ i $\Omega(f(t))$.

B.4 Notes a la secció 4 (Optimització)

Convex per desplaçaments La generalització de la convexitat euclidiana als fluxos de gradient de Wasserstein [205, capítol 23].

Distància W_2 de Wasserstein Donades dues mesures de probabilitat μ i μ' en un espai mètric M, $W_2(\mu,\mu')^2 = \inf_{\gamma\in\Gamma(\mu,\mu')} \int d(x,x')^2\, d\gamma(x,x')$, on $\Gamma(\mu,\mu')$ és el conjunt de densitats de probabilitat sobre $M \times M$ amb densitats marginals μ i μ'. Vegeu [205, capítol 6] i, en particular, la discussió a les notes bibliogràfiques sobre la terminologia en què conclou que la més adient seria *mètrica L^p mínima*, o *mètric de Kantorovich* si hagués de portar un nom. Tanmateix, Villani també usa la terminologia usual.

Funcions β-regulars Una funció f de classe C^1 és *β-regular* si $||\nabla f(x) - \nabla f(y)|| \leqslant \beta||x - y||$, és a dir, si ∇f és β-Lipschitz.

Mínim local aproximat Si f és C^2 i $\nabla^2 f$ és ρ-Lipschitz, un mínim local aproximat a menys de ε és un punt x tal que $\|\nabla f(x)\| \leqslant \varepsilon$ i $\nabla^2 f(x) + \sqrt{\rho\varepsilon}I \succ 0$.

B.5 Notes a la secció 5 (Generalització)

A [158, pàgina 18], l'autor troba una mica inapropiada la imatge «d'engrunar en bocins petits» que suscita la noció de disgregar [shatter] una classe de funcions $\mathcal{F}$ un conjunt S, i que seria millor una imatge que evoqués la facultat de $\mathcal{F}$ de seleccionar qualsevol subconjunt de S, però l'accepta perquè «almenys és vívida». Per a més precisions, vegeu [185], una referència que hem d'agrair a un dels revisors.

Per a la capacitat VC de l'espai de funcions que calcula una XN, vegeu [17].

La fita (32) de $\textsc{Rad}(A)$ es coneix com a *lema de Massart*.

El teorema 8 es pot demostrar usant la desigualtat de McDiarmid (vegeu l'apartat A.3). Per q detalls i més informació, vegeu [176, capítol 26], [105, 107, 106].

B.6 Notes a la secció 6 (Altres models i problemes oberts)

Inferència bayesiana Consignem alguns dels textos cabdals apareguts durant aquest segle: [136] (un llibre de text sobre xarxes bayesianes); [104] (models gràfics probabilistes); [51] (modelització i raonament amb xarxes bayesianes); [50] (ús de la factorització de matrius per aprendre models de Markov; una bona mostra d'aplicació de la descomposició en valors singulars); [13] (raonament bayesià i AA); [128] (models gràfics per processar dades incompletes); [84] (llenços de Markov en el cervell, és a dir, fronteres estadístiques que regeixen les interaccions entre els esdeveniments a una banda i a l'altra del llenç).

Aprenentatge amb reforç i algunes aplicacions [190] (un manual extens sobre l'aprenentatge amb reforç i sobre diversos temes relacionats); [178] (un article breu sobre l'AA en modalitat d'autoreforç per a jocs com els escacs o el go); [32] (i per al pòquer); [175] (un text recent sobre l'estat de l'art en aprenentatge amb reforç); [47] (aprenentatge amb reforç en el camp de la robòtica); [40] (aprenentatge amb reforç aplicat al problema del plegament de proteïnes).

Teoria de patrons Es pot considerar que el pare de la teoria de patrons actual és Ulf Grenander (1923–2016). La relació d'aquesta teoria amb l'AA és profunda i sembla clar que no s'ha explorat en aquesta direcció tot el seu potencial. Certament, es poden citar les *Lectures in Pattern Theory* de Grenender (Springer LNiM 18, 24, 33), però la versió que ens sembla més adient com a inspiració per a l'AA és l'aproximació de David Mumford al volum [131] i a la monografia [221].

AA i intel·ligència En línies generals, hi ha dos grans ramals en construcció enfocats a l'estudi de l'aprenentatge i la intel·ligència des del punt de vista algorísmic. Un és el representat per l'AA, amb totes les seves variants, del qual hem intentat fer una presentació panoràmica en aquest treball. El propòsit de l'altre ramal, en canvi, és esbrinar l'estructura i el funcionament dels cervells del regne animal, principalment els dels mamífers i, molt especialment, els dels humans. Aquests dos ramals estan encara lluny de convergir, però tot sembla indicar que ho faran en algun moment futur. En acabar aquest article, s'ha publicat [79], un llibre remarcable tant pel seu estil divulgatiu com per les referències als treballs de recerca essencials en què es basen els seus arguments. El recomanem vivament a tothom que tingui interès per l'evolució de la interacció entre els dos ramals, una cruïlla interdisciplinària en la qual creiem que tindran un paper destacat moltes de les tècniques que hem descrit fins aquí, potser amb modificacions i potser amb d'altres no inventades encara.

Agraïments

A Joaquim Bruna i Floris, pel seu paper catalitzador en l'establiment d'aquesta col·laboració dels autors. A Manuel Udina, per les observacions precises que ens va enviar, i als revisors i editors pels suggeriments i correccions recollits en els seus informes detallats. Escrit en bona mesura en temps de covid-19, aquest article no hauria estat possible sense les facilitats de producció en paral·lel i a distància d'Overleaf, o d'edició local com TeXstudio, o de comunicació com Skype o Zoom.

Referències

[1] Aizenberg, I. *Complex-Valued Neural Networks with Multi-Valued Neurons*. Berlín: Springer-Verlag, 2011. (Studies in Computational Intelligence; 353)

[2] Alessandretti, L.; Baronchelli, A.; He, Y.-H. «Machine learning meets number theory: The data science of Birch-Swinnerton-Dyer». Preprint (2019). arXiv:1911.02008.

[3] Alpaydin, E. *Introduction to Machine Learning*. 4a ed. Cambridge, MA: MIT Press, 2020. (Adaptive Computation and Machine Learning)

[4] Anthony, M.; Bartlett, P. L. *Neural Network Learning: Theoretical Foundations*. Cambridge: Cambridge University Press, 1999.

[5] Arbib, M. A. (ed.). *The Handbook of Brain Theory and Neural Networks*. 2a ed. Cambridge, MA: MIT Press, 2003. (A Bradford Book)

[6] Arjovsky, M.; Bottou, L.; Gulrajani, I.; Lopez-Paz, D. «Invariant risk minimization». Preprint (2019). arXiv:1907.02893.

[7] Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J. [*et al.*]. «Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI». *Information Fusion*, 58 (2020), 82–115.

[8] Arya, V.; Bellamy, R. K. E.; Chen, P.-Y. [*et al.*]. «One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques». Preprint (2019). arXiv:1909.03012.

[9] Ayesha, S.; Hanif, M.; Talib, R. «Overview and comparative study of dimensionality reduction techniques for high dimensional data». *Information Fusion*, 59 (2020), 44–58.

[10] Bach, F. «Breaking the curse of dimensionality with convex neutral networks». *J. Mach. Learn. Res.*, 18 (2017), article núm. 19, 53 p.

[11] Balas, V. E.; Roy, S. S.; Sharma, D.; Samui, P. (ed.). *Handbook of Deep Learning Applications*. Cham, Suïssa: Springer Nature Switzerland AG, 2019. (Smart Innovation, Systems and Technologies; 136)

[12] Balcan, M.-F. «Rademacher complexity». Curs *CS 8803 - Machine learning theory*. Georgia Tech (2011). Lecture Notes 11/17. https://www.cs.cmu.edu/~ninamf/ML11/lect1117.pdf.

[13] BARBER, D. *Bayesian Reasoning and Machine Learning*. Cambridge: Cambridge University Press, 2012.

[14] BAREINBOIM, E.; PEARL, J. «Causal inference and the data-fusion problem». *Proc. Nat. Acad. Sci. U.S.A.*, 113 (27) (2016), 7345–7352.

[15] BARRON, A. R. «Approximation and estimation bounds for artificial neural networks». *Mach. Learn.*, 14 (1) (1994), 115–133.

[16] BARTLETT, P. L.; BOUSQUET, O.; MENDELSON, S. «Local Rademacher complexities». *Ann. Statist.*, 33 (4) (2005), 1497–1537.

[17] BARTLETT, P. L.; MAASS, W. «Vapnik-Chervonenkis dimension of neural nets». A: ARBIB, M. (ed.). *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: The MIT Press, 2003, 1188–1192.

[18] BAYRO-CORROCHANO, E. *Geometric Computing: for Wavelet Transforms, Robot Vision, Learning, Control and Action*. Londres; Nova York: Springer, 2010.

[19] BAYRO-CORROCHANO, E. *Geometric Algebra Applications Vol. I: Computer Vision, Graphics and Neurocomputing*. Cham, Suïssa: Springer International Publishing, 2019.

[20] BAYRO-CORROCHANO, E. J.; ARANA-DANIEL, N. «Clifford support vector machines for classification, regression, and recurrence». *IEEE Trans. Neural Networks*, 21 (11) (2010), 1731–1746.

[21] BAYRO-CORROCHANO, E.; SCHEUERMANN, G. (ED.). *Geometric Algebra Computing: in Engineering and Computer Science*. Londres: Springer-Verlag, 2010.

[22] BELKIN, M.; HSU, D.; MA, S.; MANDAL, S. «Reconciling modern machine-learning practice and the classical bias-variance trade-off». *Proc. Natl. Acad. Sci. USA*, 116 (32) (2019), 15849–15854.

[23] BELKIN, M.; MA, S.; MANDAL, S. «To understand deep learning we need to understand kernel learning». Preprint (2018). arXiv:1802.01396.

[24] BIETTI, A.; MAIRAL, J. «Group invariance, stability to deformations, and complexity of deep convolutional representations». *J. Mach. Learn. Res.*, 20 (2019), article núm. 25, 49 p.

[25] BISHOP, C. M. *Pattern Recognition and Machine Learning*. Nova York: Springer, 2006. (Information Science and Statistics)

[26] BOTTOU, L.; BOUSQUET, O. «The tradeoffs of large scale learning». A: *Advances in Neural Information Processing Systems 20: Proceedings of the 2007 Conference*. Neural Information Processing Systems Foundation, Inc. (NIPS), 2008, 161–168.

[27] BOUSQUET, O.; ELISSEEFF, A. «Stability and generalization». *J. Mach. Learn. Res.*, 2 (3) (2002), 499–526.

[28] BOUSQUET, O.; KLOCHKOV, Y.; ZHIVOTOVSKIY, N. «Sharper bounds for uniformly stable algorithms». A: *Proceedings of Machine Learning Research*, 125, 2020, 610–626.

[29] BRACKX, F.; HITZER, E.; SANGWINE, S. J. «History of quaternion and Clifford-Fourier transforms and wavelets». A: *Quaternion and Clifford Fourier Transforms and Wavelets*. Basilea: Birkhäuser: Springer Basel AG, 2013, xi–xxvii. (Trends Math.)

[30] BRAVERMAN, M.; HAZAN, E.; SIMCHOWITZ, M.; WOODWORTH, B. «The gradient complexity of linear regression». A: *Proceedings of Machine Learning Research*, 125, 2020, 627–647.

[31] BRONSTEIN, M. M.; BRUNA, J.; LECUN, Y.; SZLAM, A.; VANDERGHEYNST, P. «Geometric deep learning: going beyond Euclidean data». *IEEE Signal Processing Magazine*, 34 (4) (2017), 18–42.

[32] BROWN, N.; SANDHOLM, T. «Superhuman AI for multiplayer poker». *Science*, 365 (6456) (2019), 885–890.

[33] BRUNA, J.; MALLAT, S. «Invariant scattering convolution networks». *IEEE Trans. Pattern Anal. Mach. Intell.*, 35 (8) (2013), 1872–1886.

[34] BUBECK, S. «Convex optimization: Algorithms and complexity». Preprint (2014). arXiv:1405.4980.

[35] BUCHHOLZ, S. «A theory of neural computation with Clifford algebras». Tesi doctoral. Christian-Albrechts Universität Kiel, 2005.

[36] BUCHHOLZ, S.; SOMMER, G. «Quaternionic spinor MLP». A: *ESANN'2000 Proceedings*. D-Facto, 2000, 377–382.

[37] BUCHHOLZ, S.; SOMMER, G. «Clifford algebra multilayer perceptrons». A: *Geometric Computing with Clifford Algebras*. Berlín: Springer, 2001, 315–334.

[38] BULOW, T.; SOMMER, G. «Quaternionic Gabor filters for local structure classification». A: *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*. Vol 1. IEEE, 1998, 808–810.

[39] BÜLOW, T.; SOMMER, G. «Hypercomplex signals—a novel extension of the analytic signal to the multidimensional case». *IEEE Trans. Signal Process.*, 49 (11) (2001), 2844–2852.

[40] CALLAWAY, E. «'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures». *Nature*, 588 (7837) (2020), 203–204.

[41] CARLEO, G.; CIRAC, I.; CRANMER, K. [*et al.*]. «Machine learning and the physical sciences». *Rev. Modern Phys.*, 91 (4) (2019), p. 045002.

[42] CHAN, W. L.; CHOI, H.; BARANIUK, R. «Quaternion wavelets for image analysis and processing». A: *IEEE International Conference on Image Processing (ICIP 2004)*. Vol. 5. IEEE, 2004, 3057–3060.

[43] CHEN, A.; ROTSKOFF, G. M.; BRUNA, J.; VANDEN-EIJNDEN, E. «A dynamical central limit theorem for shallow neural networks». Preprint (2020). arXiv:2008.09623.

[44] CHIZAT, L.; BACH, F. «On the global convergence of gradient descent for over-parameterized models using optimal transport». Preprint (2018). arXiv:1805.09545.

[45] CHIZAT, L.; OYALLON, E.; BACH, F. «On lazy training in differentiable programming». A: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019, 2937–2947.

[46] CHO, K. «Foundations and advances in deep learning». Tesi doctoral. Aalto University School of Science, 2014.

[47] COLOMÉ, A.; TORRAS, C. *Reinforcement Learning of Bimanual Robot Skills*. Cham, Suïssa: Springer International Publishing, 2020. (Springer Tracts in Advanced Robotics; 134)

[48] CRANMER, M.; SANCHEZ-GONZALEZ, A.; BATTAGLIA, P.; XU, R.; CRANMER, K.; SPERGEL, D.; HO, S. «Discovering symbolic models from deep learning with inductive biases». Preprint (2020). arXiv:2006.11287.

[49] CYBENKO, G. «Approximation by superpositions of a sigmoidal function». *Math. Control Signals Systems*, 2 (4) (1989), 303–314.

[50] CYBENKO, G.; CRESPI, V. «Learning hidden Markov models using non-negative matrix factorization». *IEEE Trans. Inform. Theory*, 57 (6) (2011), 3963–3970.

[51] DARWICHE, A. *Modeling and Reasoning with Bayesian Networks*. Cambridge: Cambridge University Press, 2009.

[52] DARWICHE, A. «Human-level intelligence or animal-like abilities?». *Comm. ACM*, 61 (10) (2018), 56–67.

[53] DIAKONIKOLAS, I.; KANE, D. M.; ZARIFIS, N. «Near-optimal SQ lower bounds for agnostically learning halfspaces and ReLUs under Gaussian marginals». Preprint (2020). arXiv:2006.16200.

[54] DOMINGO-ENRICH, C.; JELASSI, S.; MENSCH, A.; ROTSKOFF, G.; BRUNA, J. «A mean-field analysis of two-player zero-sum games». Preprint (2020). arXiv:2002.06277.

[55] DONOHO, D. L. [*et al.*]. «High-dimensional data analysis: The curses and blessings of dimensionality». *AMS Math Challenges Lecture*, 1 (2000), 1–32.

[56] DU, S.; LEE, J.; LI, H.; WANG, L.; ZHAI, X. «Gradient descent finds global minima of deep neural networks». A: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. PMLR: 2019, 1675–1685.

[57] DU, S. S.; ZHAI, X.; POCZOS, B.; SINGH, A. «Gradient descent provably optimizes over-parameterized neural networks». A: *7th International Conference on Learning Representations*. OpenReview.net: 2019, 19 p.

[58] ELDAN, R.; SHAMIR, O. «The power of depth for feedforward neural networks». A: *Proceedings of Machine Learning Research*, 49, 2016, 907–940.

[59] ESCALANTE, H. J.; ESCALERA, S.; GUYON, I.; BARÓ, X.; GÜÇLÜTÜRK, Y.; GÜÇLÜ, U.; VAN GERVEN, M. (ED.). *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Cham, Suïssa: Springer International Publishing, 2018. (Challenges in Machine Learning)

[60] FANG, C.; LEE, J. D.; YANG, P.; ZHANG, T. «Modeling from features: a mean-field framework for over-parameterized deep neural networks». Preprint (2020). arXiv:2007.01452.

[61] FELSBERG, M.; SOMMER, G. «The monogenic signal». *IEEE Trans. Signal Process.*, 49 (2) (2001), 3136–3144.

[62] FELZENSZWALB, P. F.; GIRSHICK, R. B.; MCALLESTER, D.; RAMANAN, D. «Object detection with discriminatively trained part-based models». *IEEE Trans. Pattern Anal. Mach. Intell.*, 32 (9) (2010), 1627–1645.

[63] FRANCHINI, S.; GENTILE, A.; SORBELLO, F.; VASSALLO, G.; VITABILE, S. «Clifford algebra based edge detector for color images». A: *2012 Sixth International Conference on Complex, Intelligent, and Software Intensive Systems*. IEEE, 2012, 84–91.

[64] FREUND, Y.; SCHAPIRE, R. E. «Game theory, on-line prediction and boosting». A: *Proceedings of the Ninth Annual Conference on Computational Learning Theory*. IEEE, 1996, 325–332.

[65] FUKUSHIMA, K. «Neocognitron: A hierarchical neural network capable of visual pattern recognition». *Neural networks*, 1 (2) (1988), 119–130.

[66] GARCIA-GASULLA, D.; CORTÉS, A.; ALVAREZ-NAPAGAO, S.; CORTÉS, U. «Signs for ethical AI: A route towards transparency». Preprint (2020). arXiv:2009.13871.

[67] GAUDET, C. J.; MAIDA, A. S. «Deep quaternion networks». A: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, 1–8.

[68] GELMAN, A. «Causality and statistical learning». Preprint (2010). arXiv:1003.2619.

[69] GHAHRAMANI, Z. «Probabilistic machine learning and artificial intelligence». *Nature*, 521 (7553) (2015), 452–459.

[70] GIRSHICK, R.; DONAHUE, J.; DARRELL, T.; MALIK, J. «Rich feature hierarchies for accurate object detection and semantic segmentation». A: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, 580–587.

[71] GOEL, S.; GOLLAKOTA, A.; JIN, Z.; KARMALKAR, S.; KLIVANS, A. «Super-polynomial lower bounds for learning one-layer neural networks using gradient descent». Preprint (2020). arXiv:2006.12011.

[72] GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge, MA: MIT Press, 2016. (Adaptive Computation and Machine Learning)

[73] GREENGARD, L. *The Rapid Evaluation of Potential Fields in Particle Systems*. Cambridge, MA: MIT Press, 1988. (ACM Distinguished Dissertations)

[74] GRILL, J.-B.; STRUB, F.; ALTCHÉ, F. [*et al.*]. «Bootstrap your own latent: A new approach to self-supervised learning». Preprint (2020). arXiv:2006.07733.

[75] GUBERMAN, N. «On complex valued convolutional neural networks». Preprint (2016). arXiv:1602.09046.

[76] GUTTAG, J. *Introduction to Computation and Programming using Python: With Application to Understanding Data*. Cambridge, MA: MIT Press, 2016.

[77] HANNEKE, S.; KPOTUFE, S. «A no-free-lunch theorem for multitask learning». Preprint (2020). arXiv:2006.15785.

[78] HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. 2a ed. Nova York: Springer, 2009. (Springer Series in Statistics)

[79] HAWKINS, J. *A Thousand Brains: A New Theory of Intelligence*. With a foreword by Richard Dawkins. Nova York: Basic Books, 2021.

[80] HAYKIN, S. *Neural Networks and Learning Machines*. Pearson, 2009.

[81] HE, Y.-H. «The Calabi–Yau landscape: from geometry, to physics, to machine-learning». Preprint (2018). arXiv:1812.02893.

[82] HE, Y.-H.; KIM, M. «Learning algebraic structures: preliminary investigations». Preprint (2019). arXiv:1905.02263.

[83] HE, Y.-H.; YAU, S.-T. «Graph Laplacians, Riemannian manifolds and their machine-learning». Preprint 2020. arXiv:2006.16619.

[84] HIPOLITO, I.; RAMSTEAD, M.; CONVERTINO, L.; BHAT, A.; FRISTON, K.; PARR, T. «Markov blankets in the brain». Preprint (2018). arXiv:2006.02741.

[85] HIROSE, A. (ED.). *Complex-Valued Neural Networks. Theories and Applications*. River Edge, NJ: World Scientific Publishing Co., Inc., 2003. (Series on Innovative Intelligence; 5)

[86] HIROSE, A. *Complex-Valued Neural Networks*. 2a ed. Berlín, Heidelberg: Springer-Verlag, 2012. (Studies in Computational Intelligence; 400)

[87] HIROSE, A. (ED.). *Complex-Valued Neural Networks. Advances and Applications*. Piscataway, NJ: IEEE Press; Hoboken, NJ: John Wiley & Sons, Inc., 2013. (IEEE Press Series on Computational Intelligence)

[88] HITZER, E.; SANGWINE, S. J. (ED.). *Quaternion and Clifford Fourier Transforms and Wavelets*. Basilea: Birkhäuser: Springer Basel AG, 2013. (Trends in Mathematics)

[89] HOFFMANN, J.; SCHMITT, S.; OSINDERO, S.; SIMONYAN, K.; ELSEN, E. «AlgebraNets». Preprint (2020). arXiv:2006.07360.

[90] HORNIK, K. «Approximation capabilities of multilayer feedforward networks». *Neural networks*, 4 (2) (1991), 251–257.

[91] HORNIK, K.; STINCHCOMBE, M.; WHITE, H. «Multilayer feedforward networks are universal approximators». *Neural Networks*, 2 (5) (1989), 359–366.

[92] ISOKAWA, T.; MATSUI, N.; NISHIMURA, H. «Quaternionic neural networks: Fundamental properties and applications». A: *Complex-Valued Neu-*

ral Networks: Utilizing High-Dimensional Parameters. IGI Global, 2009, 411–439.

[93] ISOKAWA, T.; NISHIMURA, H.; MATSUI, N. «Quaternionic multilayer perceptron with local analyticity». *Information (Switzerland)*, 3 (4) (2012), 756–770.

[94] ITEN, R.; METGER, T.; WILMING, H.; DEL RIO, L.; RENNER, R. «Discovering physical concepts with neural networks». *Phys. Rev. Lett.*, 124 (1), p. 010506 (2020).

[95] JACOT, A.; GABRIEL, F.; HONGLER, C. «Neural tangent kernel: Convergence and generalization in neural networks». A: Bengio, S. [et al.]. *Advances in Neural Information Processing Systems*. Vol. 31, 2018, 8571–8580.

[96] JAIN, A. K.; ZHONG, Y.; LAKSHMANAN, S. «Object matching using deformable templates». *IEEE Trans. Pattern Anal. Mach. Intell.*, 18 (3) (1996), 267–278.

[97] JIN, C.; NETRAPALLI, P.; GE, R., KAKADE, S. M.; JORDAN, M. I. «On nonconvex optimization for machine learning: gradients, stochasticity, and saddle points». Preprint (2019). arXiv:1902.04811.

[98] JOHNSON, R.; ZHANG, T. «Accelerating stochastic gradient descent using predictive variance reduction». A: *Advances in Neural Information Processing Systems*. Vol. 26. Curran, 2014, 315–323.

[99] KANATANI, K. «Overviews of optimization techniques for geometric estimation». *Mem. Fac. Engrg. Okayama Univ.*, 47 (2013), 1–18.

[100] KINGMA, D. P.; BA, J. «Adam: A method for stochastic optimization». Preprint (2014). arXiv:1412.6980.

[101] KLEESE VAN DAM, K.; MONDA, I.; PETERS, N.; SCHENKEL, T. «From long-distance entanglement to building a nationwide quantum internet: Report of the DOE Quantum Internet Blueprint Workshop». US, febrer 2020. [Disponible en línia a: http://doi.org/10.2172/1638794]

[102] KNILL, D. C.; RICHARDS, W. (ED.). *Perception as Bayesian Inference*. Cambridge University Press, 1996. El primer article és «Pattern theory: a unifying perspective», de D. Mumford.

[103] KOBAYASHI, M. «Rotational invariance of quaternionic Hopfield neural networks». *IEEJ Transactions on Electrical and Electronic Engineering*, 11 (4) (2016), 516–520.

[104] KOLLER, D.; FRIEDMAN, N. *Probabilistic Graphical Models. Principles and Techniques*. Cambridge, MA: MIT Press, 2009. (Adaptive Computation and Machine Learning)

[105] KOLTCHINSKII, V. «Rademacher penalties and structural risk minimization». *IEEE Trans. Inform. Theory*, 47 (5) (2001), 1902–1914.

[106] KOLTCHINSKII, V. «Rademacher complexities and bounding the excess risk in active learning». *J. Mach. Learn. Res.*, 11 (2010), 2457–2485.

[107] KOLTCHINSKII, V.; PANCHENKO, D. «Rademacher processes and bounding the risk of function learning». A: *High Dimensional Probability, II* (Seattle, WA, 1999). Boston, MA: Birkhäuser Boston, 2000, 443–457. (Progr. Probab.; 47)

[108] KULKARNI, S.; HARMAN, G. *An Elementary Introduction to Statistical Learning Theory*. Hoboken, NJ: John Wiley & Sons, Inc., 2011. (Wiley Series in Probability and Statistics)

[109] KUNG, S. Y. *Kernel Methods and Machine Learning*. Cambridge: Cambridge University Press, 2014.

[110] LAMB, L. C.; GARCEZ, A.; GORI, M.; PRATES, M.; AVELAR, P.; VARDI, M. «Graph neural networks meet neural-symbolic computing: A survey and perspective». Preprint (2020). arXiv:2003.00330.

[111] LAMPLE, G.; CHARTON, F. «Deep learning for symbolic mathematics». Preprint (2019). arXiv:1912.01412.

[112] LAVOR, C.; XAMBÓ-DESCAMPS, S.; ZAPLANA, I. *A Geometric Algebra Invitation to Space-Time Physics, Robotics and Molecular Geometry*. Cham, Suïssa: Springer, 2018. (SpringerBriefs in Mathematics)

[113] LAY, D. C.; LAY, S. R.; MCDONALD, J. J. *Linear Algebra and its Applications*. 5a ed. Pearson, 2016.

[114] LAZENDIĆ, S.; DE BIE, H.; PIŽURICA, A. «Octonion sparse representation for color and multispectral image processing». A: *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, 608–612.

[115] LAZENDIC, S.; PIZURICA, A.; DE BIE, H. «Hypercomplex algebras for dictionary learning». A: *Early Proceedings of the AGACSE 2018 Conference*. Campinas, São Paulo, Brasil: Unicamp: IMECC, 2018, 57–64.

[116] LECUN, Y.; BOSER, B.; DENKER, J. S.; HENDERSON, D.; HOWARD, R. E.; HUBBARD, W.; JACKEL, L. D. «Backpropagation applied to handwritten zip code recognition». *Neural Comput.*, 1 (4) (1989), 541–551.

[117] LEE, J. D.; LEI, Q.; SAUNSHI, N.; ZHUO, J. «Predicting what you already know helps: Provable self-supervised learning». Preprint (2020). arXiv:2008.01064.

[118] LINIAL, N.; MANSOUR, Y.; NISAN, N. «Constant depth circuits, Fourier transform, and learnability». *J. Assoc. Comput. Mach.*, 40 (3) (1993), 607–620.

[119] MACKAY, D. J. C. *Information Theory, Inference and Learning Algorithms*. Version 7.2 (fourth printing). Cambridge: Cambridge University Press, 2005.

[120] MALLAT, S. *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic Press, Inc., 1998.

[121] MALLAT, S. «Group invariant scattering». *Comm. Pure Appl. Math.*, 65 (10) (2012), 1331–1398.

[122] MALLAT, S. «Understanding deep convolutional networks». *Philos. Trans. Roy. Soc. A*, 374 (2065) (2016), p. 20150203.

[123] McCulloch, W. S.; Pitts, W. «A logical calculus of the ideas immanent in nervous activity». *Bull. Math. Biophys.*, 5 (1943), 115–133.

[124] McGrayne, S. B. *The Theory That Would Not Die. How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, & Emerged Triumphant from Two Centuries of Controversy*. New Haven, CT: Yale University Press, 2011.

[125] Mehta, P.; Schwab, D. J. «An exact mapping between the variational renormalization group and deep learning». Preprint (2014). arXiv:1410.3831.

[126] Mei, S.; Montanari, A. «The generalization error of random features regression: Precise asymptotics and double descent curve». Preprint (2019). arXiv:1908.05355.

[127] Minsky, M.; Papert, S. A. *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press, 1988. [Expanded edition of the 1969 "Perceptrons"]

[128] Mohan, K.; Pearl, J. «Graphical models for processing missing data». Preprint (2018). arXiv:1801.03583.

[129] Mönning, N.; Manandhar, S. «Evaluation of complex-valued neural networks on real-valued classification tasks». Preprint (2018). arXiv:1811.12351.

[130] Moya-Sánchez, E. U.; Xambó-Descamps, S.; Pérez, A. S.; Salazar-Colores, S.; Martínez-Ortega, J.; Cortés, U. «A bio-inspired quaternion local phase CNN layer with contrast invariance and linear sensitivity to rotation angles». *Pattern Recognition Lett.*, 131 (2021), 56–62.

[131] Mumford, D.; Desolneux, A. *Pattern Theory. The Stochastic Analysis of Real-World Signals*. Natick, MA: A K Peters, Ltd., 2010. (Applying Mathematics)

[132] Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press, 2012.

[133] Nadel, L. (ed.). *Encyclopedia of Cognitive Science*. Londres: Nature Publishing Group, 2003.

[134] Nagarajan, V.; Kolter, J. Z. «Deterministic PAC-Bayesian generalization bounds for deep networks via generalizing noise-resilience». Preprint (2019). arXiv:1905.13344.

[135] Nandy, A.; Biswas, M. *Reinforcement Learning: With Open AI, TensorFlow and Keras Using Python*. Berkeley, CA: Apress, 2017.

[136] Neapolitan, R. E. *Learning Bayesian Networks*. Upper Saddle River, NJ: Pearson Prentice Hall, 2004. (Artificial Intelligence; 38)

[137] Nemirovsky, A. S.; Yudin, D. B. *Problem Complexity and Method Efficiency in Optimization*. Nova York: John Wiley & Sons, Inc., 1983. (Wiley-Interscience Series in Discrete Mathematics)

[138] NESTEROV, YU. E. «A method for solving the convex programming problem with convergence rate $O(1/k^2)$». *Dokl. Akad. Nauk SSSR*, 269 (3) (1983), 543–547. [En rus]

[139] NGUYEN, G.; DLUGOLINSKY, S.; BOBÁK, M.; TRAN, V.; GARCÍA, A. L.; HEREDIA, I.; MALÍK, P.; HLUCHÝ, L. «Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey». *Artificial Intelligence Review*, 52 (1) (2019), 77–124.

[140] NIELSEN, M. A. *Neural Networks and Deep Learning*. San Francisco, CA, EUA: Determination Press, 2015.

[141] NITTA, T. *Complex-Valued Neural Networks: Utilizing High-Dimensional Parameters*. Hershey, PA: Information Science Reference, 2009.

[142] PARCOLLET, T.; MORCHID, M.; BOUSQUET, P.-M.; DUFOUR, R.; LINARÈS, G.; DE MORI, R. «Quaternion neural networks for spoken language understanding». A: *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, 362–368.

[143] PARCOLLET, T.; ZHANG, Y.; MORCHID, M.; TRABELSI, C.; LINARÈS, G.; DE MORI, R.; BENGIO, Y. «Quaternion convolutional neural networks for end-to-end automatic speech recognition». Preprint (2018). arXiv:1806.07789.

[144] PATEL, A. B.; NGUYEN, T.; BARANIUK, R. G. «A probabilistic theory of deep learning». Preprint (2015). arXiv:1504.00641.

[145] PEARL, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann, 1988. (The Morgan Kaufmann Series in Representation and Reasoning)

[146] PEARL, J. «The art and science of cause and effect». Conferència pública impartida el novembre de 1996 dins del *UCLA Faculty Research Lectureship Program*, 1996.

[147] PEARL, J. *Causality. Models, Reasoning, and Inference*. 2a ed. Cambridge: Cambridge University Press, 2009.

[148] PEARL, J. «Causes of effects and effects of causes». *Sociol. Methods Res.*, 44 (1) (2015), 149–164.

[149] PEARL, J. «Trygve Haavelmo and the emergence of causal calculus». *Econometric Theory*, 31 (1) (2015), 152–179.

[150] PEARL, J. «Theoretical impediments to machine learning with seven sparks from the causal revolution». Preprint (2018). arXiv:1801.04016.

[151] PEARL, J.; MACKENZIE, D. *The Book of Why. The New Science of Cause and Effect*. Nova York: Basic Books, 2018.

[152] PEARSON, J. K.; BISSET, D. L. «Neural networks in the Clifford domain». A: *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*. Vol. 3. IEEE, 1994,1465–1469.

[153] PETERS, J.; JANZING, D.; SCHÖLKOPF, B. *Elements of Causal Inference. Foundations and Learning Algorithms*. Cambridge, MA: MIT Press, 2017. (Adaptive Computation and Machine Learning)

[154] PHAM, H. T.; NGUYEN, P.-M. «A note on the global convergence of multilayer neural networks in the mean field regime». Preprint 2020. arXiv:2006.09355.

[155] PHAM, M. T.; TACHIBANA, K. «A conformal geometric algebra based clustering method and its applications». *Adv. Appl. Clifford Algebr.*, 26 (3) (2016), 1013–1032.

[156] PINKUS, A. «Density in approximation theory». *Surv. Approx. Theory*, 1 (2005), 1–45.

[157] PIRAN, Z.; SHWARTZ-ZIV, R.; TISHBY, N. «The Dual Information Bottleneck». Preprint (2020). arXiv:2006.04641.

[158] POLLARD, D. *Convergence of Stochastic Processes*. Nova York: Springer-Verlag, 1984. (Springer Series in Statistics)

[159] POPA, C.-A. «Complex-Valued Convolutional Neural Networks for Real-Valued Image Classification». A: *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, 816–822.

[160] RAHIMI, A.; RECHT, B. «Random features for large-scale kernel machines». A: *Advances in Neural Information Processing Systems*. Vol. 20. Curran Associates, Inc., 2008, 1177–1184.

[161] REBALA, G.; RAVI, A.; CHURIWALA, S. *An Introduction to Machine Learning*. Cham, Suïssa: Springer International Publishing, 2019.

[162] RIVERA-ROVELO, J.; BAYRO-CORROCHANO, E. «Medical image segmentation using a self-organizing neural network and Clifford geometric algebra». A: *The 2006 IEEE International Joint Conference on Neural Network Proceedings*. IEEE, 2006, 3538–3545.

[163] ROBBINS, H.; MONRO, S. «A stochastic approximation method». *Ann. Math. Statist.*, 22 (3) (1951), 400–407.

[164] ROSEBROCK, A. *Deep Learning for Computer Vision with Python: ImageNet Bundle*. PyImageSearch, 2017.

[165] ROSENBLATT, F. «The perceptron: a probabilistic model for information storage and organization in the brain». *Psychol. Rev.*, 65 (6) (1958), 386.

[166] ROSENBLATT, F. *Principles of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms*. Washington, DC: Spartan Books, 1962.

[167] ROTSKOFF, G.; JELASSI, S.; BRUNA, J.; VANDEN-EIJNDEN, E. «Global convergence of neuron birth-death dynamics». Preprint (2019). arXiv:1902.01843.

[168] ROTSKOFF, G. M.; VANDEN-EIJNDEN, E. «Trainability and accuracy of neural networks: An interacting particle system approach». Preprint (2018). arXiv:1805.00915.

[169] ROUX, N. L.; SCHMIDT, M.; BACH, F. R. «A stochastic gradient method with an exponential convergence-rate for finite training sets». A: *Proceedings of the 25th International Conference on Neural Information Processing Systems*. Vol. 2, 2012, 2663–2671.

[170] RUDIN, W. *Functional Analysis*. 2a ed. Nova York: McGraw-Hill, Inc., 1991. (International Series in Pure and Applied Mathematics)

[171] RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. «Learning representations by back-propagating errors». *Nature*, 323 (6088) (1986), 533–536.

[172] RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3a ed. Pearson, 2010.

[173] SAID, A.; TORRA, V. (ED.). *Data Science in Practice*. Cham, Suïssa: Springer International Publishing, 2019. (Studies in Big Data; 46)

[174] SCHMIDHUBER, J. «Deep learning in neural networks: An overview». *Neural networks*, 61 (2015), 85–117.

[175] SEWAK, M. *Deep Reinforcement Learning. Frontiers of Artificial Intelligence*. Springer, 2019.

[176] SHALEV-SHWARTZ, S.; BEN-DAVID, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge: Cambridge University Press, 2014.

[177] SHWARTZ-ZIV, R.; TISHBY, N. «Opening the black box of deep neural networks via information». Preprint (2017). arXiv:1703.00810.

[178] SILVER, D.; HUBERT, T.; SCHRITTWIESER, J. [*et al.*]. «A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play». *Science*, 362 (6419) (2018), 1140–1144.

[179] SILVER, N. *The Signal and the Noise: Why so Many Predictions Fail–But Some Don't*. Penguin, 2012.

[180] SIRIGNANO, J.; SPILIOPOULOS, K. «Mean field analysis of deep neural networks». Preprint (2019). arXiv:1903.04440.

[181] SLOMAN, S. *Causal Models: How People Think About the World and its Alternatives*. Oxford: Oxford University Press, 2005.

[182] SOMMER, G. (ED.). *Geometric Computing with Clifford Algebras. Theoretical Foundations and Applications in Computer Vision and Robotics*. Berlín: Springer-Verlag, 2001.

[183] SPANOS, A. *Probability Theory and Statistical Inference. Econometric Modeling with Observational Data*. 2a ed. Cambridge: Cambridge University Press, 2019.

[184] SRA, S.; NOWOZIN, S.; WRIGHT, S. J. (ED.). *Optimization for Machine Learning*. Cambridge, MA: MIT Press, 2012. (Neural Information Processing Series)

[185] STEELE, J. M. «Empirical discrepancies and subadditive processes». *Ann. Probability*, 6 (1) (1978), 118–127.

[186] STONE, L. D. *Theory of Optimal Search*. Nova York; Londres: Academic Press [Harcourt Brace Jovanovich, Publishers], 1975. (Mathematics in Science and Engineering; 118)

[187] STONE, C. J. «Consistent nonparametric regression». *Ann. Statist.*, 5 (4) (1977), 595–645.

[188] STONE, P.; BROOKS, R.; BRYNJOLFSSON, E. [*et al.*]. «Artificial intelligence and life in 2030». *One Hundred Year Study on Artificial Intelligence: Report of the 2015 Study Panel*, Setembre 2016.

[189] STRANG, G. *Linear Algebra and Learning from Data*. Wellesley-Cambridge Press, 2019.

[190] SUTTON, R. S.; BARTO, A. G. *Reinforcement Learning: An Introduction*. 2a ed. Cambridge, MA: MIT Press, 2018. (Adaptive Computation and Machine Learning)

[191] SZEGEDY, C.; ZAREMBA, W.; SUTSKEVER, I.; BRUNA, J.; ERHAN, D.; GOODFELLOW, I.; FERGUS, R. «Intriguing properties of neural networks». Preprint (2014). arXiv:1312.6199.

[192] THEODORIDIS, S. *Machine Learning: A Bayesian and Optimization Perspective*. Academic Press, 2015.

[193] TISHBY, N.; PEREIRA, F. C.; BIALEK, W. «The information bottleneck method». Preprint (2020). arXiv:physics/0004057.

[194] TISHBY, N.; ZASLAVSKY, N. «Deep learning and the information bottleneck principle». Preprint (2015). arXiv:1503.02406.

[195] TOSH, C.; KRISHNAMURTHY, A.; HSU, D. «Contrastive learning, multi-view redundancy, and linear models». Preprint (2020). arXiv:2008.10150.

[196] TRABELSI, C.; BILANIUK, O.; ZHANG, Y. [*et al.*]. «Deep complex networks». Preprint (2017). arXiv:1705.09792.

[197] TSYBAKOV, A. B. «Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes». *Ann. Statist.*, 26 (6) (1998), 2420–2469.

[198] TYGERT, M.; BRUNA, J.; CHINTALA, S.; LECUN, Y.; PIANTINO, S.; SZLAM, A. «A mathematical motivation for complex-valued convolutional networks». *Neural Comput.*, 28 (5) (2016), 815–825.

[199] VAN DEN BROECK, G.; MOHAN, K.; CHOI, A.; DARWICHE, A.; PEARL, J. «Efficient algorithms for Bayesian network parameter learning from incomplete data». A: *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI)*. 2015, 161–170.

[200] VÂN LÊ, H. «Mathematical foundations of machine learning». (2020). [Disponible en línia a: https://users.math.cas.cz/hvle/MFML.pdf]

[201] VAPNIK, V. N. *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998. (Adaptive and Learning Systems for Signal Processing, Communications, and Control; 1)

[202] VAPNIK, V. N. «An overview of statistical learning theory». *IEEE Trans. Neural Networks*, 10 (5) (1999), 988–999.

[203] VAPNIK, V. N.; CHERVONENKIS, A. YA. «On the uniform convergence of relative frequencies of events to their probabilities». *Theor. Probability Appl.*, 16 (1971), 264–280.

[204] VIDAL, R.; BRUNA, J.; GIRYES, R.; SOATTO, S. «Mathematics of deep learning». Preprint (2017). arXiv:1712.04741.

[205] VILLANI, C. *Optimal Transport. Old and New*. Berlín: Springer-Verlag, 2009. (Grundlehren der Mathematischen Wissenschaften; 338)

[206] VON LUXBURG, U.; BOUSQUET, O. «Distance-based classification with Lipschitz functions». *J. Mach. Learn. Res.*, 5 (2004), 669–695.

[207] VON LUXBURG, U.; SCHÖLKOPF, B. «Statistical learning theory: Models, concepts, and results». A: *Handbook of the History of Logic*. Elsevier, 2011, 651–706. (Inductive logic; 10)

[208] WAINWRIGHT, M. J. *High-dimensional Statistics. A Non-asymptotic Viewpoint*. Cambridge: Cambridge University Press, 2019. (Cambridge Series in Statistical and Probabilistic Mathematics; 48)

[209] WANG, R.; SHI, Y.; CAO, W. «GA-SURF: A new speeded-up robust feature extraction algorithm for multispectral images based on geometric algebra». *Pattern Recognition Lett.*, 127 (2019), 11–17.

[210] WEI, C.; SHEN, K.; CHEN, Y.; MA, T. «Theoretical analysis of self-training with deep networks on unlabeled data». Preprint (2020). arXiv:2010.03622.

[211] WIDROW, B.; HOFF, M. E. «Adaptive switching circuits». Technical report Stanford Electronics Labs, Stanford University, CA (1960).

[212] WOLF, M. M. «Mathematical foundations of supervised learning (growing lecture notes)». (2018). [Disponible en línia a: https://www-m5.ma.tum.de/foswiki/pub/M5/Allgemeines/MA4801_2018S/ML_notes_main.pdf]

[213] WU, J.; XU, L.; WU, F.; KONG, Y.; SENHADJI, L.; SHU, H. «Deep octonion networks». *Neurocomputing*, 397 (2020), 179–181.

[214] WU, T.; TEGMARK, M. «Toward an AI Physicist for unsupervised learning». Preprint (2018). arXiv:1810.10525.

[215] XAMBÓ-DESCAMPS, S. *Real Spinorial Groups. A Short Mathematical Introduction*. SpringerBriefs in Mathematics. Cham: Springer, 2018. (SBMAC SpringerBriefs)

[216] XAMBÓ-DESCAMPS, S. «Calculi on geometric spaces». En premsa (2021).

[217] XAMBÓ-DESCAMPS, S.; MOYA-SÁNCHEZ, E. U. «Geometric calculi and automatic learning–An outline». A: DELSHAMS, A.; XAMBÓ-DESCAMPS, S. (ed.). *Systems, Patterns and Data Engineering with Geometric Calculi*. Springer, 2021, en premsa. (ICIAM2019 SEMA SIMAI Springer Series)

[218] YING, Z.; BOURGEOIS, D.; YOU, J.; ZITNIK, M.; LESKOVEC, J. «Gnnexplainer: Generating explanations for graph neural networks». A: Bengio, S. [et al.]. *Advances in Neural Information Processing Systems*. Vol. 32, 2019, 9240–9251.

[219] ZHAO, Y.; BIRDAL, T.; LENSSEN, J. E.; MENEGATTI, E.; GUIBAS, L.; TOMBARI, F. «Quaternion equivariant capsule networks for 3D point clouds». Preprint (2019). arXiv:1912.12098.

[220] ZHOU, D.-X. «Universality of deep convolutional neural networks». *Appl. Comput. Harmon. Anal.*, 48 (2) (2020), 787–794.

[221] Zhu, S.-C.; Mumford, D. «A Stochastic Grammar of Images». *Foundations and Trends in Computer Graphics and Vision*, 2 (4) (2006), 259–362.

[222] Zhu, X.; Xu, Y.; Xu, H.; Chen, C. «Quaternion convolutional neural networks». A: *Computer Vision - ECCV 2018*. Springer, 2018, 631–647.

[223] Zweig, A.; Bruna, J. «A functional perspective on learning symmetric functions with neural networks». Preprint (2020). arXiv:2008.06952.

Joan Bruna
Courant Institute & Center for Data Science, New York University
60 Fifth Avenue, Office 612
Nova York, USA
bruna@cims.nyu.edu

Sebastià Xambó
Departament de Matemàtiques, UPC, i Visitant BSC
Edifici OMEGA, c/ Jordi Girona 1-3
08034 Barcelona, Spain
sebastia.xambo@upc.edu