What will it be about?

- Statistical approach to neural network learning
- Specificity of the expectation-based learning
- Strong law of large numbers for network learing
- Central limit theorem for artificial neural networks
- A central limit theorem application to network pruning

Basic framework

- A MLP with n input neurons, m output neurons
- The *training pairs* $z_i = (x_i, y_i)$ with $x_i \in \mathbb{R}^n, y_i \in \mathbb{R}^m$ are viewed as realizations of *random vectors* Z_i, respectively X_i and Y_i
- All random vectors Z_i are assumed mutually independent and identically distributed with a distribution μ
  - X_i and Y_i have the marginal distributions μ_x and μ_y of μ

Assumption about moments

♦ Z_i and $F(X_i)$ have finite 2nd moments: $\mathbb{E}\|Z_i\|^2, \mathbb{E}\|F(X_i)\|^2 < +\infty$

- equivalently in terms of function spaces: $Z_i \in L_2(\mu), F(X_i) \in L_2(\mu_x)$
- for a bounded F, $F(X_i) \in L_2(\mu_x)$ already follows from $Z_i \in L_2(\mu)$

⟹1. For expectation and variance: $\mathbb{E}\, Z_i \in \mathbb{R}^{n+m}, \mathrm{Var}\, Z_i \in \mathbb{R}^{n+m,n+m}$

- for inputs and outputs: $\mathbb{E}\, X_i \in \mathbb{R}^n, \mathbb{E}\, Y_i \in \mathbb{R}^m, \mathrm{Var}\, X_i \in \mathbb{R}^{n,n}, \mathrm{Var}\, Y_i \in \mathbb{R}^{n,n}$

⟹ 2. For conditional moments: $\mathbb{E}(Y_i|X_i) \in L_2(\mu_x), \mathrm{Var}\,(Y_i|X_i) \in L_1(\mu_x)$

Expectation-based learning

- Expectation consideres all possible inputs + their probability
- This learning yields weights w and biasses b minimizing an

  expected loss $\mathbb{E}\mathcal{L}$ of network predictions $F_{(w,b)}(X)$ to outputs Y:

$$(w^*, b^*) = \arg\min_{(w,b)} \mathbb{E}_\mu \mathcal{L}\left(F_{(w,b)}(X), Y\right)$$

  - The *most common* loss − sum of squares (SSE):

$$(w^*, b^*) = \arg\min_{(w,b)} \text{SSE} = \arg\min_{(w,b)} \mathbb{E}_\mu \left\| F_{(w,b)}(X) - Y \right\|^2$$

Random-sample-based learning

- Typically, $\mathbb{E}_\mu \mathcal{L}\left(F_{(w,b)}(X), Y\right)$ cannot be computed $\Leftarrow \mu$ is unknown

- But for a random sample $(x_1, y_1), \ldots, \left(x_p, y_p\right)$, the mean

  $\frac{1}{p}\sum_{k=1}^{p} \mathcal{L}\left(F_{(w,b)}(x_k), y_k\right)$ is an *unbiased estimate of* $\mathbb{E}_\mu \mathcal{L}\left(F_{(w,b)}(X), Y\right)$

  - as $(x_1, y_1), \ldots, \left(x_p, y_p\right)$ can serve all / some training data

- Coincides with the traditional way of learning because

$$\min_{w,b} \frac{1}{p} \sum_{k=1}^{p} \mathcal{L}\left(F_{(w,b)}(x_k), y_k\right) = \min_{w,b} \sum_{k=1}^{p} \mathcal{L}\left(F_{(w,b)}(x_k), y_k\right)$$

Specificity of SSE learning

- Notation: $\langle\cdot,\cdot\rangle_{L_2(\mu)} \,|\, \|\cdot\|_{L_2(\mu)}$ – scalar product | norm in $L_2(\mu)$

- $\mathrm{SSE} = \left\|F_{(w,b)}(X) - Y\right\|^2_{L_2(\mu)} = \left\|F_{(w,b)}(X) - \mathbb{E}(Y|X)\right\|^2_{L_2(\mu)} +$

$+ \left\|\mathbb{E}(Y|X) - Y\right\|^2_{L_2(\mu)} + \left\langle F_{(w,b)}(X) - \mathbb{E}(Y|X), \mathbb{E}(Y|X) - Y\right\rangle_{L_2(\mu)}$

$$\left\langle F_{(w,b)}(X) - \mathbb{E}(Y|X), \mathbb{E}(Y|X) - Y \right\rangle_{L_2(\mu)}$$

$$\left\langle F_{(w,b)}(X) - \mathbb{E}(Y|X), \mathbb{E}(Y|X) - Y \right\rangle_{L_2(\mu)} =$$

$$= \mathbb{E}_{\mu} \left(F_{(w,b)}(X) - \mathbb{E}(Y|X) \right)^{\top} (\mathbb{E}(Y|X) - Y) =$$

$$= \mathbb{E}_{\mu_X} \left[\mathbb{E} \left(\left(F_{(w,b)}(X) - \mathbb{E}(Y|X) \right)^{\top} (\mathbb{E}(Y|X) - Y) \middle| X \right) \right] =$$

$$= \mathbb{E}_{\mu_X} \left[\left(F_{(w,b)}(X) - \mathbb{E}(Y|X) \right)^{\top} \mathbb{E}(\mathbb{E}(Y|X) - Y|X) \right] =$$

$$= \mathbb{E}_{\mu_X} \left[\left(F_{(w,b)}(X) - \mathbb{E}(Y|X) \right)^{\top} \left(\mathbb{E}(Y|X) - \mathbb{E}(Y|X) \right) \right] = 0$$

$\|\mathbb{E}(Y|X) - Y\|_{L_2(\mu)}^2$

$$\|\mathbb{E}(Y|X) - Y\|_{L_2(\mu)}^2 = \mathbb{E}_\mu \|\mathbb{E}(Y|X) - Y\|^2 =$$

$$= \mathbb{E}_{\mu_x} \mathbb{E}\left(\|\mathbb{E}(Y|X) - Y\|^2 \middle| X \right) =$$

$$= \mathbb{E}_{\mu_x} \left(\sum \left(\mathbb{E}_\mu (Y|X)_i - Y_i \right)^2 \middle| X \right) =$$

$$= \mathbb{E}_{\mu_x} \sum \mathrm{Var}(Y|X)_{i,i} =$$

$$= \mathbb{E}_{\mu_x} \mathrm{trace}\, \mathrm{Var}(Y|X)$$

Specificity of SSE learning

- Notation: $\langle\cdot,\cdot\rangle_{L_2(\mu)} \,|\, \|\cdot\|_{L_2(\mu)}$ – scalar product | norm in $L_2(\mu)$

- $\mathrm{SSE} = \left\|F_{(w,b)}(X) - Y\right\|^2_{L_2(\mu)} = \left\|F_{(w,b)}(X) - \mathbb{E}(Y|X)\right\|^2_{L_2(\mu)} +$

$$+ \left\|\mathbb{E}(Y|X) - Y\right\|^2_{L_2(\mu)} + \left\langle F_{(w,b)}(X) - \mathbb{E}(Y|X), \mathbb{E}(Y|X) - Y\right\rangle_{L_2(\mu)}$$

$$= \left\|F_{(w,b)}(X) - \mathbb{E}(Y|X)\right\|^2_{L_2(\mu)} + \mathbb{E}_{\mu_x}\mathrm{trace}\,\mathrm{Var}(Y|X)$$

- Thus $\arg\min\limits_{(w,b)} \mathrm{SSE} = \arg\min\limits_{(w,b)} \left\|F_{(w,b)}(X) - \mathbb{E}(Y|X)\right\|^2_{L_2(\mu)}$, and if exist

$(w,b)_{(Y|X)}$ such that $F_{(w,b)_{(Y|X)}}(X) = \mathbb{E}(Y|X)$, then $\arg\min\limits_{(w,b)} \mathrm{SSE} = (w,b)_{(Y|X)}$

Random-sample ⋈ expectation

- Expectation-based learning is not common because

  the distribution of learning samples is typically unknown

- But random sample empirical *mean estimates expectation*

  1. in an *unbiased* way: $\mathbb{E}_\mu \frac{1}{p}\sum_{k=1}^{p} \mathcal{L}(F(x_k), y_k) = \mathbb{E}_\mu \mathcal{L}(F(X), Y)$

  2. in a *consistent* way: $\frac{1}{p}\sum_{k=1}^{p} \mathcal{L}(F(x_k), y_k) \longrightarrow \mathbb{E}_\mu \mathcal{L}(F(X), Y)$

Laws of large numbers

- The consistence property, that $\frac{1}{p}\sum_{k=1}^{p}\mathcal{L}(F(x_k), y_k)$ converges to $\mathbb{E}_\mu \mathcal{L}(F(X), Y)$ is called law of large numbers.

- *Weak* law: convergence of random variables in *probability*

$$\forall \varepsilon > 0\colon \lim_{p\to\infty} \mu\left(\left|\frac{1}{p}\sum_{k=1}^{p}\mathcal{L}(F(x_k), y_k) - \mathbb{E}_\mu \mathcal{L}(F(X), Y)\right| > \varepsilon\right) = 0$$

- *Strong* law ($\Longrightarrow$ weak law): convergence *almost everywhere*

$$\mu\left(\lim_{p\to\infty}\frac{1}{p}\sum_{k=1}^{p}\mathcal{L}(F(x_k), y_k) = \mathbb{E}_\mu \mathcal{L}(F(X), Y)\right) = 1$$

Complete probability space

- Laws of large numbers cannot be directly applied

  to MLPs $\Longleftarrow \sum_{k=1}^{p} \mathcal{L}\left(F_{(w,b)}(x_k), y_k\right)$ changed by minimum

  - therefore, for MLPs, specific *additional assumptions* are needed

- Let $Z_i, i \in \mathbb{N}$, be Borel-measurable, $Z_i: (\Omega, \mathcal{A}, P) \to (\mathbb{R}^{n+m}, \mathcal{B}, \mu)$

- The probability space $(\Omega, \mathcal{A}, P)$ is assumed being *complete*:

$$A \in \mathcal{A} \& B \subset \Omega \& (A \backslash \mathrm{B}) \cup (B \backslash \mathrm{A}) \subset C \in \mathcal{A} \& P(C) = 0 \Longrightarrow B \in \mathcal{A}$$

Assumptions for the strong law

1. $(\Omega, \mathcal{A}, \mathrm{P})$ is a complete probability space
2. $(X_i, Y_i), \mathrm{i} \in \mathbb{N}$, are *i.i.d.* (independent and identically distributed)
3. $W = \{$admissible $(w, b) | w -$ weights$, b -$ bias$\}$ is a *compact* set
4. $(\forall (w,b) \in W) \mathcal{L}\left(F_{(w,b)}(x), y\right)$ is a Borel-*measurable* function of (x,y)
5. $(\forall (x,y) \in \mathbb{R}^{n+m}) \mathcal{L}\left(F_{(w,b)}(x), y\right)$ is a *W-continuous* function of (w, b)
6. $\mathcal{L}\left(F_{(w,b)}(X), Y\right)$ has an $\mathbb{R}^{n+m}$-integrable *majorizer* over W

Statement of the strong law

- Consider the set of expectation-based learning results

$$W^* = \left\{(w^*, b^*) \in W \,\middle|\, \mathbb{E}_\mu \mathcal{L}\left(F_{(w^*,b^*)}(X), Y\right) = \min_{(w,b)} \mathbb{E}_\mu \mathcal{L}\left(F_{(w,b)}(X), Y\right)\right\}$$

+ random-sample-based learning results for $(x_i, y_i)_{p=1}^{\infty}$

$$(\forall p \in \mathbb{N})\left(\widehat{w}_p, \hat{b}_p\right) = \arg\min_{w,b} \sum_{k=1}^{p} \mathcal{L}\left(F_{(w,b)}(x_k), y_k\right)$$

- Then $\left(\widehat{w}_p, \hat{b}_p\right)_{p=1}^{\infty}$ converges *almost everywhere to W^**

$$\mu\left(\lim_{p\to\infty} \inf_{(w^*,b^*)\in W^*} \left\|\left(\widehat{w}_p, \hat{b}_p\right) - (w^*, b^*)\right\| = 0\right) = 1$$

Network pruning

- *Removing connections* from fully connected networks
  - decreases the risk of overtraining + computational costs
- If all input connections / all output connections of a *hidden neuron* h are pruned, then h is removed
- *Formalised:* $S(w,b)=0$ with a 0/1-valued matrix S, rows contain for the 1 connection / for neuron's all connections + bias

Statistical approach to pruning

- Because $(\hat{w}_p, \hat{b}_p)$ that results from learning is only an estimate (unbiased + consistent) of (w^*, b^*), what we actually need is to know *whether* $S(w^*, b^*) = 0$
  - cannot be directly checked $\Longleftarrow (w^*, b^*)$ is not known
- Statistical approach to checking statements for estimated values: *hypotheses testing* using their *estimator* $\left((\hat{w}_p, \hat{b}_p)\right)$

Hypotheses testing recalled

- Testing a null hypotheses H_0 against H_1: checking whether $T \in \mathfrak{C}$
  - T – *test statistics*: random variable with some assumed distribution
  - $\mathfrak{C}$ – *critical set*: $\mathfrak{C} \subset \mathrm{Val}\, T$ with $H_0 \Rightarrow P(T \in \mathfrak{C} | H_0 \vee \neg H_1) \leq \alpha$ – significance
- The assumed T distribution can always asymptotically rely on

  the normality of $\frac{1}{\sqrt{p}} \sum_{k=1}^{p} \mathcal{L}(F(x_k), y_k) \Leftarrow$ *CLT* (central limit theorem)
  - not directly applicable $\Leftarrow \sum_{k=1}^{p} \mathcal{L}\left(F_{(w,b)}(x_k), y_k\right)$ changed by minimum

CLT for MLPs: assumptions 1.– 6.

1. $(\Omega, \mathcal{A}, \mathrm{P})$ is a complete probability space

2. $(X_i, Y_i), i \in \mathbb{N}$, are *i.i.d.* (independent and identically distributed)

3. $W = \{\text{admissible } (w, b) | w - \text{weights}, b - \text{bias}\}$ is a *compact* set

4. $W^* = \{(w^*, b^*)\}$ with (w^*, b^*) an *inner* point of W

5. $(\forall (w,b) \in W) \mathcal{L}\left(F_{(w,b)}(x), y\right)$ is a Borel-*measurable* function of (x,y)

6. $\mathcal{L}Z\left(F_{(w,b)}(X), Y\right)$ has an $\mathbb{R}^{n+m}$-integrable *majorizer* over W

CLT for MLPs: auxiliary notation

- $\nabla_{(\mathrm{w,b})}\mathcal{L} = \nabla_{(\mathrm{w,b})}\mathcal{L}\big(\mathrm{F}_{(\mathrm{w,b})}(\mathrm{X}), \mathrm{Y}\big)$: a random vector such that

  $(\forall (x,y) \in \mathbb{R}^{n+m})\, \nabla_{(\mathrm{w,b})}\mathcal{L} =$ the *gradient* of $\mathcal{L}\big(F_{(w,b)}(x), y\big)$ w.r.t. to (w, b)

- $\nabla^2_{(w,b)}\mathcal{L}\big(F_{(w,b)}(X), Y\big)$: a random matrix such that $(\forall (x,y) \in \mathbb{R}^{n+m})$

  $\nabla^2_{(\mathrm{w,b})}\mathcal{L}\big(\mathrm{F}_{(\mathrm{w,b})}(\mathrm{x}), \mathrm{y}\big) =$ the *Hessian* of $\mathcal{L}\big(F_{(w,b)}(x), y\big)$ w.r.t. to (w, b)

CLT for MLPs: assumptions 7.– 12.

7. $(\forall x, y)\mathcal{L}\left(F_{(w,b)}(x), y\right)$ has W-continuous Hessian w.r. to (w, b)
8. The matrix A^* defined $A^* = \mathbb{E}_\mu\left(\nabla^2_{(w^*,b^*)}\mathcal{L}\left(\mathrm{F}_{(\mathrm{w,b})}(\mathrm{X}), \mathrm{Y}\right)\right)$ is regular
9. $\nabla^2_{(\mathrm{w,b})}\mathcal{L}\left(\mathrm{F}_{(\mathrm{w,b})}(\mathrm{x}), \mathrm{y}\right)$ has an $\mathbb{R}^{n+m}$-integrable majorizer over W
10. The matrix B^* defined $B^* = \mathbb{E}_\mu\left(\nabla_{(\mathrm{w,b})}\mathcal{L}^\intercal\nabla_{(\mathrm{w,b})}\mathcal{L}\right)$ is regular
11. $\left\|\mathcal{L}\left(\mathrm{F}_{(\mathrm{w,b})}(\mathrm{x}), y\right)\right\|^2$ has an $\mathbb{R}^{n+m}$-integrable majorizer over W
12. A $\{0,1\}$-valued matrix *S has s = rank S rows*

CLT for MLPs: auxiliary notation

- $\nabla_{(\mathrm{w,b})}\mathcal{L} = \nabla_{(\mathrm{w,b})}\mathcal{L}\big(\mathrm{F}_{(\mathrm{w,b})}(\mathrm{X}), \mathrm{Y}\big)$: a random vector such that

  $(\forall(x,y) \in \mathbb{R}^{n+m})\nabla_{(\mathrm{w,b})}\mathcal{L} =$ the *gradient* of $\mathcal{L}\big(F_{(w,b)}(x), y\big)$ w.r.t. to (w, b)

- $\nabla^2_{(w,b)}\mathcal{L}\big(F_{(w,b)}(X), Y\big)$: a random matrix such that $(\forall(x, y) \in \mathbb{R}^{n+m})$

  $\nabla^2_{(\mathrm{w,b})}\mathcal{L}\big(\mathrm{F}_{(\mathrm{w,b})}(\mathrm{x}), \mathrm{y}\big) =$ the *Hessian* of $\mathcal{L}\big(F_{(w,b)}(x), y\big)$ w.r.t. to (w, b)

- $\hat{A}_p = \frac{1}{p}\sum_{i=1}^{p} \nabla^2_{(\mathrm{w,b})}\mathcal{L}\left(F_{(\hat{w}_p, \hat{b}_p)}(x_i), y_i\right)$

- $\hat{B}_p = \frac{1}{p}\sum_{i=1}^{p} \nabla_{(\mathrm{w,b})}\mathcal{L}\left(\mathrm{F}_{(\hat{w}_p, \hat{b}_p)}(x_i), y_i\right)\mathcal{L}\left(\mathrm{F}_{(\hat{w}_p, \hat{b}_p)}(x_i), y_i\right)^{\top}$

CLT for MLPs: exact covariance

- $\left(\sqrt{p}\left((\hat{w}_p,\hat{b}_p)-(w^*,b^*)\right)\right)_{p=1}^{\infty}$ converges to the *distribution* $N(0,C^*)$, the covariance matrix of which is $C^* = {A^*}^{-1}B^*{A^*}^{-1}$

- If $S(w^*,b^*) = 0$, then $\left(\sqrt{p}S(\hat{w}_p,\hat{b}_p)\right)_{p=1}^{\infty}$ converges to $N\left(0,SC^*S^\intercal\right)$

- If $S(w^*,b^*) = 0$, then the quadratic forms of $\left(\sqrt{p}S(\hat{w}_p,\hat{b}_p)\right)_{p=1}^{\infty}$

$$\left(p(\hat{w}_p,\hat{b}_p)^\intercal S^\intercal(SC^*S^\intercal)^{-1}S(\hat{w}_p,\hat{b}_p)\right)_{p=1}^{\infty}$$

converge to the *distribution* χ^2_s with s degrees of freedom

CLT for MLPs: estimated covariance

- Define an estimate $\hat{C}_p = \begin{cases} \hat{A}_p^{-1}\hat{B}_p\hat{A}_p^{-1} & \text{if } p \in \mathbb{N}, \hat{A}_p \text{ is regular} \\ \hat{B}_p & \text{if } p \in \mathbb{N}, \hat{A}_p \text{ is singular} \end{cases}$

- Then $\hat{C}_p \longrightarrow C^*$ in probability $\left(\Longrightarrow \lim_{p\to\infty} \mu\left(\hat{C}_p \text{ is regular}\right) = 1 \right)$

- If $S\,(w^*, b^*) = 0$, then also the quadratic forms

$\left(p\left(\widehat{w}_p, \widehat{b}_p\right)^{\top} S^{\top}\left(S\hat{C}_p S^{\top}\right)^{-1} S\left(\widehat{w}_p, \widehat{b}_p\right) \right)_{p=1}^{\infty}$ converge to the *distribution* χ_s^2

Test procedure for $S(w^*, b^*) = 0$

1. For given observations $(x_1,y_1),\ldots,(x_p,y_p)$, get [= compute] $(\hat{w}_p,\hat{b}_p)$
2. For $i = 1,\ldots,p$, get $\nabla_{(\mathrm{w,b})}\mathcal{L}\left(\mathrm{F}_{(\hat{w}_p,\hat{b}_p)}(x_i),y_i\right)$ and $\nabla^2_{(\mathrm{w,b})}\mathcal{L}\left(F_{(\hat{w}_p,\hat{b}_p)}(x_i),y_i\right)$
3. Get $\hat{A}_p, \hat{B}_p$ and check whether any is singular
4. Then the test cannot proceed, else get $\hat{C}_p$
5. Get $p(\hat{w}_p,\hat{b}_p)^{\top} S^{\top}(S\hat{C}_p S^{\top})^{-1} S(\hat{w}_p,\hat{b}_p)$ and compare with the distribution χ^2_s
6. If > the quantile $\chi^2_s(1-\alpha), a \in (0,1)$, ***reject*** $S\ (w^*, b^*) = 0$