Moderne Methoden der Datenanalyse
– Ereignisklassifikation –

Roger Wolf

23. Juli 2020

Inhalt der Vorlesung

- Wann war das Training erfolgreich?
- Konfidenz in die Entscheidung eines MLP.
- Ausflug ins unboxing eines MLPs.

Erfolg nach Training

- Der Erfolg eines Trainings bei der Anpassung einer MLP Architektur an eine vorgegebene Aufgabe (*task*) wird i.a. durch Vergleich mit den *Labels* auf dem (*gelabelten*) Validierungs-datensatz ($\mathcal{V}$) eingeschätzt:

- Dabei geht es um die Überprüfung der MLP Vorhersage y_i mit dem *Label* p_i der Grund-gesamtheit.

- Stimmt die Vorhersage „häufig genug“ mit dem *Label* überein war die Anpassung des MLP Modells (an den Trainingsdatensatz) erfolgreich.

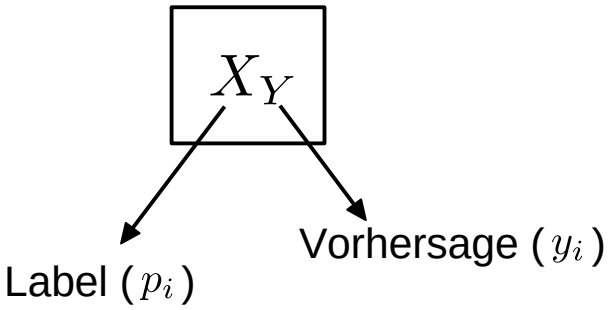
Binäre Klassifikation

- Für den Spezialfall binärer Klassifikation lässt sich diese Einschätzung auf die Diskussion binärer Hypothesentests zurückführen:

- Hierzu definieren wir:

H_1 : Zu testende Hypothese, Signal

H_0 : Alternative Hypothese, Untergrund



y_i \ p_i	H_0	H_1
$\hat{H}_0$	T_N	F_N
$\hat{H}_1$	F_P	T_P

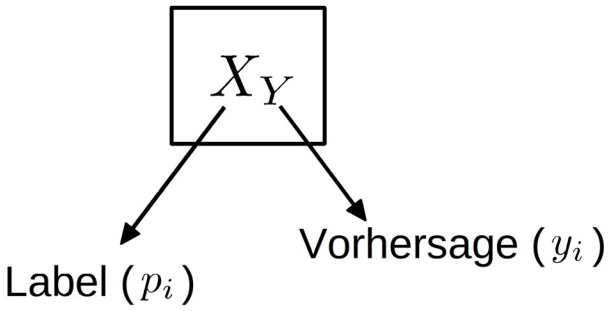
H_0 ist wahr (d.h. kein Singal)

- Für den Spezialfall binärer Klassifikation lässt sich diese Einschätzung auf die Diskussion binärer Hypothesentests zurückführen:

- Hierzu definieren wir:

H_1 : Zu testende Hypothese, Signal

H_0 : Alternative Hypothese, Untergrund



y_i \ p_i	H_0	H_1
$\hat{H}_0$	T_N $t_N = \frac{T_N}{F_P + T_N}$ • Spezifität • Specifity • **True negative rate (TNR)**	F_N
$\hat{H}_1$	F_P	T_P

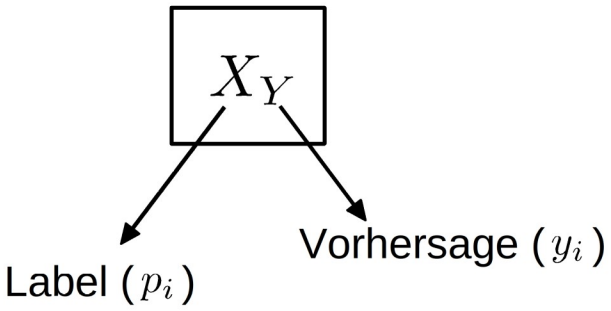
H_0 ist wahr (d.h. kein Singal)

- Für den Spezialfall binärer Klassifikation lässt sich diese Einschätzung auf die Diskussion binärer Hypothesentests zurückführen:

- Hierzu definieren wir:

H_1 : Zu testende Hypothese, Signal

H_0 : Alternative Hypothese, Untergrund



y_i \ p_i	H_0	H_1
$\hat{H}_0$	T_N	F_N
$\hat{H}_1$	F_P $f_P = \frac{F_P}{F_P + T_N}$ • Ausfallrate • **False positive rate (FPR)** • Fallout	T_P

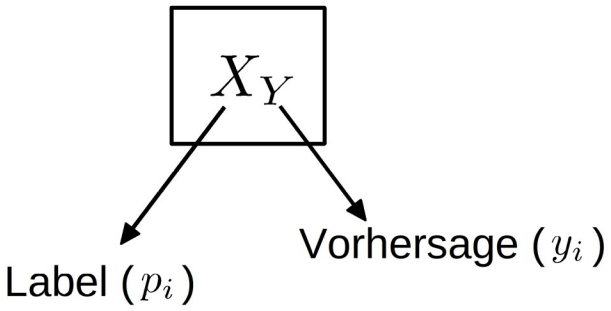
H_1 ist wahr (d.h. Signal)

- Für den Spezialfall binärer Klassifikation lässt sich diese Einschätzung auf die Diskussion binärer Hypothesentests zurückführen:

- Hierzu definieren wir:

H_1 : Zu testende Hypothese, Signal

H_0 : Alternative Hypothese, Untergrund



y_i \ p_i	H_0	H_1
$\hat{H}_0$	T_N	F_N $f_N = \frac{F_N}{T_P + F_N}$ • **False negative rate (FNR)** • Miss rate
$\hat{H}_1$	F_P	T_P

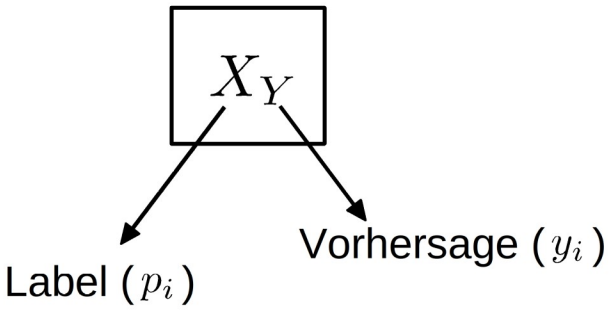
H_1 ist wahr (d.h. Signal)

- Für den Spezialfall binärer Klassifikation lässt sich diese Einschätzung auf die Diskussion binärer Hypothesentests zurückführen:

- Hierzu definieren wir:

H_1 : Zu testende Hypothese, Signal

H_0 : Alternative Hypothese, Untergrund



y_i \ p_i	H_0	H_1
$\hat{H}_0$	T_N	F_N
$\hat{H}_1$	F_P	T_P $t_P = \frac{T_P}{T_P + F_N}$ • Sensitivität • Empfindlichkeit • **True positive rate (TPR)** • Recall • Hit rate

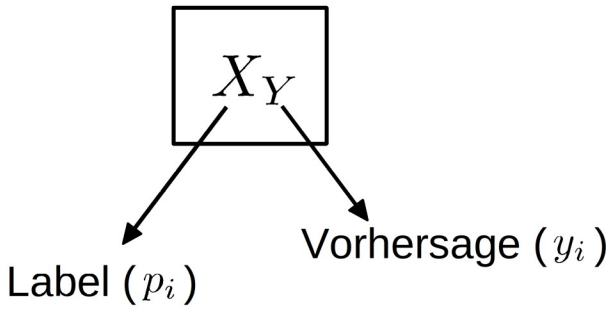
$\hat{H}_0$ wurde klassifiziert

- Für den Spezialfall binärer Klassifikation lässt sich diese Einschätzung auf die Diskussion binärer Hypothesentests zurückführen:

- Hierzu definieren wir:

H_1 : Zu testende Hypothese, Signal

H_0 : Alternative Hypothese, Untergrund



y_i \ p_i	H_0	H_1
$\hat{H}_0$	T_N $\tau_N = \frac{T_N}{F_N + T_N}$ • Trennfähigkeit • Segreganz • **Negative predictive value (NPV)**	F_N
$\hat{H}_1$	F_P	T_P

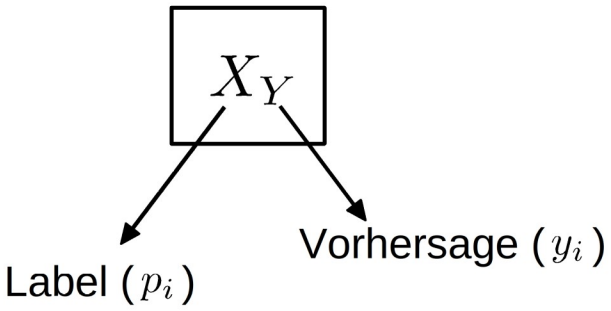
$\hat{H}_1$ wurde klassifiziert

- Für den Spezialfall binärer Klassifikation lässt sich diese Einschätzung auf die Diskussion binärer Hypothesentests zurückführen:

- Hierzu definieren wir:

H_1 : Zu testende Hypothese, Signal

H_0 : Alternative Hypothese, Untergrund



y_i \ p_i	H_0	H_1
$\hat{H}_0$	T_N	F_N
$\hat{H}_1$	F_P	T_P $\tau_P = \frac{T_P}{T_P + F_P}$ • Relevanz • Genauigkeit • **Positive predictive value (PPV)** • Precision

Fehler 1. und 2. Art

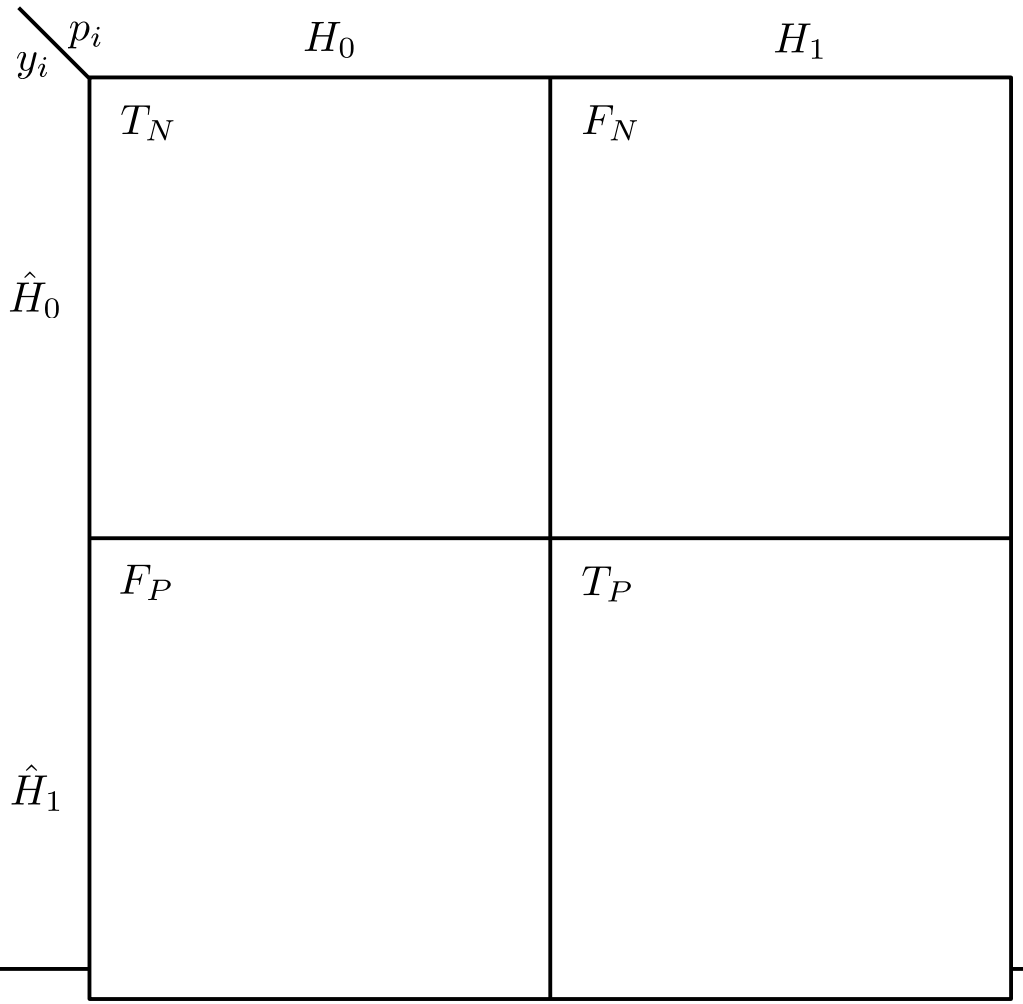
- Für den Spezialfall binärer Klassifikation lässt sich diese Einschätzung auf die Diskussion binärer Hypothesentests zurückführen:

- Hierzu definieren wir:

  H_1 : Zu testende Hypothese, Signal

  H_0 : Alternative Hypothese, Untergrund

- Machen Sie sich nochmal den Zusammenhang zwischen diesen Begriffen und dem Fehler 1. Art (α) und 2. Art (β) beim Hypothesentest klar.



Fehler 1. und 2. Art

- Für den Spezialfall binärer Klassifikation lässt sich diese Einschätzung auf die Diskussion binärer Hypothesentests zurückführen:

- Hierzu definieren wir:

H_1 : Zu testende Hypothese, Signal

H_0 : Alternative Hypothese, Untergrund

- Machen Sie sich nochmal den Zusammenhang zwischen diesen Begriffen und dem Fehler 1. Art (α) und 2. Art (β) beim Hypothesentest klar.

y_i \ p_i	H_0	H_1
$\hat{H}_0$	T_N	F_N $\boxed{\beta}$ $f_N = \frac{F_N}{T_P + F_N}$ • **False negative rate (FNR)** • Miss rate
$\hat{H}_1$	F_P $\boxed{\alpha}$ $f_P = \frac{F_P}{F_P + T_N}$ • Ausfallrate • **False positive rate (FPR)** • Fallout	T_P

Gütefunktion/Trennschärfe

- Für den Spezialfall binärer Klassifikation lässt sich diese Einschätzung auf die Diskussion binärer Hypothesentests zurückführen:

- Hierzu definieren wir:

  H_1 : Zu testende Hypothese, Signal
  H_0 : Alternative Hypothese, Untergrund

- Machen Sie sich nochmal den Zusammenhang zwischen diesen Begriffen und dem Fehler 1. Art (α) und 2. Art (β) beim Hypothesentest klar.

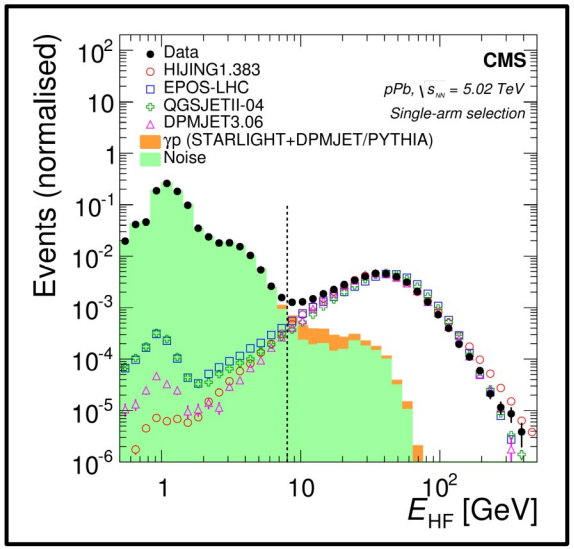
- Die Funktion $1 - \beta(\alpha, c, n)$ bezeichnet man als **Gütefunktion** oder **Trennschärfe** des Hypothesentests.

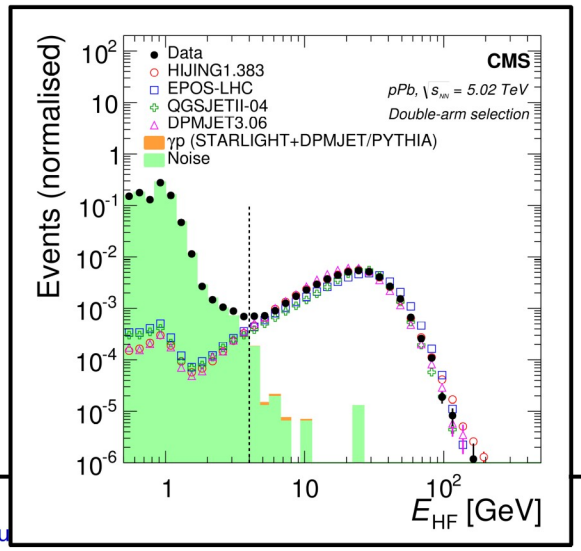
y_i \ p_i	H_0	H_1
$\hat{H}_0$	T_N	F_N β $f_N = \frac{F_N}{T_P + F_N}$ • **False negative rate (FNR)** • Miss rate
$\hat{H}_1$	F_P α $f_P = \frac{F_P}{F_P + T_N}$ • Ausfallrate • **False positive rate (FPR)** • Fallout	T_P

Dabei ist c der kritische Wert von y_i der über die Annahme von $\hat{H}_{0/1}$ entscheidet.

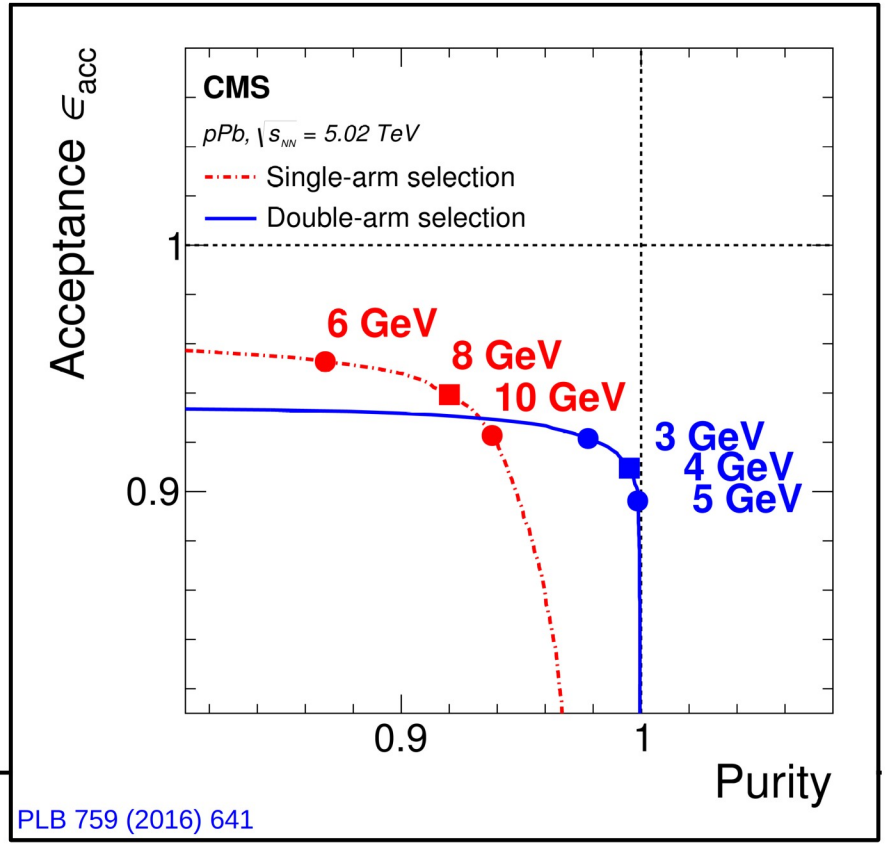
ROC Kurve

- Bei binärer Klassifikation wird die **Trennschärfe/Gütefunktion** eines Hypothesentests oft durch die *Reciever Operating Characteristics* (**ROC**) Kurve angegeben.





Ein Beispiel aus der Teilchenphysik:

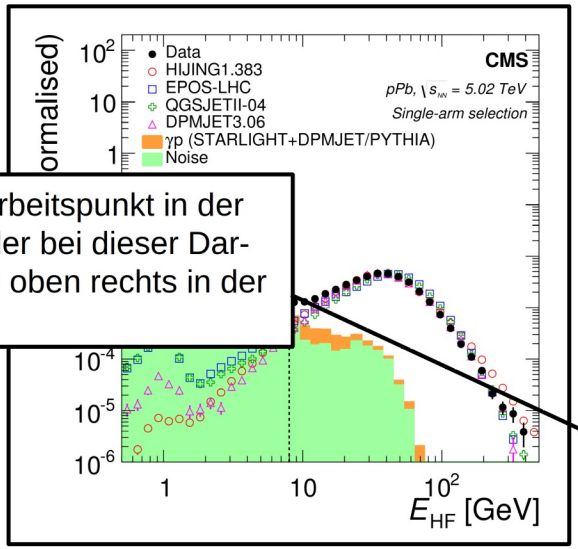


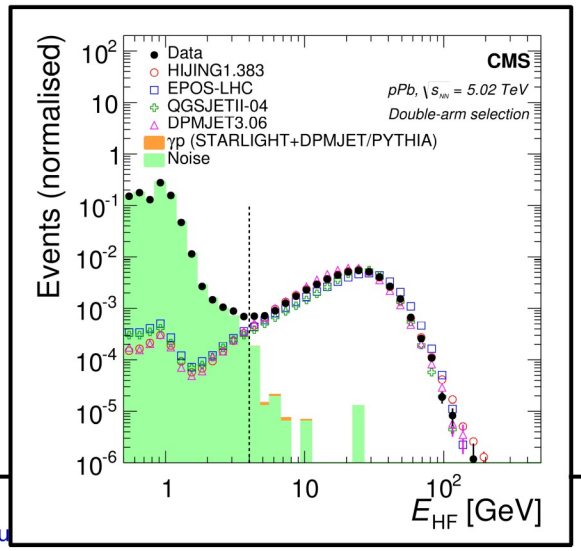
PLB 759 (2016) 641

ROC Kurve

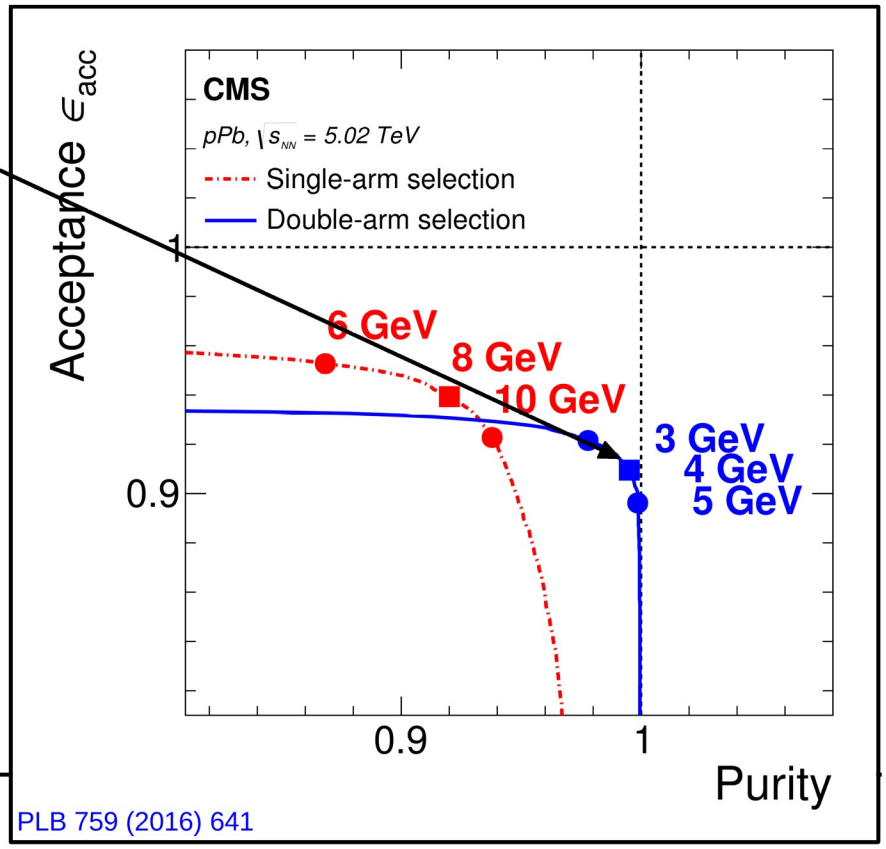
- Bei binärer Klassifikation wird die **Trennschärfe/Gütefunktion** eines Hypothesentests oft durch die *Reciever Operating Characteristics* (**ROC**) Kurve angegeben.

Sie würden i.a. den Arbeitspunkt in der ROC Kurve wählen, der bei dieser Darstellung am weitesten oben rechts in der Kurve liegt.





Ein Beispiel aus der Teilchenphysik:

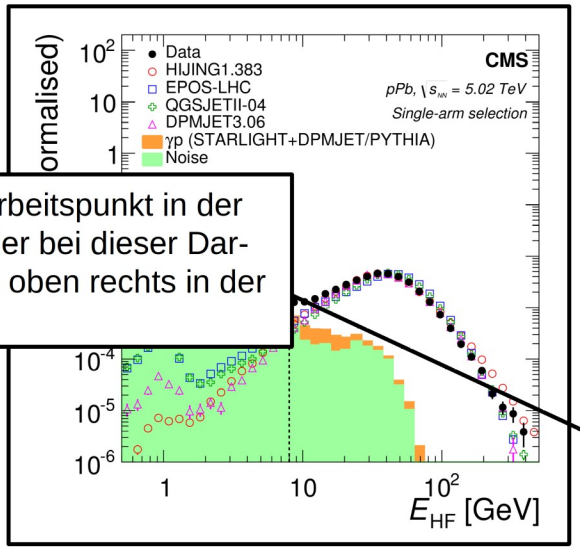


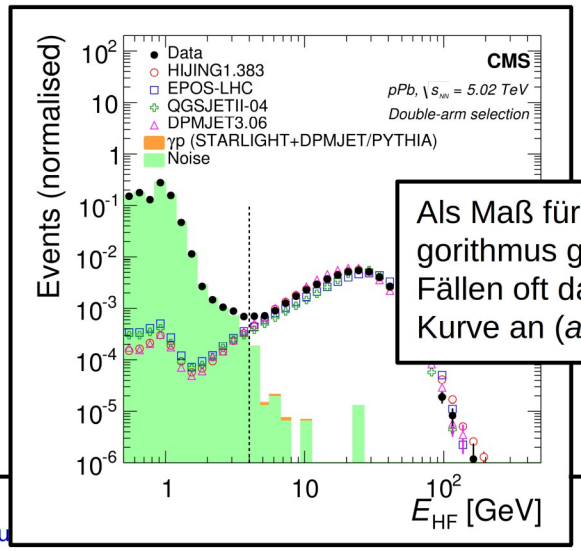
PLB 759 (2016) 641

ROC Kurve

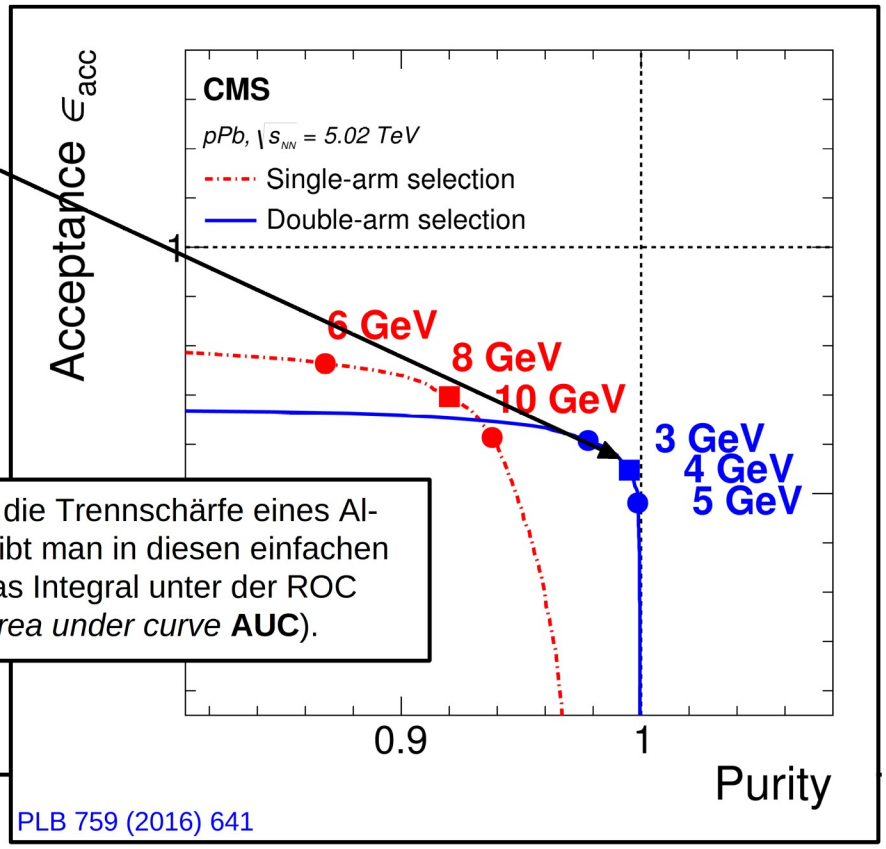
- Bei binärer Klassifikation wird die **Trennschärfe/Gütefunktion** eines Hypothesentests oft durch die *Reciever Operating Characteristics* (**ROC**) Kurve angegeben.

Sie würden i.a. den Arbeitspunkt in der ROC Kurve wählen, der bei dieser Darstellung am weitesten oben rechts in der Kurve liegt.





Ein Beispiel aus der Teilchenphysik:

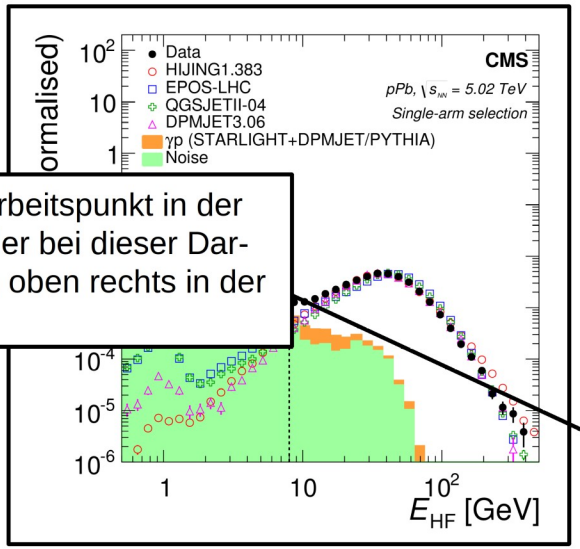


Als Maß für die Trennschärfe eines Algorithmus gibt man in diesen einfachen Fällen oft das Integral unter der ROC Kurve an (*area under curve* **AUC**).

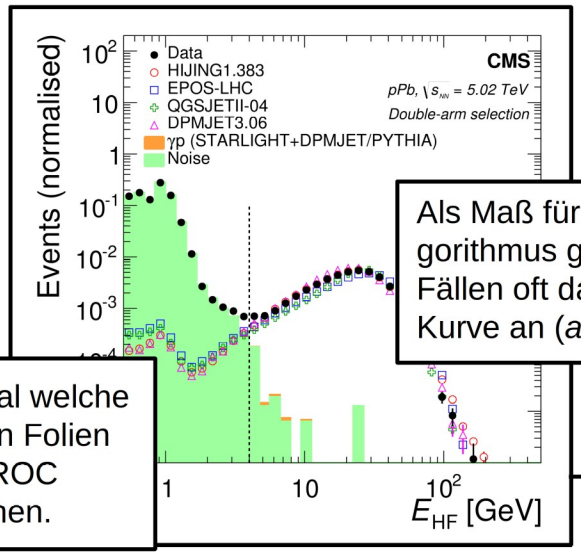
PLB 759 (2016) 641

ROC Kurve

- Bei binärer Klassifikation wird die **Trennschärfe/Gütefunktion** eines Hypothesentests oft durch die *Reciever Operating Characteristics* (**ROC**) Kurve angegeben.

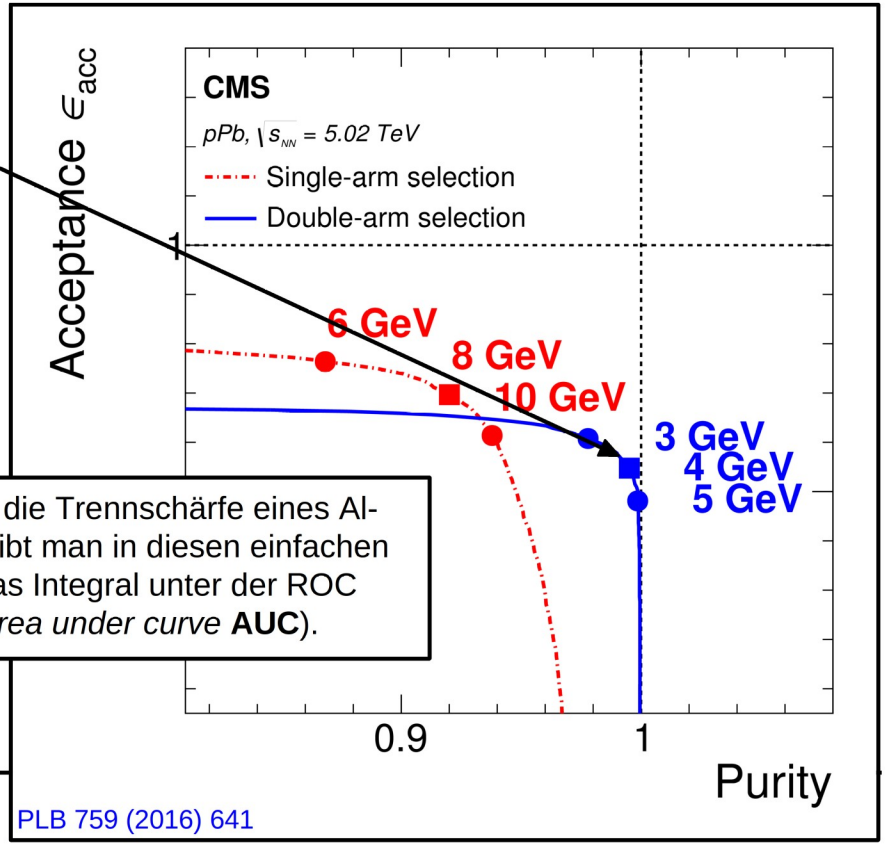


Sie würden i.a. den Arbeitspunkt in der ROC Kurve wählen, der bei dieser Darstellung am weitesten oben rechts in der Kurve liegt.



Überlegen Sie sich nochmal welche Größen aus den vorherigen Folien bei diesem Beispiel einer ROC Kurve auf den Achsen stehen.

Ein Beispiel aus der Teilchenphysik:



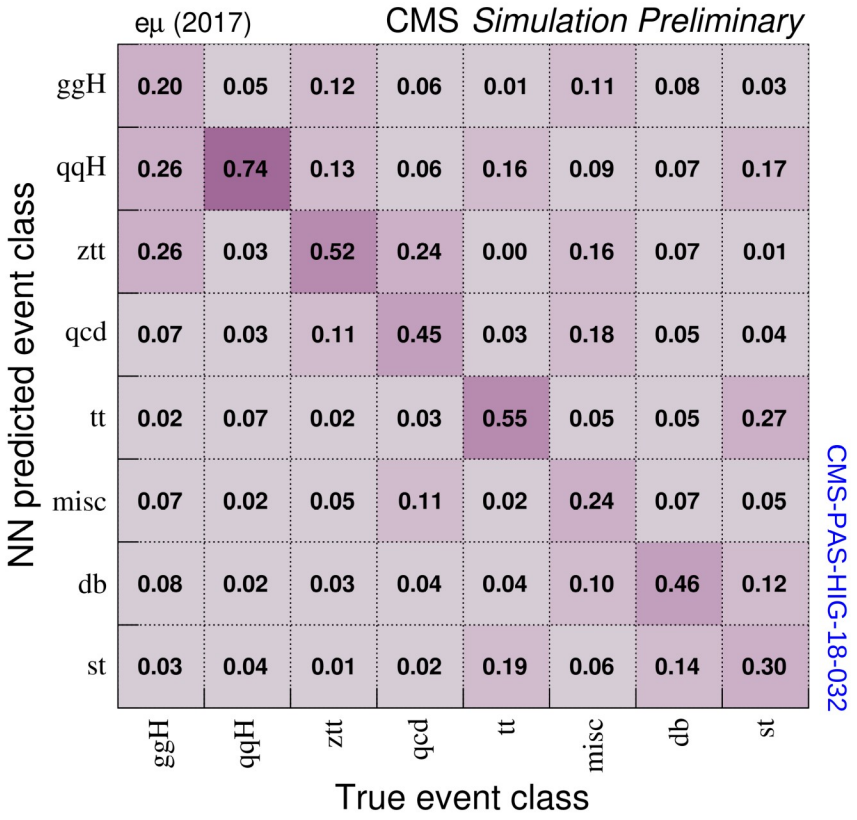
Als Maß für die Trennschärfe eines Algorithmus gibt man in diesen einfachen Fällen oft das Integral unter der ROC Kurve an (*area under curve* **AUC**).

Konfusionsmatrix

- Zur Anwendung einer ROC Kurve im Fall der **Multiklassifikation** muss diese ggf. auf paarweise binäre Klassifikation reduziert werden.

- Alternativ erfolgt die Bewertung in Form einer Konfusionsmatrix (*confusion matrix*).

- Hier erwarten Sie bevorzugt hohe Einträge auf der Diagonalen.

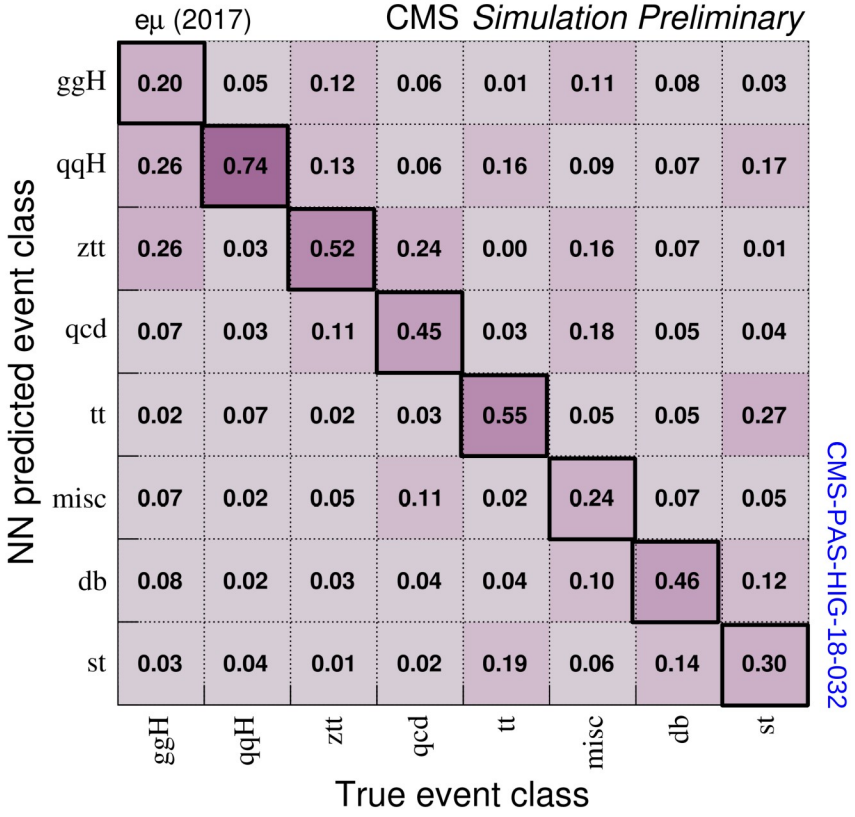


Konfusionsmatrix

- Zur Anwendung einer ROC Kurve im Fall der **Multiklassifikation** muss diese ggf. auf paarweise binäre Klassifikation reduziert werden.

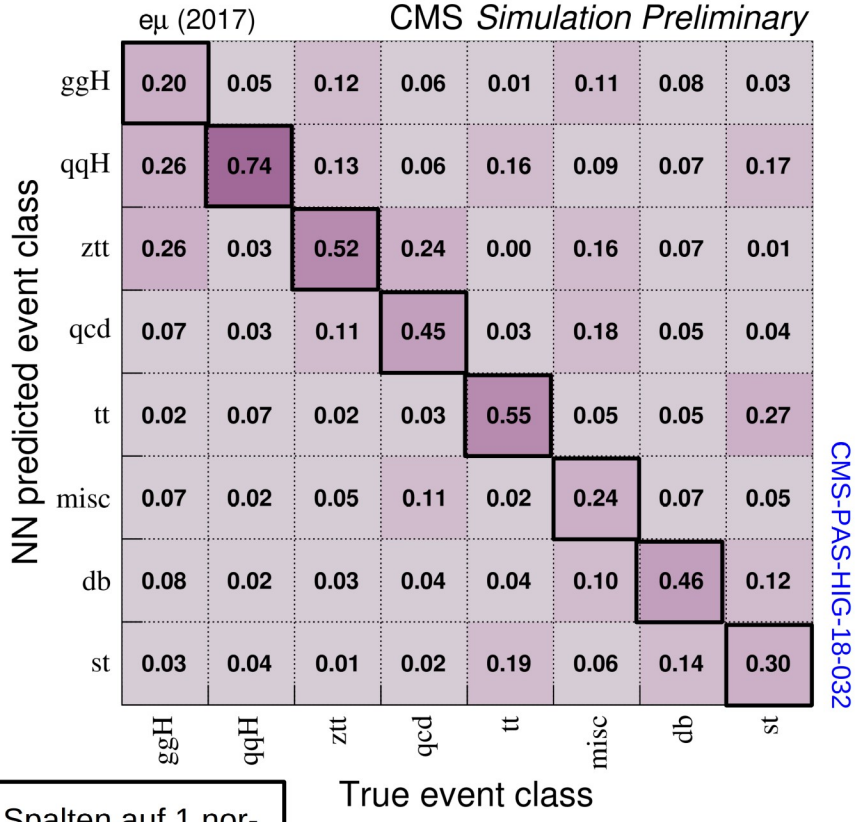
- Alternativ erfolgt die Bewertung in Form einer Konfusionsmatrix (*confusion matrix*).

- Hier erwarten Sie bevorzugt hohe Einträge auf der Diagonalen.



Konfusionsmatrix

- Zur Anwendung einer ROC Kurve im Fall der **Multiklassifikation** muss diese ggf. auf paarweise binäre Klassifikation reduziert werden.
- Alternativ erfolgt die Bewertung in Form einer Konfusionsmatrix (*confusion matrix*).
- Hier erwarten Sie bevorzugt hohe Einträge auf der Diagonalen.
- Es gibt mehrere Varianten von Konfusionsmatrizen abhängig davon, wie die Einträge normiert wurden.

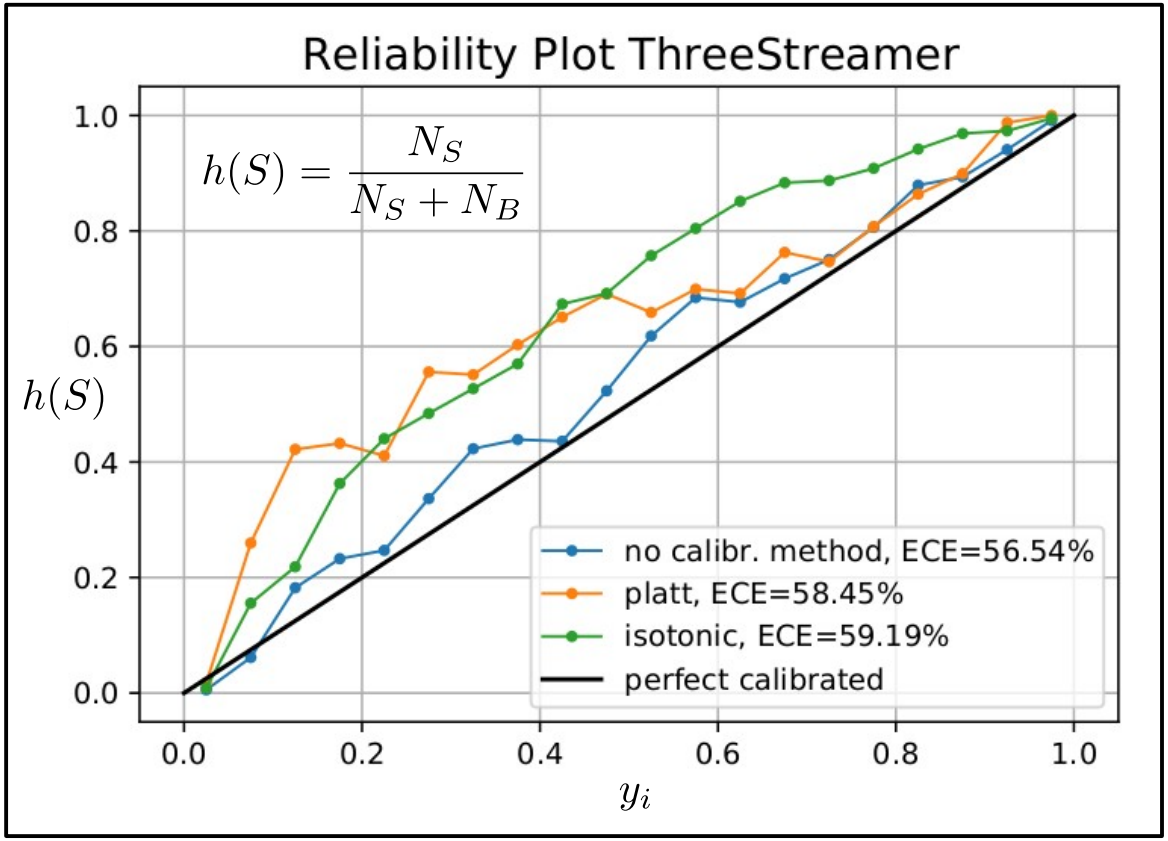


Hier wurden die Spalten auf 1 normiert, d.h. der Diagonaleneintrag entspricht der TPR (Reinheit).

Kalibration der MLP Antwort

- Eine weitere oftmals wünschenswerte Eigenschaft des MLP Antwortverhaltens ist, dass die Verteilung der y_i Werte ($g(y_i)$) auf $\mathcal{V}$ wirklich der (frequentistischen) rel. Häufigkeitsverteilung $h(S)$ für Signal entspricht:

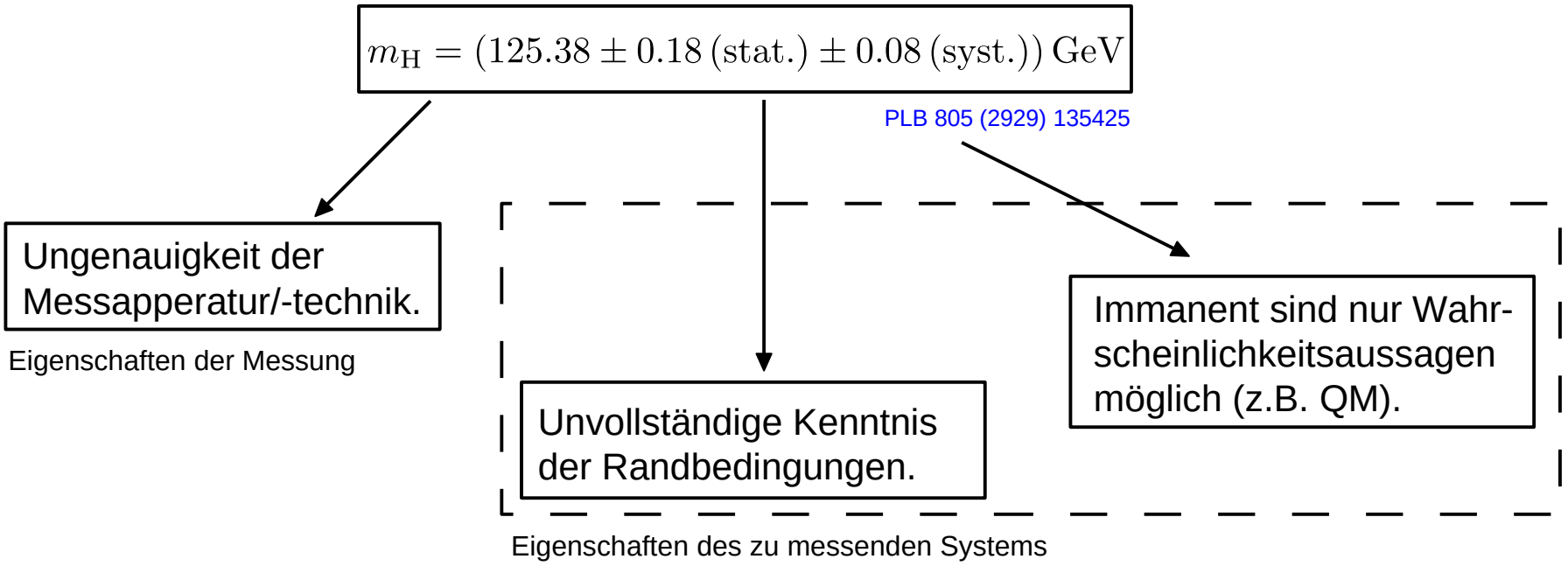
- Ist dies nicht der Fall können Sie den Versuch unternehmen y_i zu kalibieren.



Konfidenz

- Unser Anspruch in der Physik ist die Angabe von Messwerten unter Angabe von **Konfidenz-intervallen** (*confidence/credibility interval*).

- Hier ein Beispiel aus der Teilchenphysik:



Konfidenz ins Antwortverhalten des MLP?

- Welche (zusätzlichen) Messunsicherheiten ergeben sich aus der Entscheidung des MLP?

Konfidenz ins Antwortverhalten des MLP?

- Welche (zusätzlichen) Messunsicherheiten ergeben sich aus der Entscheidung des MLP?

- Unterscheidung von Unsicherheiten in Bereich des *machine learning*:

Machine learning

Aleatorische (*aleatoric*) Unsicherheiten:

Epistemische (*epistemic*) Unsicherheiten:

Konfidenz ins Antwortverhalten des MLP?

- Welche (zusätzlichen) Messunsicherheiten ergeben sich aus der Entscheidung des MLP?

- Unterscheidung von Unsicherheiten in Bereich des *machine learning*:

Machine learning

Aleatorische (*aleatoric*) Unsicherheiten:

Inhärente Unsicherheiten aufgrund statistischer Varianz.

Epistemische (*epistemic*) Unsicherheiten:

Konfidenz ins Antwortverhalten des MLP?

- Welche (zusätzlichen) Messunsicherheiten ergeben sich aus der Entscheidung des MLP?

- Unterscheidung von Unsicherheiten in Bereich des *machine learning*:

Machine learning

Aleatorische (*aleatoric*) Unsicherheiten:

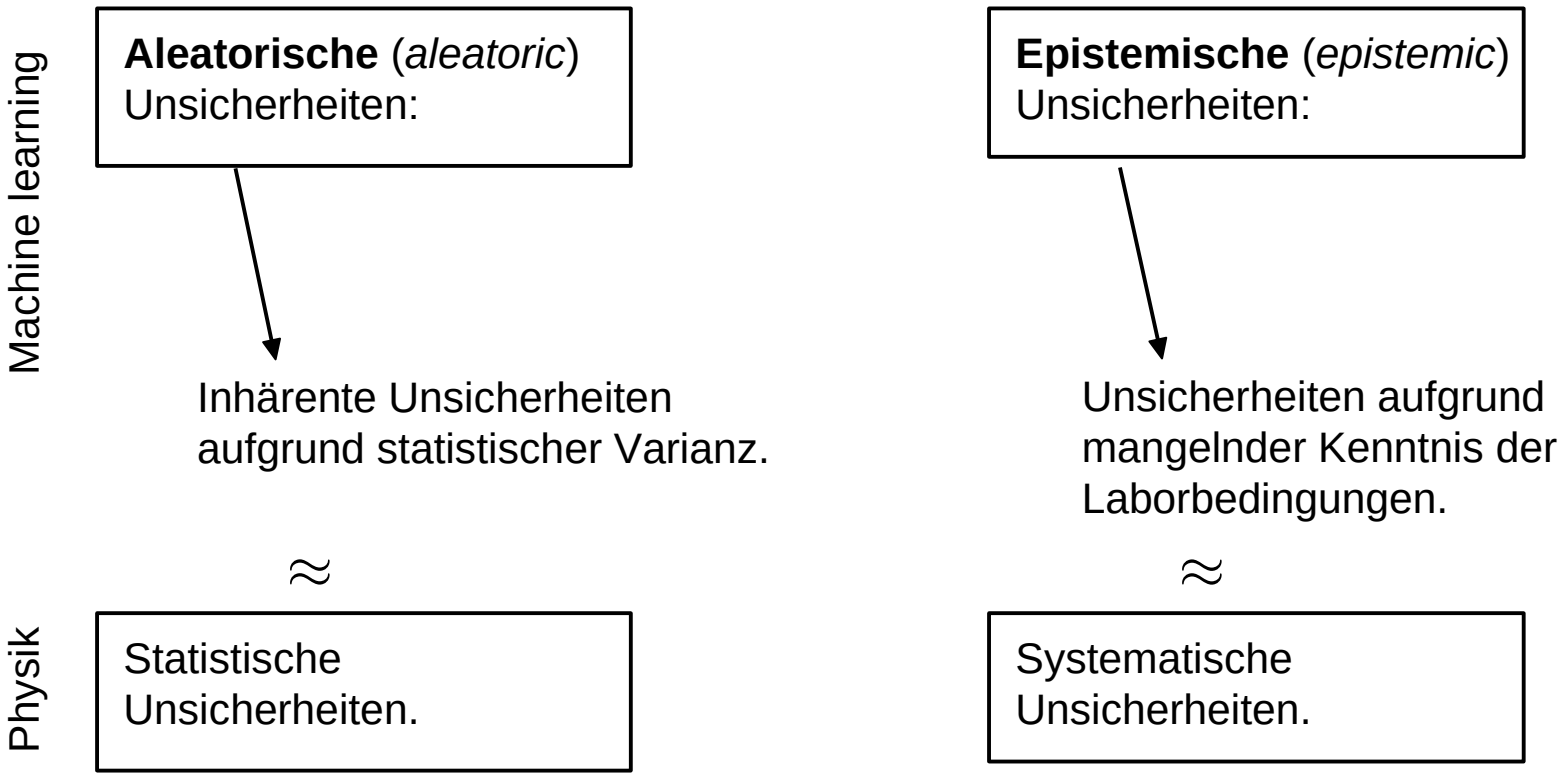
Inhärente Unsicherheiten aufgrund statistischer Varianz.

Epistemische (*epistemic*) Unsicherheiten:

Unsicherheiten aufgrund mangelnder Kenntnis der Laborbedingungen.

Konfidenz ins Antwortverhalten des MLP?

- Welche (zusätzlichen) Messunsicherheiten ergeben sich aus der Entscheidung des MLP?
- Unterscheidung von Unsicherheiten in Bereich des *machine learning*:



NB: Wir sprechen hier nur von Unsicherheiten der MLP Entscheidung selbst, nicht von Unsicherheiten einer physikalischen Messung, die auf einer MLP-Entscheidung beruht.

Epistemische Unsicherheiten

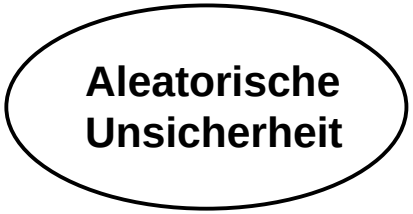
- Für ein MLP ergeben sich epistemische Usicherheiten daraus, dass das verwendete Modell nicht genügend Freiheitsgrade besitzt, um die Grundgesamtheit abbilden zu können (→ ***underfitting***) …

Epistemische Unsicherheiten

- Für ein MLP ergeben sich epistemische Usicherheiten daraus, dass das verwendete Modell nicht genügend Freiheitsgrade besitzt, um die Grundgesamtheit abbilden zu können (→ ***underfitting***), oder daraus, dass das verwendete Modell spezifische Eigenschaften des Trainingsdatensatzes statt allgemeinen Eigenschaften der Grundgesamtheit wiedergibt (→ ***overfitting***).

Aleatorische Unsicherheiten

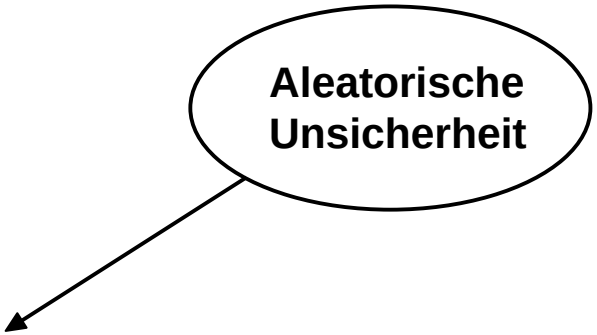
- Aleatorische Unsicherheiten wohnen dem MLP-Ansatz aufgrund seiner statistischen Natur inne:



Wo entdecken Sie statistische Prozesse bei der Verwendung von MLPs?

Aleatorische Unsicherheiten

- Aleatorische Unsicherheiten wohnen dem MLP-Ansatz aufgrund seiner statistischen Natur inne:

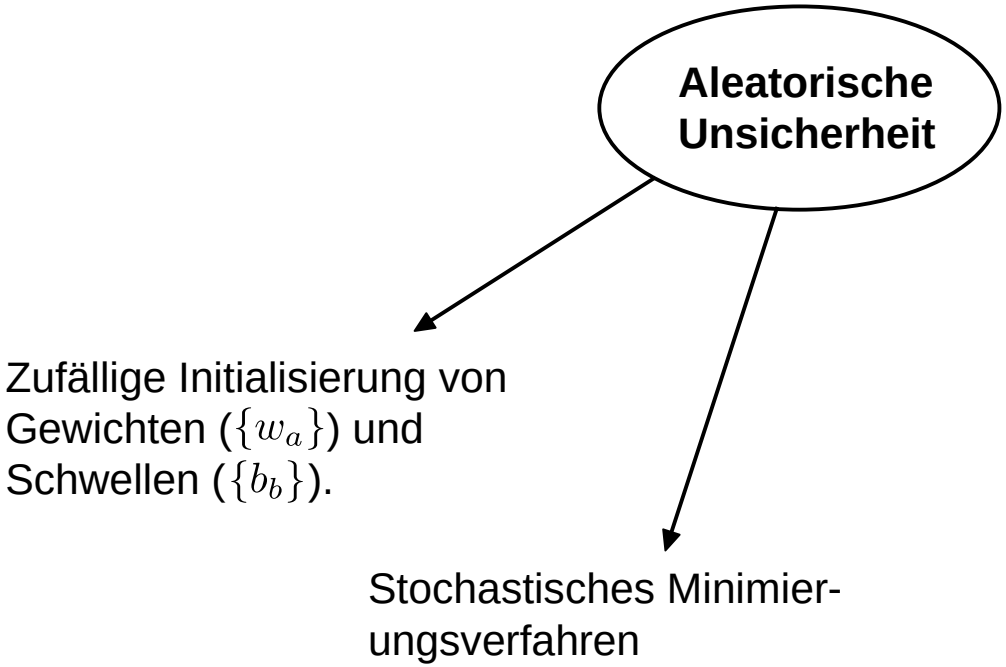


Zufällige Initialisierung von Gewichten ($\{w_a\}$) und Schwellen ($\{b_b\}$).

Wo entdecken Sie statistische Prozesse bei der Verwendung von MLPs?

Aleatorische Unsicherheiten

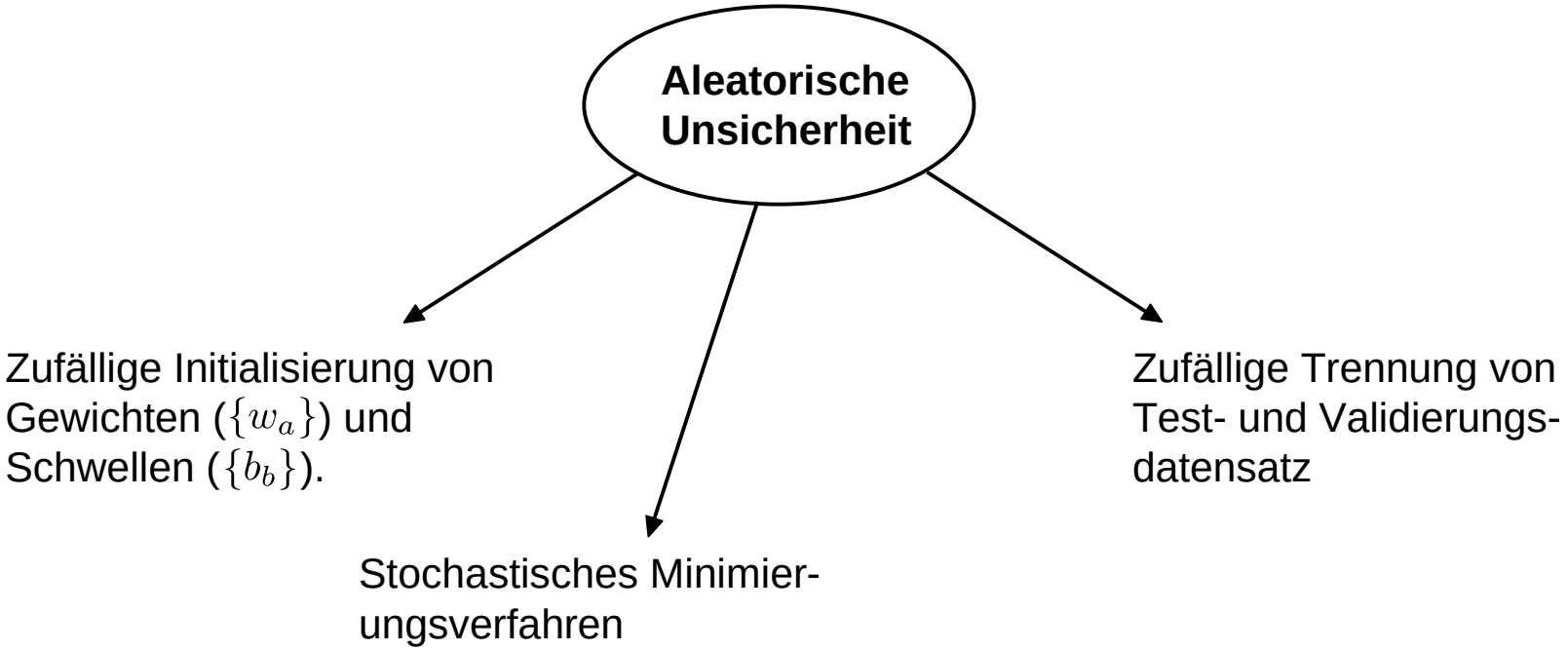
- Aleatorische Unsicherheiten wohnen dem MLP-Ansatz aufgrund seiner statistischen Natur inne:



Wo entdecken Sie statistische Prozesse bei der Verwendung von MLPs?

Aleatorische Unsicherheiten

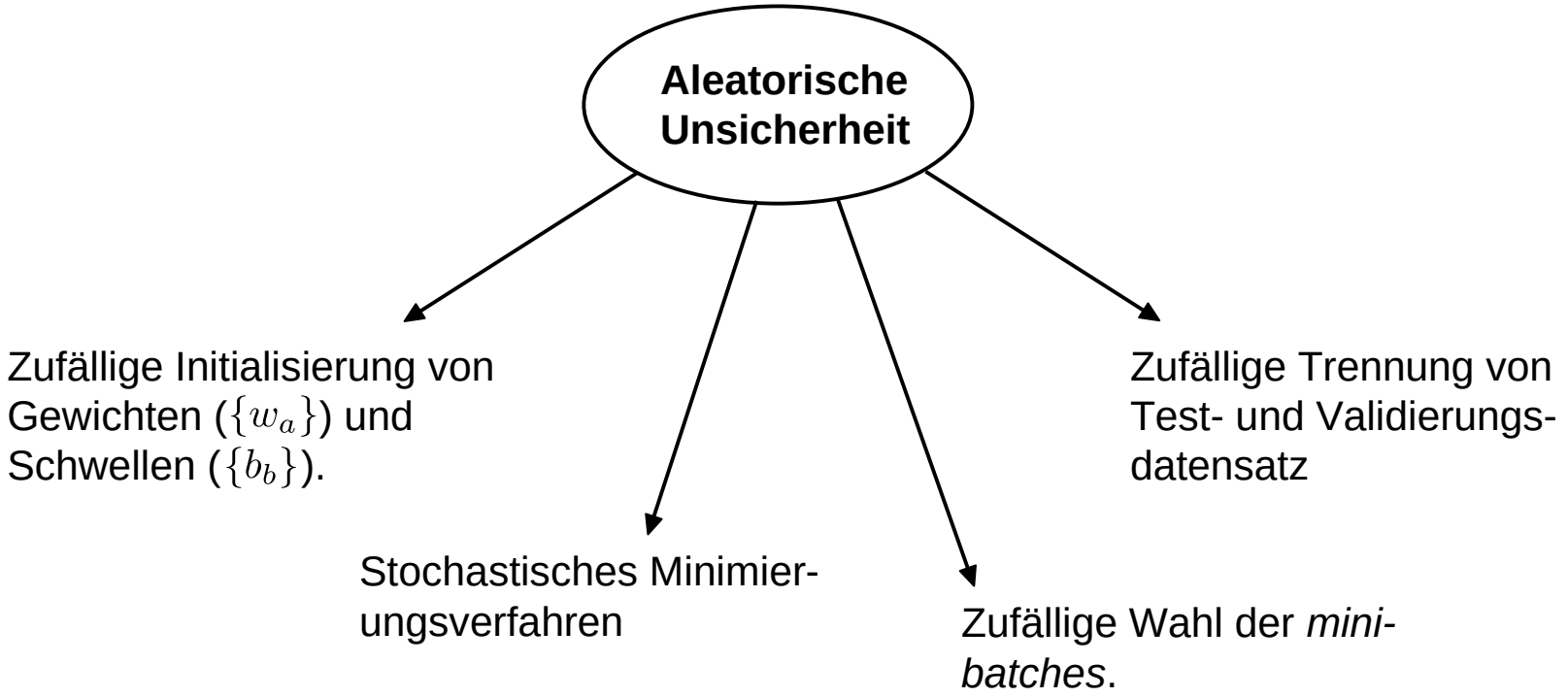
- Aleatorische Unsicherheiten wohnen dem MLP-Ansatz aufgrund seiner statistischen Natur inne:



Wo entdecken Sie statistische Prozesse bei der Verwendung von MLPs?

Aleatorische Unsicherheiten

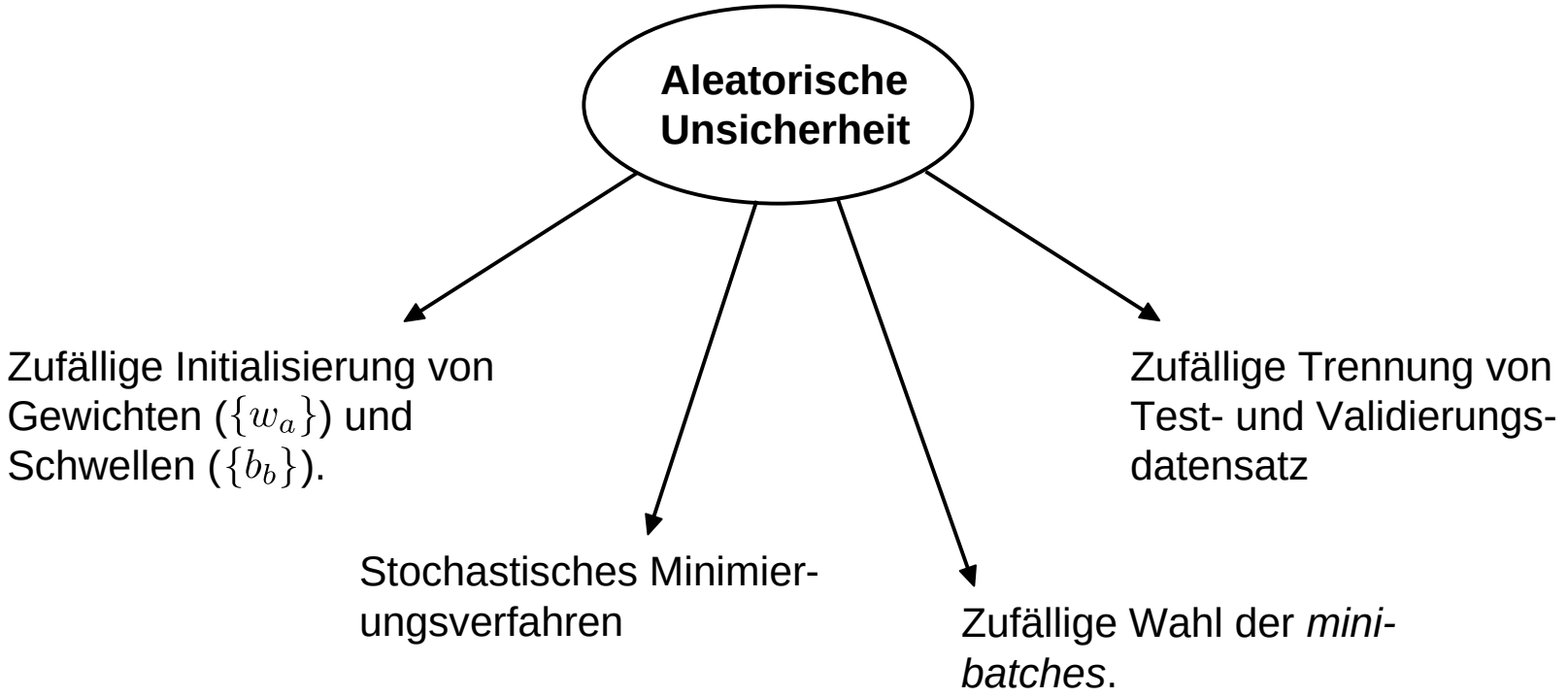
- Aleatorische Unsicherheiten wohnen dem MLP-Ansatz aufgrund seiner statistischen Natur inne:



Wo entdecken Sie statistische Prozesse bei der Verwendung von MLPs?

Aleatorische Unsicherheiten

- Aleatorische Unsicherheiten wohnen dem MLP-Ansatz aufgrund seiner statistischen Natur inne:



- Außerdem sind Trainings- und Testdatensatz selbst **nur Stichproben** aus der Grundgesamtheit.

Wo entdecken Sie statistische Prozesse bei der Verwendung von MLPs?

Abschätzung aleatorischer Unsicherheiten

- Derzeit werden drei Hauptströmungen diskutiert, wie man aleatorische Unsicherheiten abschätzen kann:

Ensemble Tests (a.k.a. Deep Ensemble):

- Initialisiere identische MLP Architektur immer wieder neu (mit variierendem *random seed*) und bestimmte die Verteilung des Ausgangs.
- Frequentistischer Ansatz offensichtlich.

Abschätzung aleatorischer Unsicherheiten

- Derzeit werden drei Hauptströmungen diskutiert, wie man aleatorische Unsicherheiten abschätzen kann:

Ensemble Tests (a.k.a. Deep Ensemble):

- Initialisiere identische MLP Architektur immer wieder neu (mit variierendem *random seed*) und bestimmte die Verteilung des Ausgangs.
- Frequentistischer Ansatz offensichtlich.

Unter Verwendung **Bayesianischer** Methoden:

Abschätzung aleatorischer Unsicherheiten

- Derzeit werden drei Hauptströmungen diskutiert, wie man aleatorische Unsicherheiten abschätzen kann:

Ensemble Tests (a.k.a. Deep Ensemble):

- Initialisiere identische MLP Architektur immer wieder neu (mit variierendem *random seed*) und bestimmte die Verteilung des Ausgangs.
- Frequentistischer Ansatz offensichtlich.

Unter Verwendung **Bayesianischer** Methoden:

Mit Hilfe von **Dropout**:

arxiv:1506.02142

Abschätzung aleatorischer Unsicherheiten

- Derzeit werden drei Hauptströmungen diskutiert, wie man aleatorische Unsicherheiten abschätzen kann:

Ensemble Tests (a.k.a. Deep Ensemble):

- Initialisiere identische MLP Architektur immer wieder neu (mit variierendem *random seed*) und bestimmte die Verteilung des Ausgangs.
- Frequentistischer Ansatz offensichtlich.

Unter Verwendung **Bayesianischer** Methoden:

- Siehe nächste Folien.

Mit Hilfe von **Dropout**:

- Siehe nächste Folien.

arxiv:1506.02142

Bayesianische Wahrscheinlichkeitsinterpretation

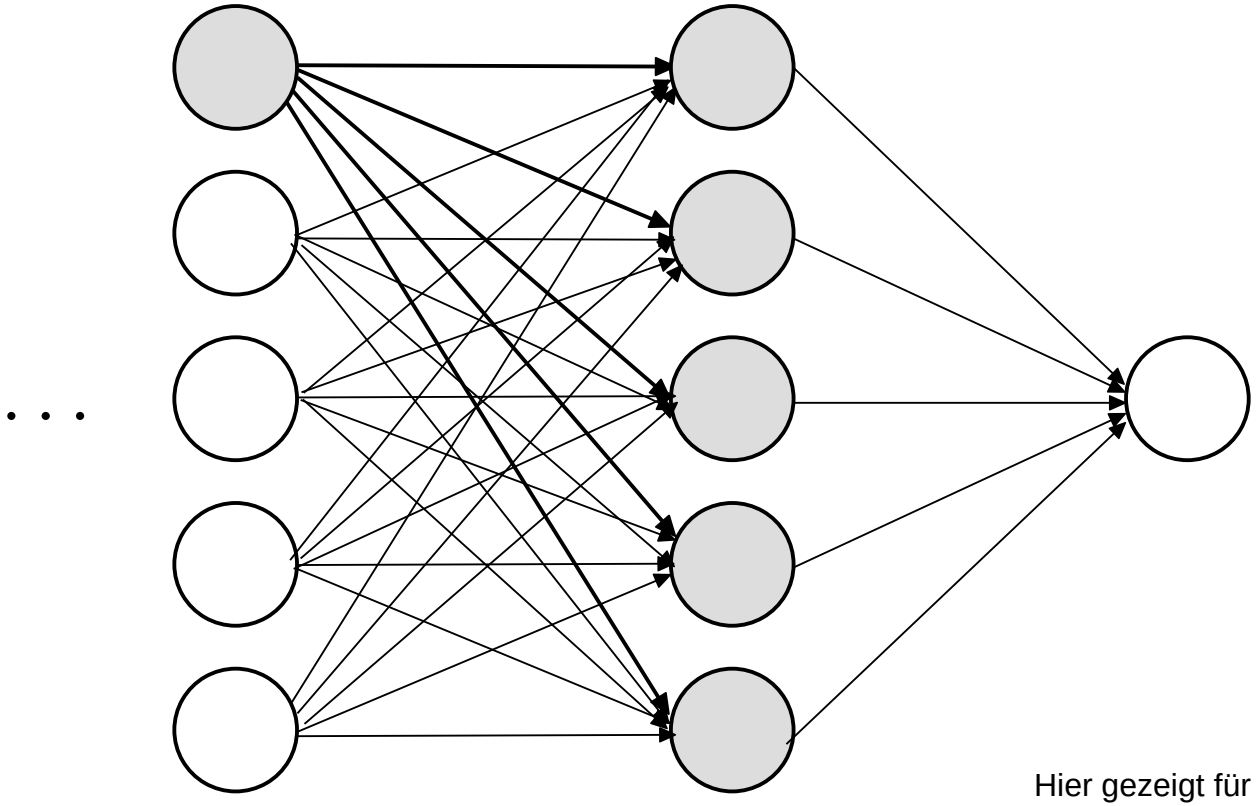
- In diesem Fall werden die Elemente des Ergebnisraums als Hypothesen bezeichnet. Der Ereignisraum $\mathfrak{P}(\Omega)$ heißt Hypothesenraum.

- Die Wahrscheinlichkeit $P(A)$ ist ein Maß für das **subjektive Fürwahrhalten** der Hypothese A .

Das Bayesianische NN

- Zu jeder Wahl eines Gewichts w_a gehört ein Konfidenzintervall als **Maß des Fürwahrhaltens**.

- Ersetze jedes Gewicht w_a im (klassichen) MLP durch eine Wahrscheinlichkeitsdichte $p(w_a|y(\{x_k\}))$:

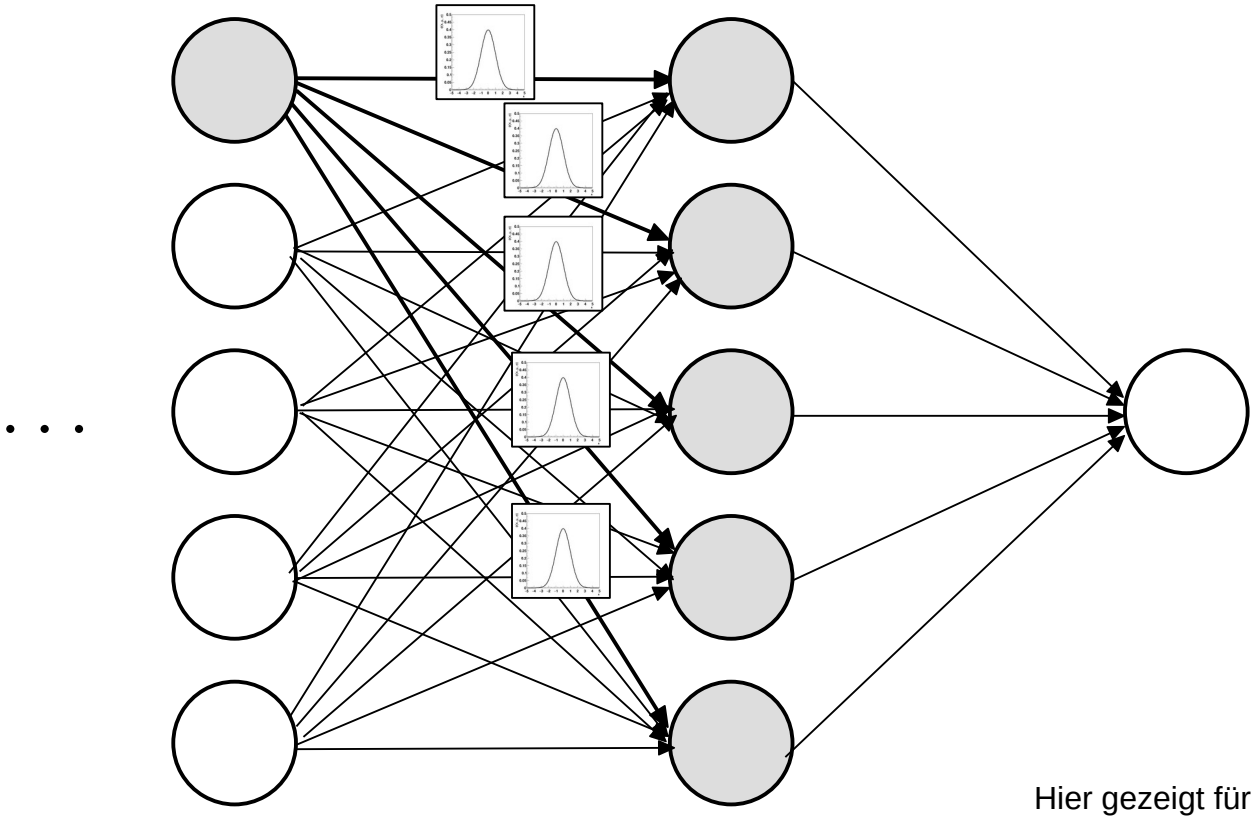


Hier gezeigt für die hervorgehobenen Verbindungen.

Das Bayesianische NN

- Zu jeder Wahl eines Gewichts w_a gehört ein Konfidenzintervall als **Maß des Fürwahrhaltens**.

- Ersetze jedes Gewicht w_a im (klassichen) MLP durch eine Wahrscheinlichkeitsdichte $p(w_a|y(\{x_k\}))$:

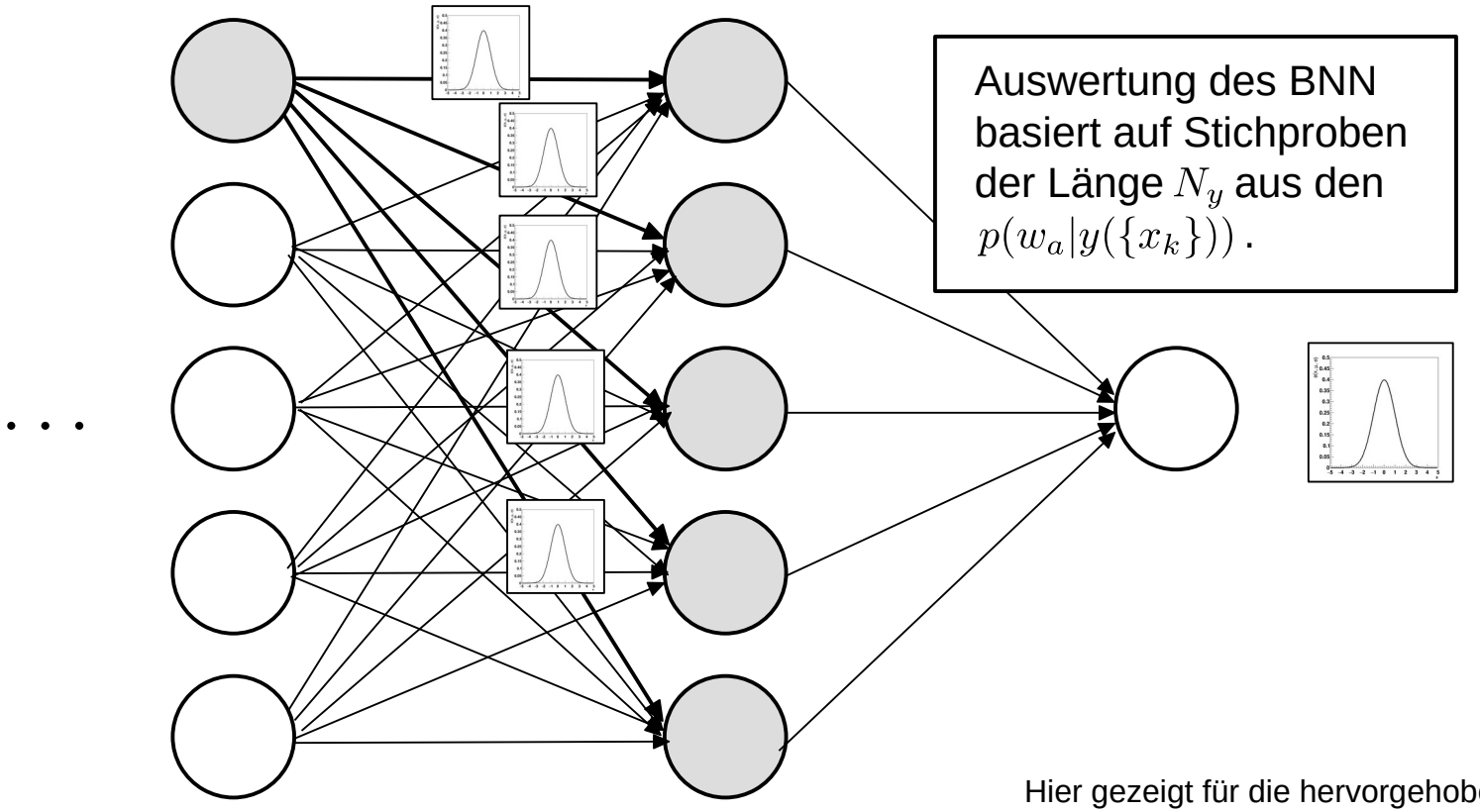


Hier gezeigt für die hervorgehobenen Verbindungen.

Das Bayesianische NN

- Zu jeder Wahl eines Gewichts w_a gehört ein Konfidenzintervall als **Maß des Fürwahrhaltens**.

- Ersetze jedes Gewicht w_a im (klassichen) MLP durch eine Wahrscheinlichkeitsdichte $p(w_a|y(\{x_k\}))$:

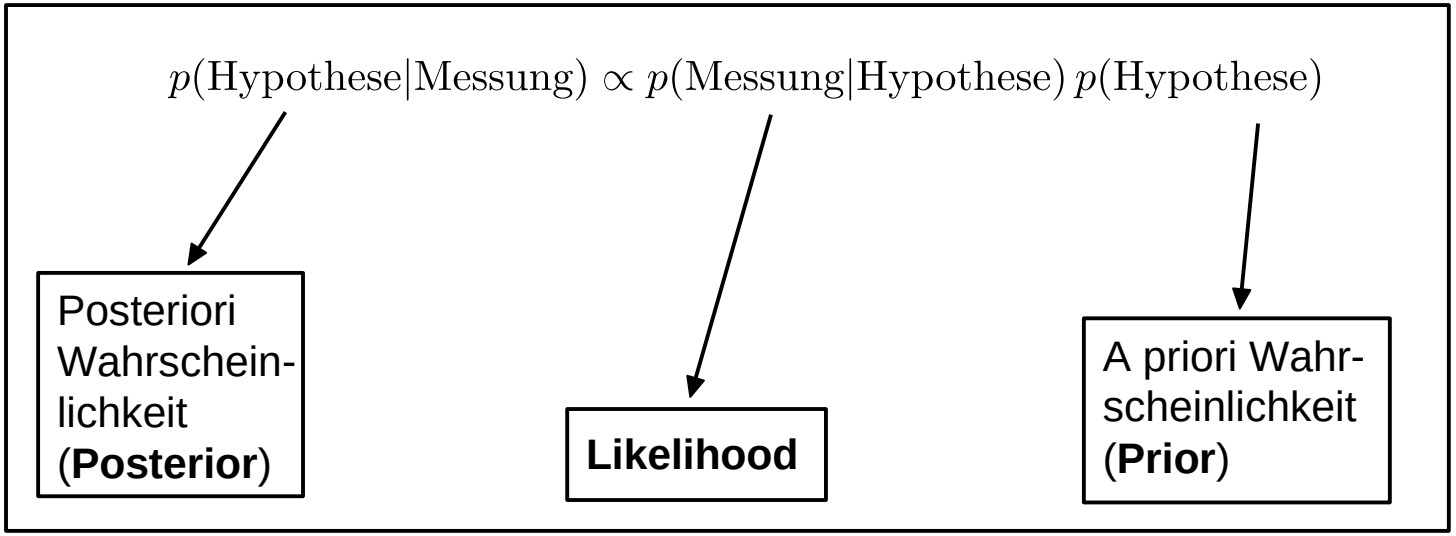


Hier gezeigt für die hervorgehobenen Verbindungen.

Bayesianische Fragestellung

- Zentrale Gesetzmäßigkeit der Bayesiansichen Wahrscheinlichkeitsinterpretation → Satz von Bayes:

Hypothese wird mit Messung abgeglichen:



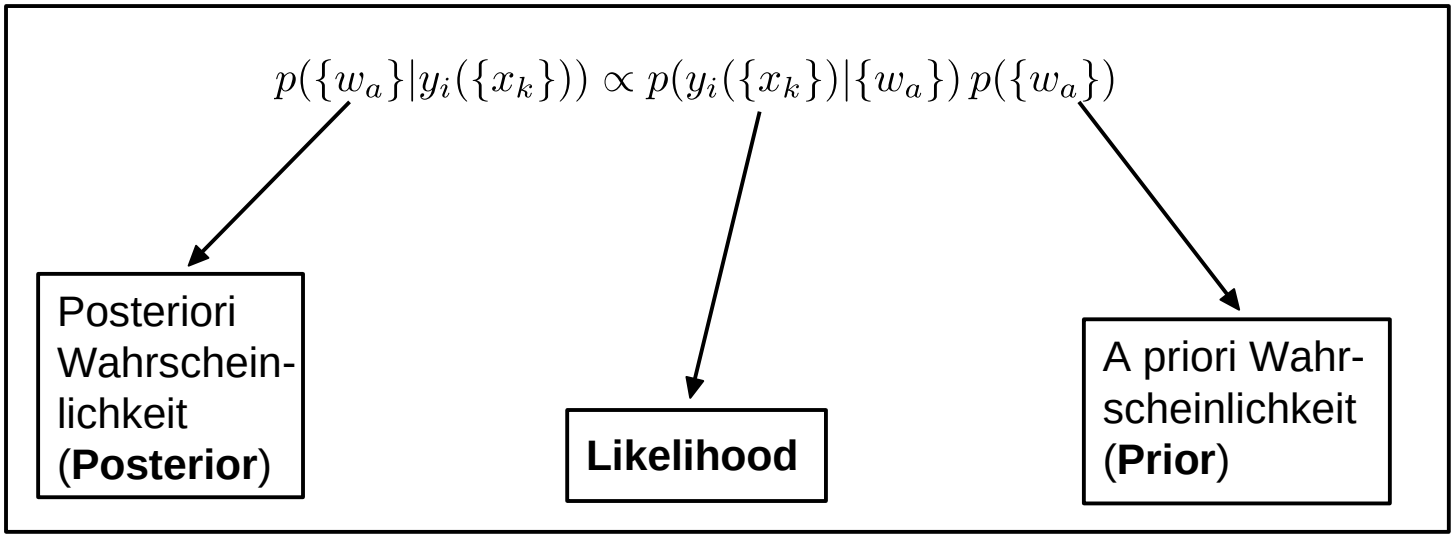
Bayesianische Fragestellung

- Zentrale Gesetzmäßigkeit der Bayesiansichen Wahrscheinlichkeitsinterpretation → Satz von Bayes:

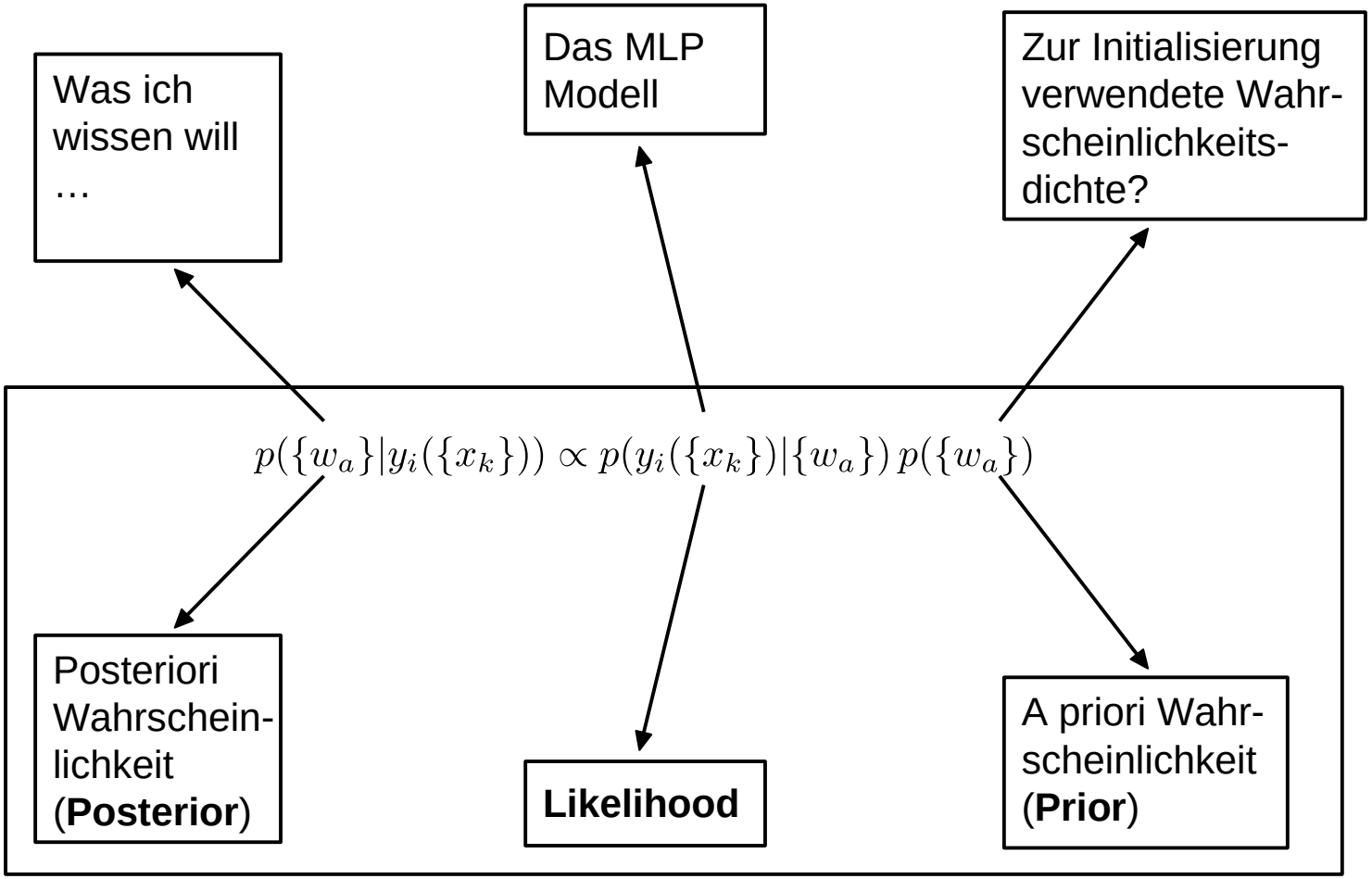
- Übersetzung des Problems eines MLP-Trainings in die **Sprache der Bayesianischen Statistik**:

  Wie sind die Gewichte $\{w_a\}$ des MLP zu wählen, um ein optimales Antwortverhalten y_i des MLP zu erhalten?

Hypothese wird mit Messung abgeglichen:



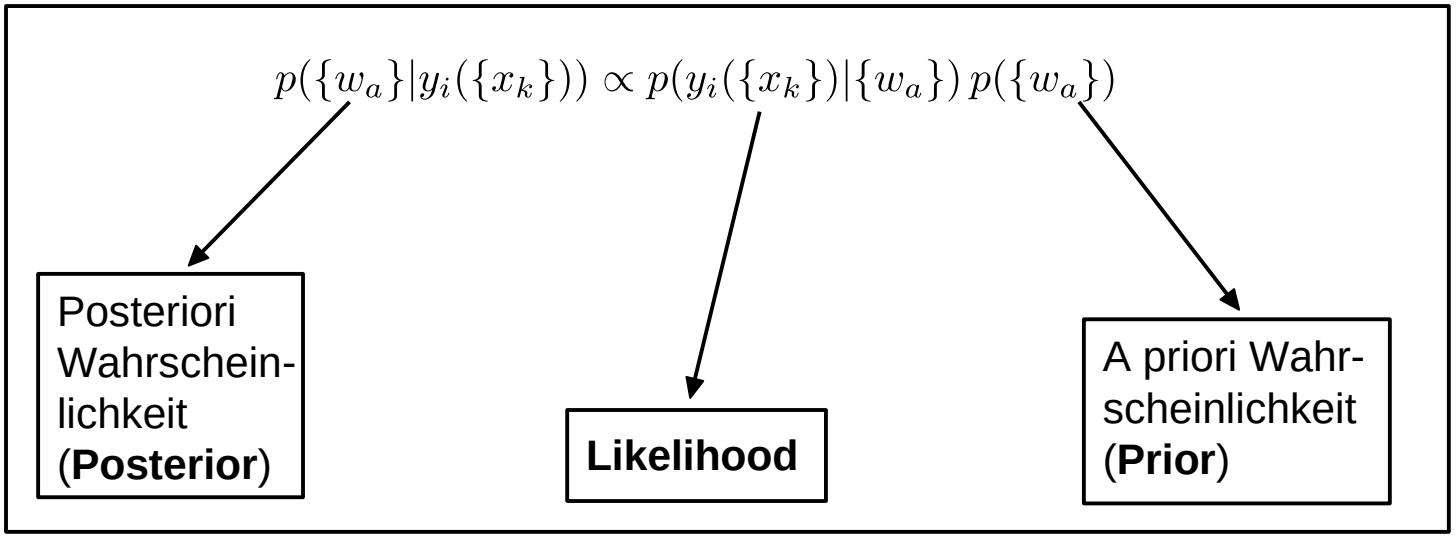
Was ist was?


Was ich wissen will …
Das MLP Modell
Zur Initialisierung verwendete Wahrscheinlichkeitsdichte?
$p(\{w_a\}|y_i(\{x_k\})) \propto p(y_i(\{x_k\})|\{w_a\})\, p(\{w_a\})$
Posteriori Wahrscheinlichkeit (**Posterior**)
Likelihood
A priori Wahrscheinlichkeit (**Prior**)

Wie kann ich $p(\{w_a\}|y_i(\{x_k\}))$ bestimmen?

- Für Probleme mit geringer Dimension ließe sich $p(\{w_a\}|y_i(\{x_k\}))$ noch durch Integration der Wahrscheinlichkeitsdichten $p(y_i(\{x_k\})|\{w_a\})$ und $p(\{w_a\})$, z.B. mit Hilfe einer MC Methode bestimmen.

- Für ein MLP mit vielen 1000 Gewichten und Schwellen ist dies jedoch **praktisch unmöglich**.



Variational inference

- Man versucht daher eine handhabbare Approximation $q_\theta(\{w_a\})$ mit Parametern $\{\theta_j\}$ zu „erraten“, die sich so gut wie möglich an $p(\{w_a\}|y_i(\{x_k\}))$ anpassen lässt.

- Den Vorgang dieser Anpassung bezeichnet man als *variational inference*. Er **entspricht dem klassischen MLP-Training**.

Variational inference

- Man versucht daher eine handhabbare Approximation $q_\theta(\{w_a\})$ mit Parametern $\{\theta_j\}$ zu „erraten“, die sich so gut wie möglich an $p(\{w_a\}|y_i(\{x_k\}))$ anpassen lässt.

- Den Vorgang dieser Anpassung bezeichnet man als *variational inference*. Er **entspricht dem klassischen MLP-Training**.

- Für die weitere Diskussion führen wir die folgende Vereinfachung der Notation ein:

$$\begin{aligned} \boldsymbol{w} &\equiv \{w_a\} && : \text{Gewichte} \\ D &\equiv y_i(\{x_k\})) && : \text{Beobachtete MLP-Antwort (Daten)} \end{aligned}$$

Vergleichbarkeit von Verteilungen

- Für die *variational inference* müssen wir Verteilungen vergleichen.
- Hierfür ist prinzipiell jede sinnvolle und zielführende Metrik verwendbar.
- Es wird sich jedoch gleich zeigen, das die Wahl der Kullback-Leibler Divergenz (auch KL-Divergenz oder *information gain*) eine sehr intuitive Korrespondenz zum klassischen NN-Training erlaubt:

$$KL[q_\theta(\boldsymbol{w}), p(\boldsymbol{w}|D)] = \int q_\theta(\boldsymbol{w}) \ln \left(\frac{q_\theta(\boldsymbol{w})}{p(\boldsymbol{w}|D)} \right) \mathrm{d}\boldsymbol{w}$$

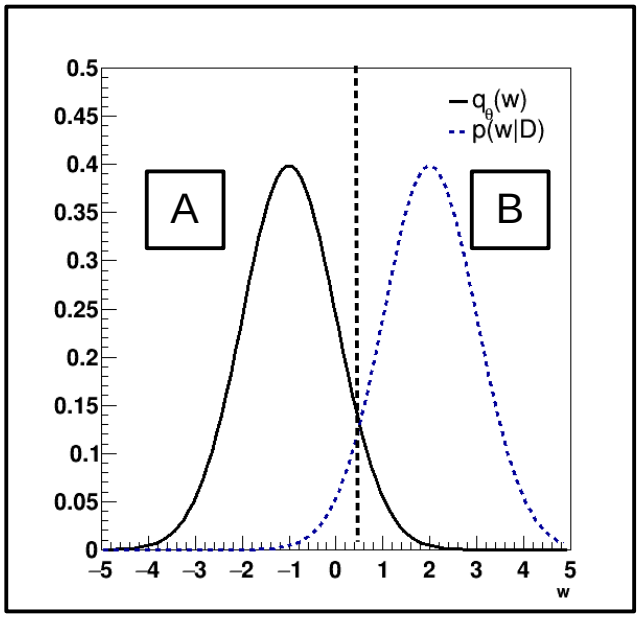
- Die KL-Divergenz **entspricht der Verlustfunktion** im klassischen MLP-Training.

KL-Divergenz

- Die KL-Divergenz ist immer ≥ 0 – Warum?

$$KL[q_\theta(\boldsymbol{w}), p(\boldsymbol{w}|D)] = \int q_\theta(\boldsymbol{w}) \underbrace{\ln\left(\frac{q_\theta(\boldsymbol{w})}{p(\boldsymbol{w}|D)}\right)}_{\substack{\text{kann Werte} > 1 \text{ oder} \\ < 1 \text{ annehmen}}} \mathrm{d}\boldsymbol{w}$$

- Anschauliches Argument:



A

$$q_\theta(\boldsymbol{w}) > p(\boldsymbol{w}|D)$$

$$\ln\left(\frac{q_\theta(\boldsymbol{w})}{p(\boldsymbol{w}|D)}\right) > 0$$

B

$$q_\theta(\boldsymbol{w}) < p(\boldsymbol{w}|D)$$

$$\ln\left(\frac{q_\theta(\boldsymbol{w})}{p(\boldsymbol{w}|D)}\right) < 0$$

Wertebereich aus A trägt mit höherem Gewicht zum Integral bei, als der Wertebereich in B.

Umformungen der KL-Divergenz

KL-Divergenz:

$$KL[q_\theta(\boldsymbol{w}), p(\boldsymbol{w}|D)] = \int q_\theta(\boldsymbol{w}) \ln\left(\frac{q_\theta(\boldsymbol{w})}{p(\boldsymbol{w}|D)}\right) \mathrm{d}\boldsymbol{w}$$

Satz von Bayes:

$$p(\boldsymbol{w}|D) = \frac{p(D|\boldsymbol{w})\, p(\boldsymbol{w})}{p(D)}$$

$$\begin{aligned} KL[q_\theta(\boldsymbol{w}), p(\boldsymbol{w}|D)] &= \int q_\theta(\boldsymbol{w}) \ln\left(\frac{q_\theta(\boldsymbol{w})\, p(D)}{p(D|\boldsymbol{w})\, p(\boldsymbol{w})}\right) \mathrm{d}\boldsymbol{w} \\ &= \underbrace{-\int q_\theta(\boldsymbol{w}) \ln\left(p(D|\boldsymbol{w})\right) \mathrm{d}\boldsymbol{w}}_{\equiv \text{ Kreuzentropie}} + \underbrace{\int q_\theta(\boldsymbol{w}) \ln\left(\frac{q_\theta(\boldsymbol{w})}{p(\boldsymbol{w})}\right) \mathrm{d}\boldsymbol{w}}_{\equiv KL[q_\theta(w), p(\boldsymbol{w})]} + \underbrace{\ln\left(p(D)\right)}_{= \text{const.}} \end{aligned}$$

NB: Für kontinuierliche Größen

Regularisierung

Umformungen der KL-Divergenz

KL-Divergenz:

$$KL[q_\theta(\boldsymbol{w}), p(\boldsymbol{w}|D)] = \int q_\theta(\boldsymbol{w}) \ln\left(\frac{q_\theta(\boldsymbol{w})}{p(\boldsymbol{w}|D)}\right) \mathrm{d}\boldsymbol{w}$$

Satz von Bayes:

$$p(\boldsymbol{w}|D) = \frac{p(D|\boldsymbol{w})\, p(\boldsymbol{w})}{p(D)}$$

$$KL[q_\theta(\boldsymbol{w}), p(\boldsymbol{w}|D)] = \int q_\theta(\boldsymbol{w}) \ln\left(\frac{q_\theta(\boldsymbol{w})\, p(D)}{p(D|\boldsymbol{w})\, p(\boldsymbol{w})}\right) \mathrm{d}\boldsymbol{w}$$

$$= \underbrace{-\int q_\theta(\boldsymbol{w}) \ln\left(p(D|\boldsymbol{w})\right) \mathrm{d}\boldsymbol{w}}_{\equiv \text{ Kreuzentropie}} + \underbrace{\int q_\theta(\boldsymbol{w}) \ln\left(\frac{q_\theta(\boldsymbol{w})}{p(\boldsymbol{w})}\right) \mathrm{d}\boldsymbol{w}}_{\equiv KL[q_\theta(w), p(\boldsymbol{w})]} + \underbrace{\ln\left(p(D)\right)}_{= \text{const.}}$$

Regularisierung

NB: Für kontinuierliche Größen

Näherung für den posterior

Umformungen der KL-Divergenz

KL-Divergenz:

$$KL[q_\theta(\boldsymbol{w}), p(\boldsymbol{w}|D)] = \int q_\theta(\boldsymbol{w}) \ln\left(\frac{q_\theta(\boldsymbol{w})}{p(\boldsymbol{w}|D)}\right) \mathrm{d}\boldsymbol{w}$$

Satz von Bayes:

$$p(\boldsymbol{w}|D) = \frac{p(D|\boldsymbol{w})\, p(\boldsymbol{w})}{p(D)}$$

$$KL[q_\theta(\boldsymbol{w}), p(\boldsymbol{w}|D)] = \int q_\theta(\boldsymbol{w}) \ln\left(\frac{q_\theta(\boldsymbol{w})\, p(D)}{p(D|\boldsymbol{w})\, p(\boldsymbol{w})}\right) \mathrm{d}\boldsymbol{w}$$

$$= \underbrace{-\int q_\theta(\boldsymbol{w}) \ln\left(p(D|\boldsymbol{w})\right) \mathrm{d}\boldsymbol{w}}_{\equiv \text{ Kreuzentropie}} + \underbrace{\int q_\theta(\boldsymbol{w}) \ln\left(\frac{q_\theta(\boldsymbol{w})}{p(\boldsymbol{w})}\right) \mathrm{d}\boldsymbol{w}}_{\equiv KL[q_\theta(w), p(\boldsymbol{w})]} + \underbrace{\ln\left(p(D)\right)}_{= \text{const.}}$$

Regularisierung

NB: Für kontinuierliche Größen

Näherung für den posterior

NN output

Umformungen der KL-Divergenz

KL-Divergenz:

$$KL[q_\theta(\boldsymbol{w}), p(\boldsymbol{w}|D)] = \int q_\theta(\boldsymbol{w}) \ln\left(\frac{q_\theta(\boldsymbol{w})}{p(\boldsymbol{w}|D)}\right) \mathrm{d}\boldsymbol{w}$$

Satz von Bayes:

$$p(\boldsymbol{w}|D) = \frac{p(D|\boldsymbol{w})\, p(\boldsymbol{w})}{p(D)}$$

$$KL[q_\theta(\boldsymbol{w}), p(\boldsymbol{w}|D)] = \int q_\theta(\boldsymbol{w}) \ln\left(\frac{q_\theta(\boldsymbol{w})\, p(D)}{p(D|\boldsymbol{w})\, p(\boldsymbol{w})}\right) \mathrm{d}\boldsymbol{w}$$

$$= \underbrace{-\int q_\theta(\boldsymbol{w}) \ln\left(p(D|\boldsymbol{w})\right) \mathrm{d}\boldsymbol{w}}_{\equiv \text{ Kreuzentropie}} + \underbrace{\int q_\theta(\boldsymbol{w}) \ln\left(\frac{q_\theta(\boldsymbol{w})}{p(\boldsymbol{w})}\right) \mathrm{d}\boldsymbol{w}}_{\equiv KL[q_\theta(w), p(\boldsymbol{w})]} + \underbrace{\ln\left(p(D)\right)}_{= \text{const.}}$$

Regularisierung

NB: Für kontinuierliche Größen

Näherung für den posterior

NN output

Stellt sicher, dass sich $q_\theta(\boldsymbol{w})$ nicht allzuweit von $p(\boldsymbol{w})$ entfernt

Umformungen der KL-Divergenz

KL-Divergenz:

$$KL[q_\theta(\boldsymbol{w}), p(\boldsymbol{w}|D)] = \int q_\theta(\boldsymbol{w}) \ln\left(\frac{q_\theta(\boldsymbol{w})}{p(\boldsymbol{w}|D)}\right) \mathrm{d}\boldsymbol{w}$$

Satz von Bayes:

$$p(\boldsymbol{w}|D) = \frac{p(D|\boldsymbol{w})\, p(\boldsymbol{w})}{p(D)}$$

Sind $q_\theta(\boldsymbol{w})$ und $p(\boldsymbol{w})$ Produkte aus Normalverteilungen, dann entspricht $KL[q_\theta(\boldsymbol{w}), p(\boldsymbol{w})]$ einer mit $q_\theta(\boldsymbol{w})$ gewichteten L2-Regularisierung.

$$KL[q_\theta(\boldsymbol{w}), p(\boldsymbol{w}|D)] = \int q_\theta(\boldsymbol{w}) \ln\left(\frac{q_\theta(\boldsymbol{w})\, p(D)}{p(D|\boldsymbol{w})\, p(\boldsymbol{w})}\right) \mathrm{d}\boldsymbol{w}$$

$$= \underbrace{-\int q_\theta(\boldsymbol{w}) \ln\left(p(D|\boldsymbol{w})\right) \mathrm{d}\boldsymbol{w}}_{\equiv \text{ Kreuzentropie}} + \underbrace{\int q_\theta(\boldsymbol{w}) \ln\left(\frac{q_\theta(\boldsymbol{w})}{p(\boldsymbol{w})}\right) \mathrm{d}\boldsymbol{w}}_{\equiv KL[q_\theta(w), p(\boldsymbol{w})]} + \underbrace{\ln\left(p(D)\right)}_{= \text{const.}}$$

Regularisierung

NB: Für kontinuierliche Größen

Näherung für den posterior

NN output

Stellt sicher, dass sich $q_\theta(\boldsymbol{w})$ nicht allzuweit von $p(\boldsymbol{w})$ entfernt

Anmerkungen zum Dropout

- Wendet man Dropout nicht (nur) zum Training, sondern bei der **Anwendung auf dem Test-datensatz** an führt dies ebenfalls zu einer Verteilung des NN Antwortverhaltens.

- Diese Verteilung ist eine Näherung zur Abschätzung der Unsicherheit eines BNN. – Warum?

Anmerkungen zum Dropout

- Wendet man Dropout nicht (nur) zum Training, sondern bei der **Anwendung auf dem Test-datensatz** an führt dies ebenfalls zu einer Verteilung des NN Antwortverhaltens.

- Diese Verteilung ist eine Näherung zur Abschätzung der Unsicherheit eines BNN. – Warum?

  - Die Knoten des MLP entsprechen einer Reihenentwicklung der zu approximierenden Kontour/Funktion im Funktionenraum.

Anmerkungen zum Dropout

- Wendet man Dropout nicht (nur) zum Training, sondern bei der **Anwendung auf dem Test-datensatz** an führt dies ebenfalls zu einer Verteilung des NN Antwortverhaltens.

- Diese Verteilung ist eine Näherung zur Abschätzung der Unsicherheit eines BNN. – Warum?

  - Die Knoten des MLP entsprechen einer Reihenentwicklung der zu approximierenden Kontour/Funktion im Funktionenraum.

  - Bei der Auswertung zufällig Knoten aus dem MLP zu entfernen ist äquivalent zur Approximation von $p(\boldsymbol{w}|D)$ durch eine Funktion $q_\theta(\boldsymbol{w})$ die mit Hilfe eines Gauss‘schen Prozesses erzeugt wurde.

Anmerkungen zu Unsicherheiten von MLPs

- Es zeigt sich das sowohl BNNs als auch Dropout Probleme mit der Abdeckung (*coverage*) der abgeleiteten Konfidenzintervalle aufweisen.

Anmerkungen zu Unsicherheiten von MLPs

- Es zeigt sich das sowohl BNNs als auch Dropout Probleme mit der Abdeckung (*coverage*) der abgeleiteten Konfidenzintervalle aufweisen.

- Wenn es um die Abschätzung von Unsicherheiten eines MLP geht ist meiner Meinung nach das Verfahren, das Abdeckung am sichersten gewährleistet der (frequentistische) **Ensemble-test**.

Anmerkungen zu Unsicherheiten von MLPs

- Es zeigt sich das sowohl BNNs als auch Dropout Probleme mit der Abdeckung (*coverage*) der abgeleiteten Konfidenzintervalle aufweisen.

- Wenn es um die Abschätzung von Unsicherheiten eines MLP geht ist meiner Meinung nach das Verfahren, das Abdeckung am sichersten gewährleistet der (frequentistische) **Ensemble-test**.

- Glücklicherweise ist die Frage nach der (intrinsichen) Unsicherheit eines MLP im konkreten Fall einer physikalischen Messung meistens nicht von Relevanz. – Warum?

Anmerkungen zu Unsicherheiten von MLPs

- Es zeigt sich das sowohl BNNs als auch Dropout Probleme mit der Abdeckung (*coverage*) der abgeleiteten Konfidenzintervalle aufweisen.

- Wenn es um die Abschätzung von Unsicherheiten eines MLP geht ist meiner Meinung nach das Verfahren, das Abdeckung am sichersten gewährleistet der (frequentistische) **Ensemble-test**.

- Glücklicherweise ist die Frage nach der (intrinsichen) Unsicherheit eines MLP im konkreten Fall einer physikalischen Messung meistens nicht von Relevanz. – Warum?

  - Nach dem Training sind alle Gewichte und Schwellen wohldefiniert und fest vorgegeben.

Anmerkungen zu Unsicherheiten von MLPs

- Es zeigt sich das sowohl BNNs als auch Dropout Probleme mit der Abdeckung (*coverage*) der abgeleiteten Konfidenzintervalle aufweisen.

- Wenn es um die Abschätzung von Unsicherheiten eines MLP geht ist meiner Meinung nach das Verfahren, das Abdeckung am sichersten gewährleistet der (frequentistische) **Ensemble-test**.

- Glücklicherweise ist die Frage nach der (intrinsichen) Unsicherheit eines MLP im konkreten Fall einer physikalischen Messung meistens nicht von Relevanz. – Warum?

  - Nach dem Training sind alle Gewichte und Schwellen wohldefiniert und fest vorgegeben.

  - Das MLP ist in diesem Sinne eine deterministische high-level Observable, die für identische Eingabewerte $\{x_k\}$ immer die gleiche Antwort zurück gibt. .

Anmerkungen zu Unsicherheiten von MLPs

- Es zeigt sich das sowohl BNNs als auch Dropout Probleme mit der Abdeckung (*coverage*) der abgeleiteten Konfidenzintervalle aufweisen.

- Wenn es um die Abschätzung von Unsicherheiten eines MLP geht ist meiner Meinung nach das Verfahren, das Abdeckung am sichersten gewährleistet der (frequentistische) **Ensemble-test**.

- Glücklicherweise ist die Frage nach der (intrinsichen) Unsicherheit eines MLP im konkreten Fall einer physikalischen Messung meistens nicht von Relevanz. – Warum?

  - Nach dem Training sind alle Gewichte und Schwellen wohldefiniert und fest vorgegeben.

  - Das MLP ist in diesem Sinne eine deterministische high-level Observable, die für identische Eingabewerte $\{x_k\}$ immer die gleiche Antwort zurück gibt. .

  - Wenn Sie niemals den Fehler begehen, die trainierten Gewichte und Schwellen ihres MLP zu löschen ist die Frage nach einer intrinsichen Unsicherheit des MLP irrelevant, weil Sie nicht irgendein MLP verwenden sondern ein sehr konkretes. Es gibt ja auch nur einen LHC.

Anmerkungen zu Unsicherheiten von MLPs

- Es zeigt sich das sowohl BNNs als auch Dropout Probleme mit der Abdeckung (*coverage*) der abgeleiteten Konfidenzintervalle aufweisen.

- Wenn es um die Abschätzung von Unsicherheiten eines MLP geht ist meiner Meinung nach das Verfahren, das Abdeckung am sichersten gewährleistet der (frequentistische) **Ensemble-test**.

- Glücklicherweise ist die Frage nach der (intrinsichen) Unsicherheit eines MLP im konkreten Fall einer physikalischen Messung meistens nicht von Relevanz. – Warum?

  - Nach dem Training sind alle Gewichte und Schwellen wohldefiniert und fest vorgegeben.

  - Das MLP ist in diesem Sinne eine deterministische high-level Observable, die für identische Eingabewerte $\{x_k\}$ immer die gleiche Antwort zurück gibt. .

  - Wenn Sie niemals den Fehler begehen, die trainierten Gewichte und Schwellen ihres MLP zu löschen ist die Frage nach einer intrinsichen Unsicherheit des MLP irrelevant, weil Sie nicht irgendein MLP verwenden sondern ein sehr konkretes. Es gibt ja auch nur einen LHC.

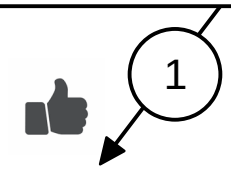
  - Wie so oft in der Statistik hängt die Berechnung der Unsicherheiten von der konkreten Fragestellung ab!

NN scrutiny

What additional uncertainties have to be taken into account due to the presence of the NN, to trust our measurement?

NN scrutiny

What additional uncertainties have to be taken into account due to the presence of the NN, to trust our measurement?



After training NN acts like a deterministic high-level variable
→ no additional uncert. required.

NN scrutiny

What additional uncertainties have to be taken into account due to the presence of the NN, to trust our measurement?

1

After training NN acts like a deterministic high-level variable
→ no additional uncert. required.

2

Intrinsic (stat.) uncertainties of NN training only of importance for reproducibility of training.

NN scrutiny

What additional uncertainties have to be taken into account due to the presence of the NN, to trust our measurement?

1

After training NN acts like a deterministic high-level variable
→ no additional uncert. required.

2

Intrinsic (stat.) uncertainties of NN training only of importance for reproducibility of training.

3

NN exploits input variable space much deeper than e.g. cut-based
→ thorough control of input space.

Unboxing the NN

Comput. Softw. Big Sci. 2 (2018) 5

- Exploit Taylor expansion of NN output function y_i in $\{x_j\}$ with fixed $\{w_a\}$ and $\{b_b\}$ after training to identify x_j's w/ largest influence on y_i :

$$\langle t_\alpha \rangle = \frac{1}{N} \sum_{k=1}^{N} \left| t_\alpha \left(\left\{ x_j^{(k)} \right\} \right) \right|$$

N : Sample size

t_α : Taylor coefficient labeled by α

- $\langle t_\alpha \rangle$ is the arithmetic mean of $|t_\alpha|$, evaluated on the whole input space that is sampled by the test data set.

- Introduce nomenclature of *generalized features* of the input feature space:

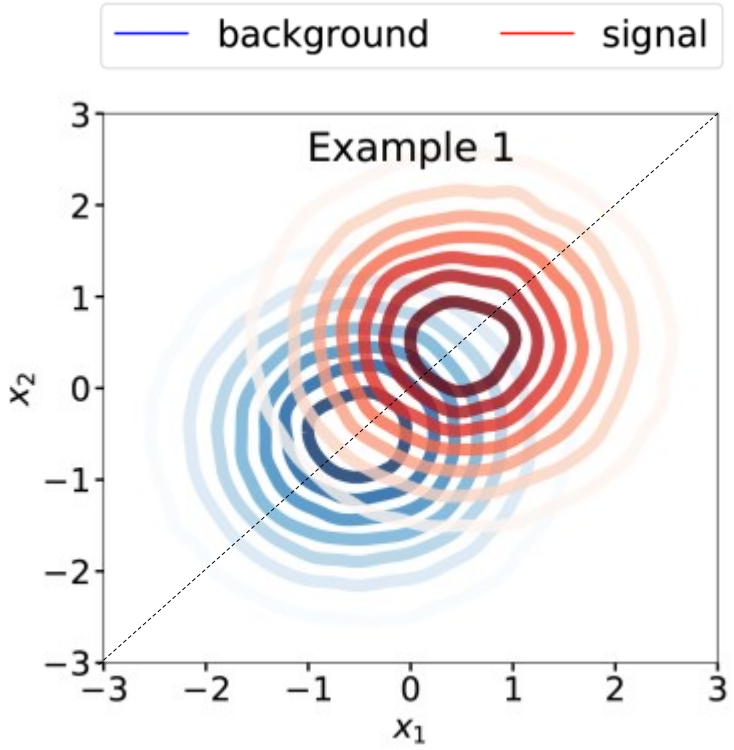
$\alpha = x_1, x_2, \ldots$	1. order feature of input space (~ 1. order derivative)
$\alpha = x_1x_1, x_1x_2, \ldots$	2. order feature of input space (~ 2. order derivative)
$\alpha = x_1x_1x_1, x_1x_1x_2, \ldots$	3. order feature of input space (~ 3. order derivative)
$\vdots$	$\vdots$

Meaning of *generalized features*

- 1. order feature: $\alpha = x_1, x_2, \ldots$
  - Physical location of feature/marginal distributions, e.g. signal @ small x_1 background @ large x_1.
- 2. order feature: $\alpha = x_1 x_1, x_1 x_2, \ldots$
  - Linear correlations across two features ($x_i x_j$).
  - “Self-correlations” ($x_i x_i$), i.e. curvature of NN output function.

Simplistic task

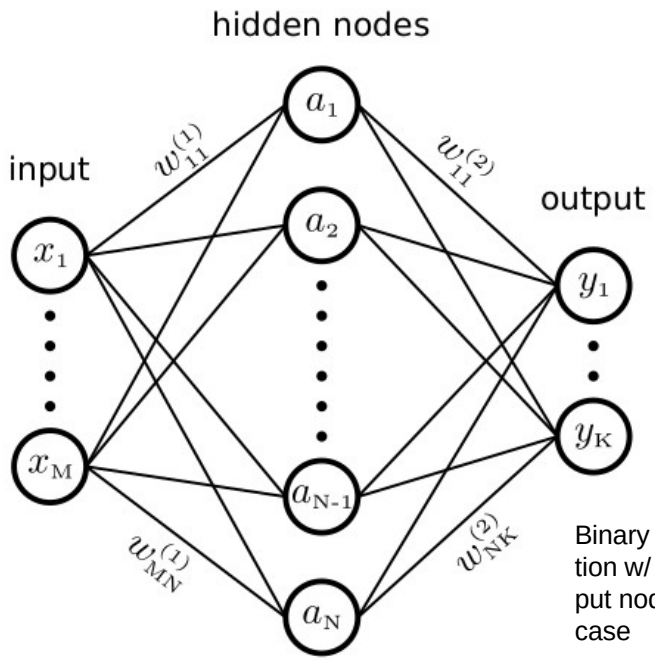
- Illustration with a simplistic task in 2d:



- Two input features x_1 and x_2 .
- Binary classification.
- Signal ($\mu_{\mathrm{S}} = (0.5,\, 0.5)$) and background ($\mu_{\mathrm{BG}} = (-0.5,\, -0.5)$) sampled from normal distributions w/ different means, but otherwise identical.
- Note the symmetry of the task.

Simple NN

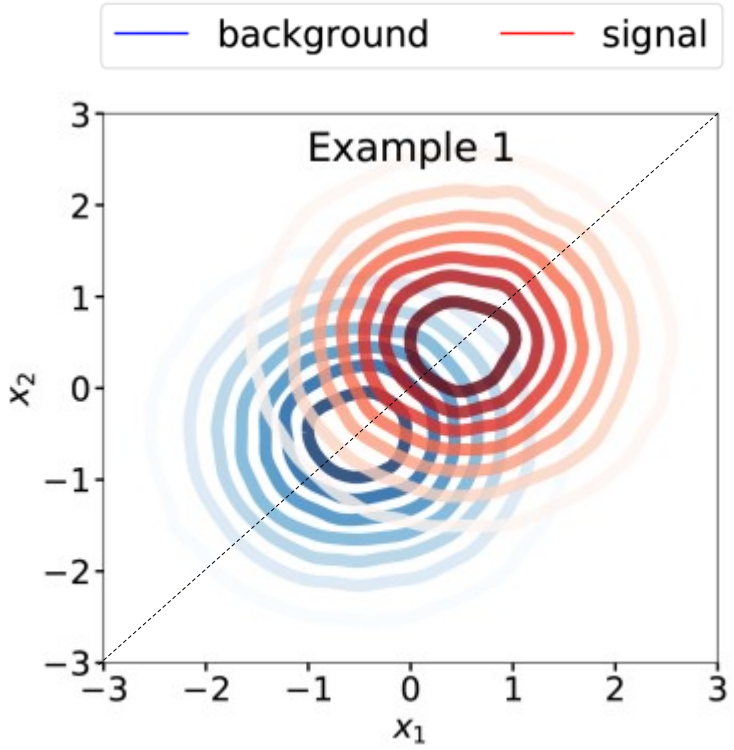
- NN architecture and configuration:



Binary classification w/ single output node in this case

- One hidden layer with 100 nodes.
- Activation functions: $\tanh(\cdot)$ (hidden layer), sigmoid (output layer).
- Loss function: cross entropy.
- Minimization: Adam (10^{-4}).
- Mini-batch training with early stopping after 30 epochs (for what will be shown in the following).

Taylor coefficient analysis…



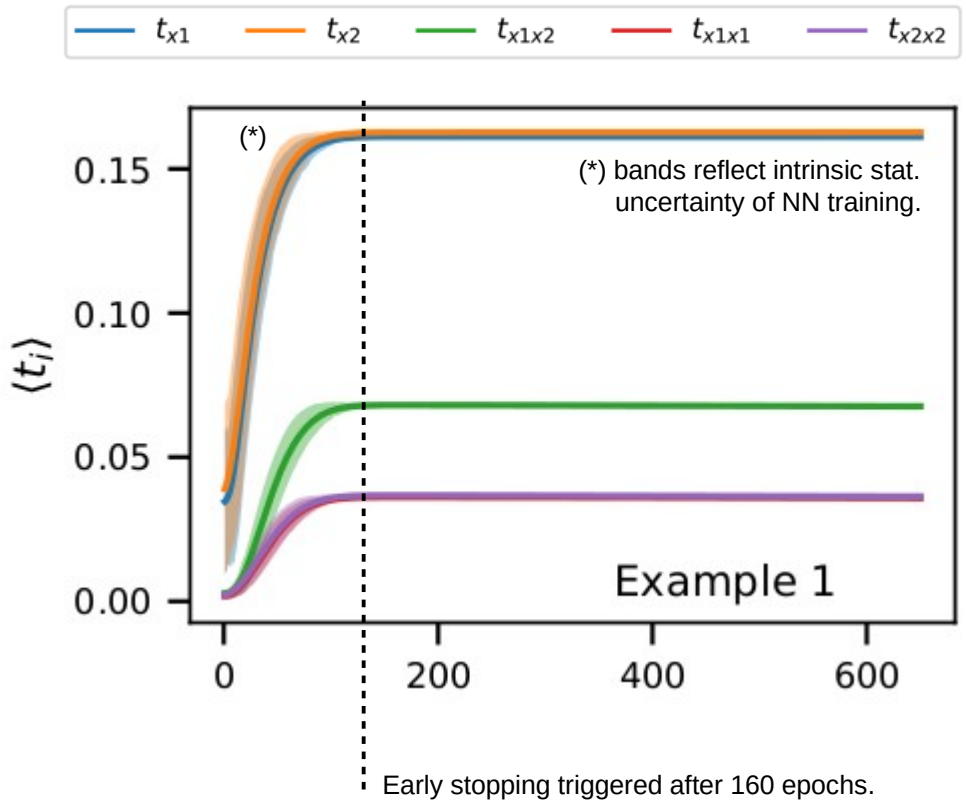
Result after training:

α	Value after training
x_1	0.16
x_2	0.16
x_1x_2	0.07
x_1x_1	0.04
x_2x_2	0.04

- x_1 and x_2 found to be most influential (distinction of S and BG by location).
- x_ix_j indicate that correlation plays a role (for S and BG x_1 and x_2 are linearly correlated).

… as a function of time…

- Evaluation after each training epoch:



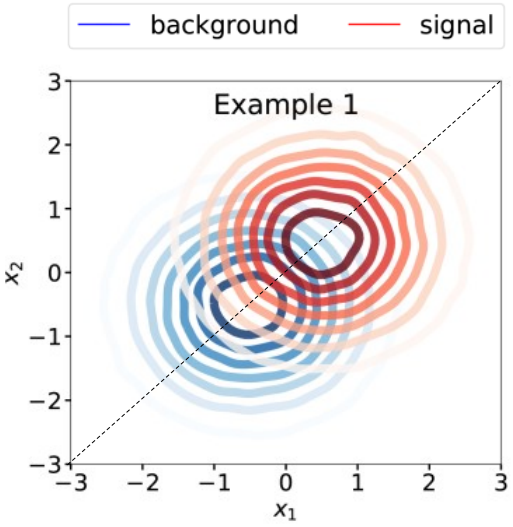
- Allows monitoring of training process.

Result after training:

α	Value after training
x_1	0.16
x_2	0.16
x_1x_2	0.07
x_1x_1	0.04
x_2x_2	0.04

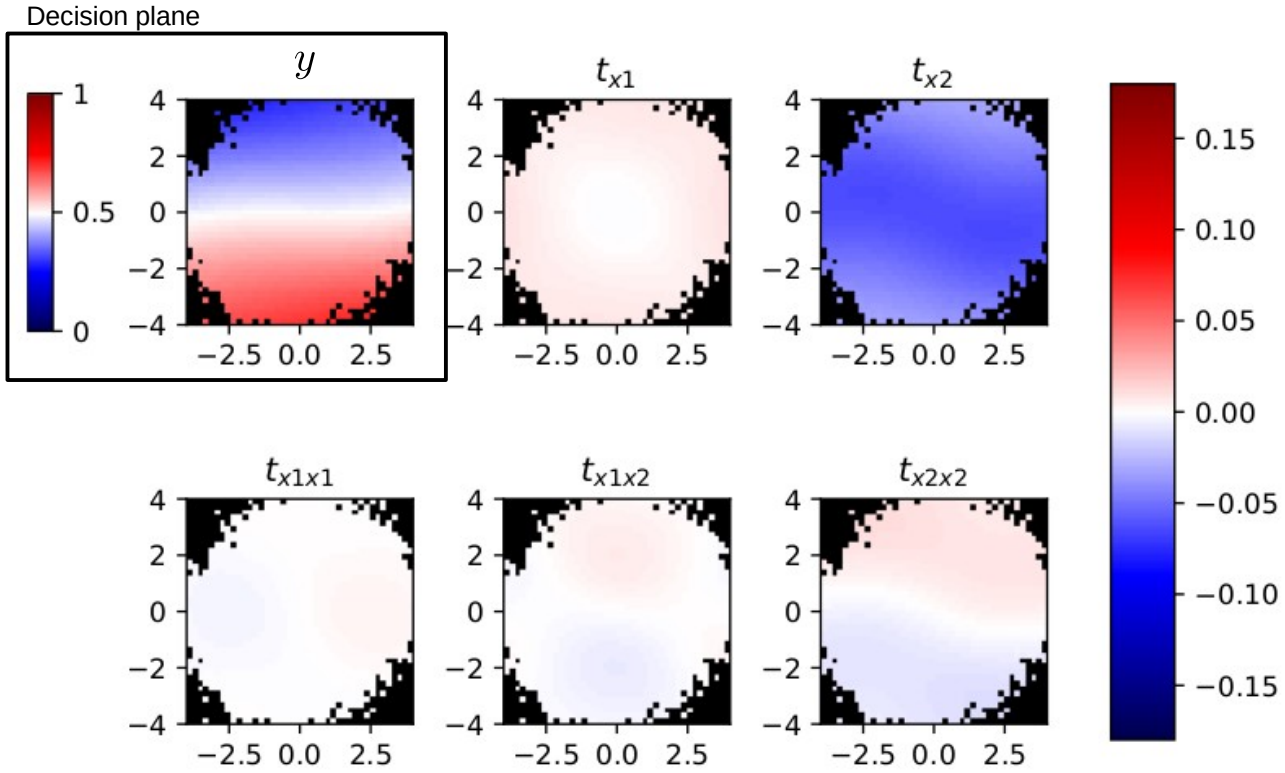
- x_1 and x_2 found to be most influential (distinction of S and BG by location).
- x_ix_j indicate that correlation plays a role (for S and BG x_1 and x_2 are linearly correlated).
- After convergence $\langle t_\alpha \rangle$ show stable and reproducible behavior.

… as a function of time and sample space

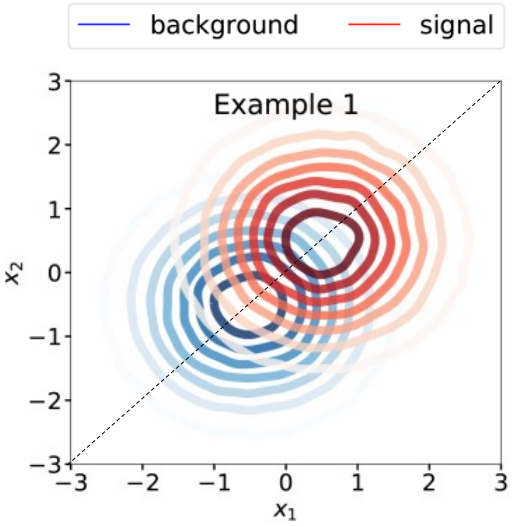


- Monitor what phase space regions the NN identifies as important and at what point in the training it starts to investigate them:

- After epoch 1:

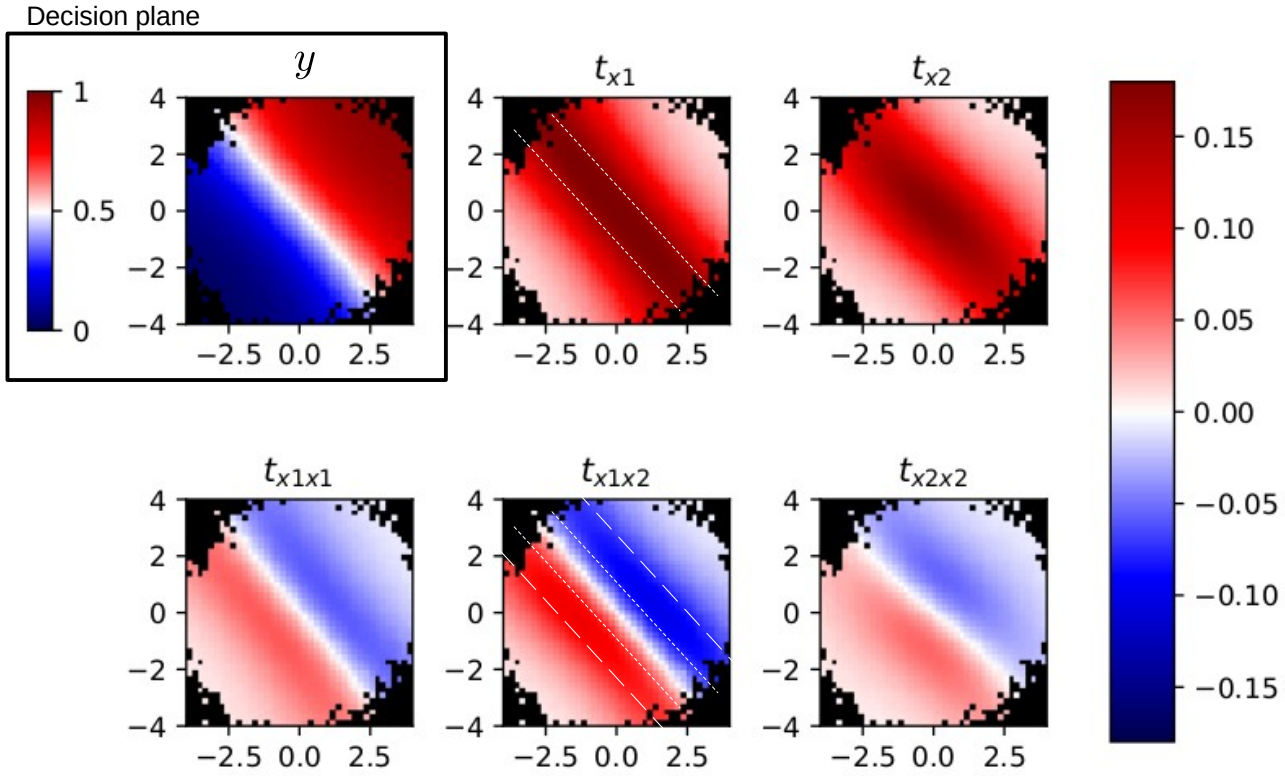


… as a function of time and sample space



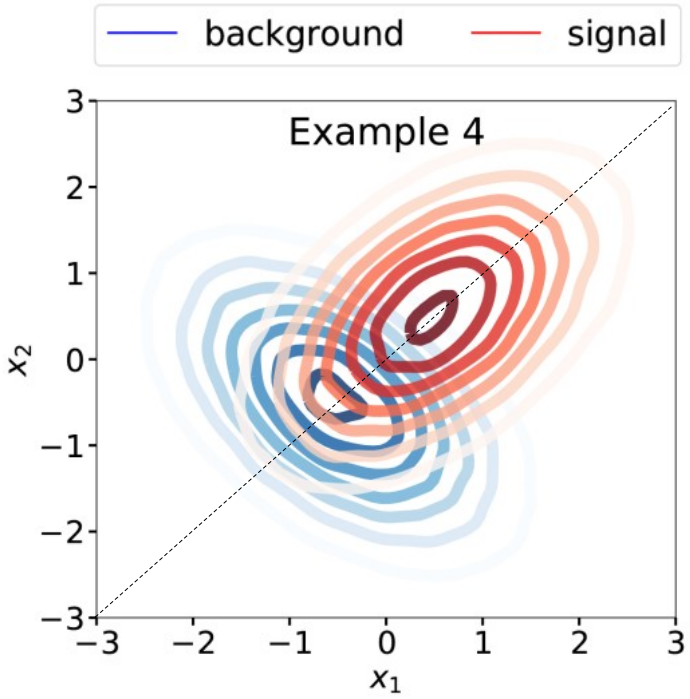
- Monitor what phase space regions the NN identifies as important and at what point in the training it starts to investigate them:

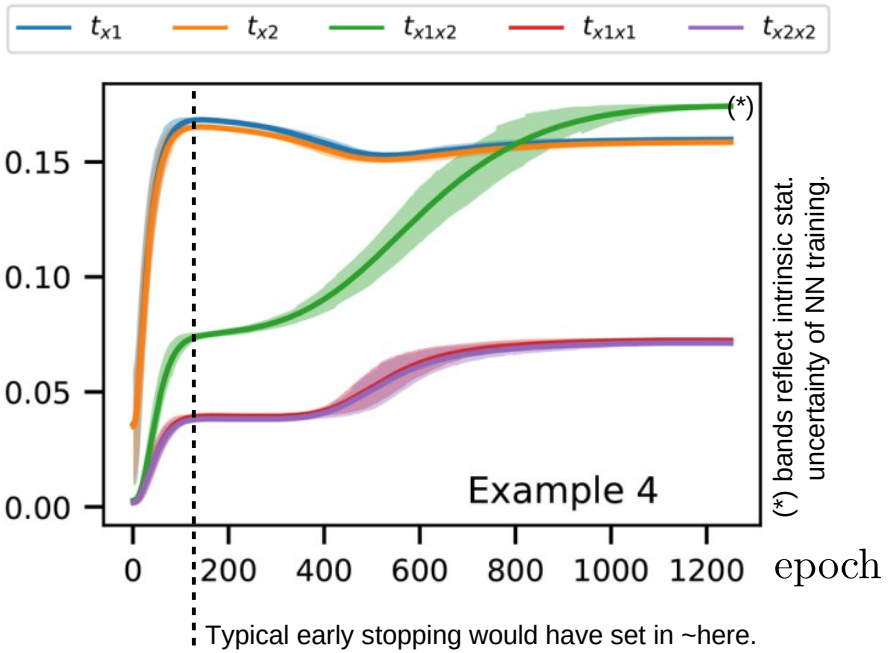
- After epoch 50:



Slightly more complex task

- Adding different linear correlations to S and BG:

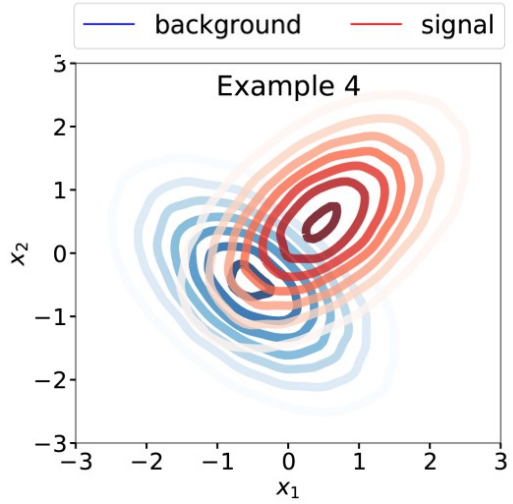




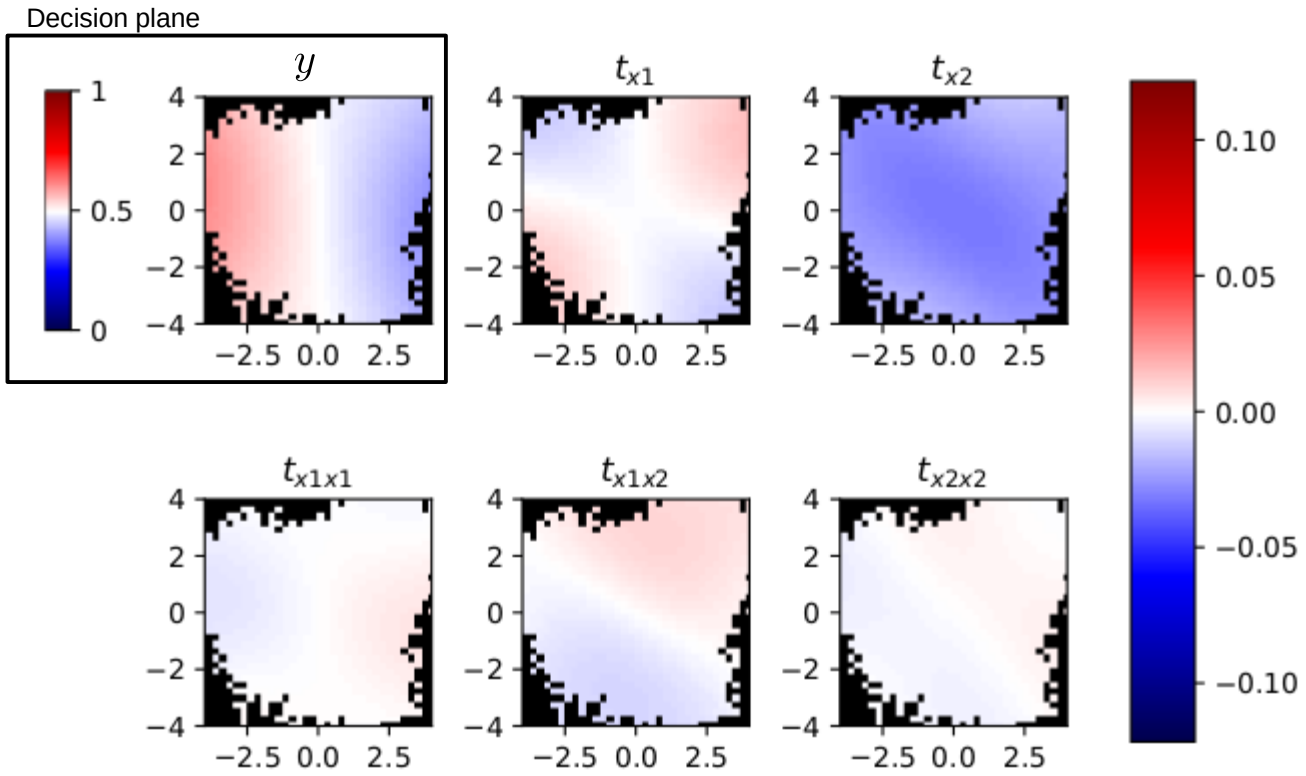
- Steep learning curve to identify x_1 and x_2 as important.

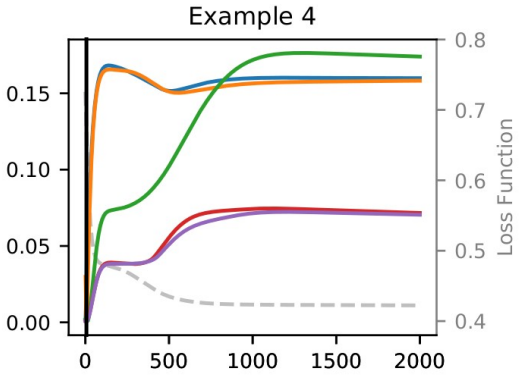
- After a phase of “contemplation” x_1 and x_2 are recognized as being slightly overrated. Later the additional importance of $x_1 x_2$ is identified (improving ROC-AUC by ~10%).

Watch the NN learn

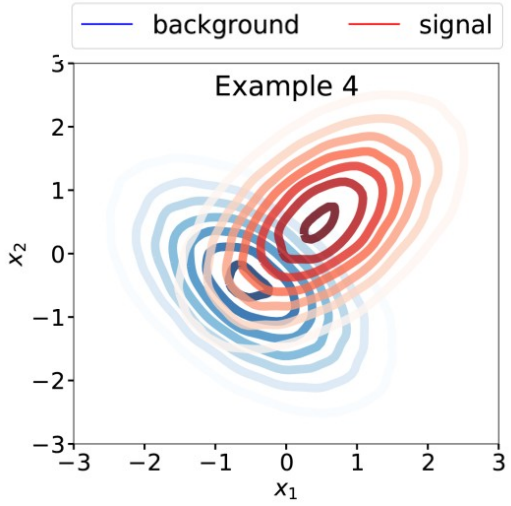


- After epoch 1:

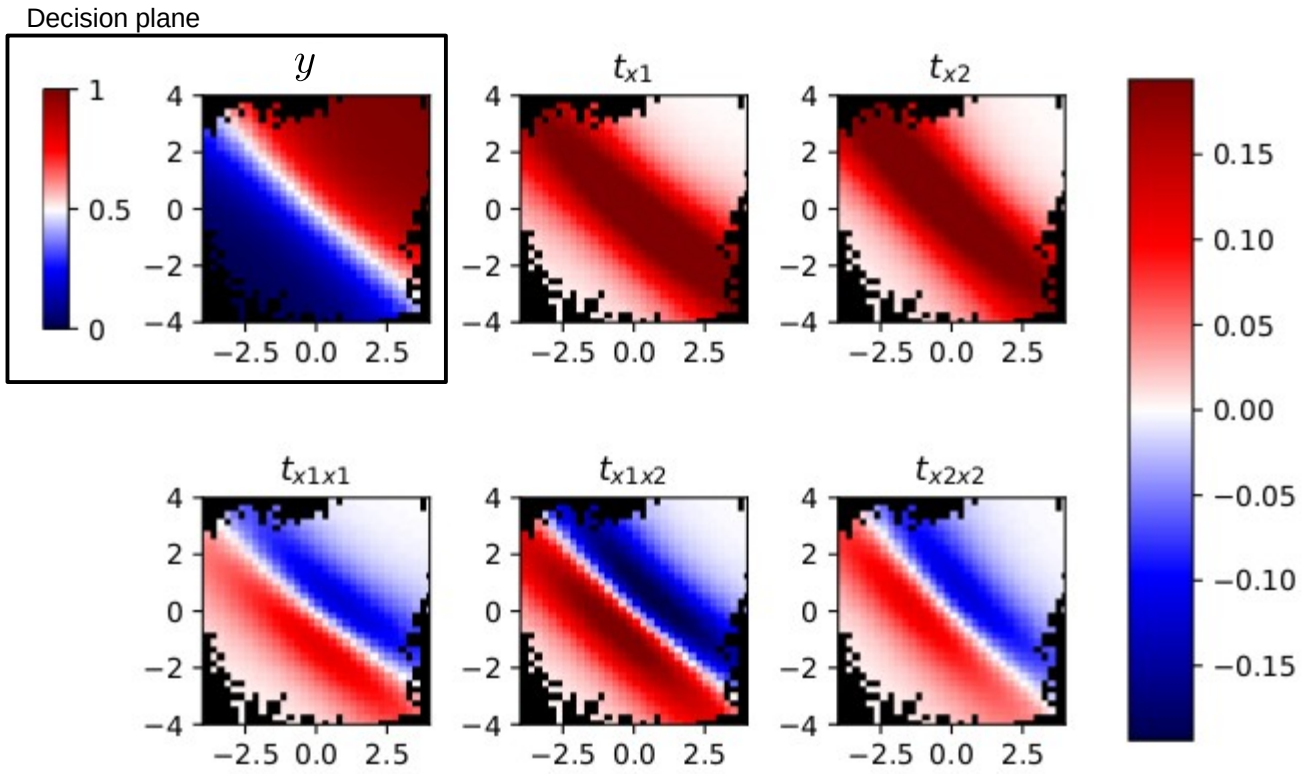


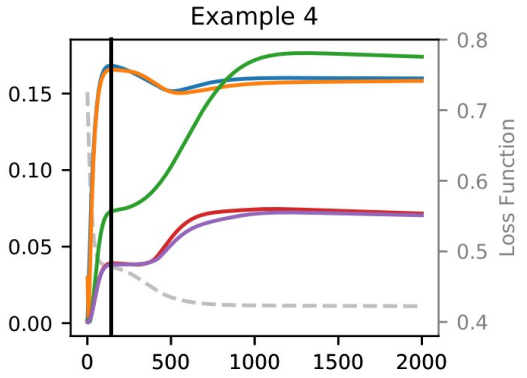


Watch the NN learn



- After epoch 100:



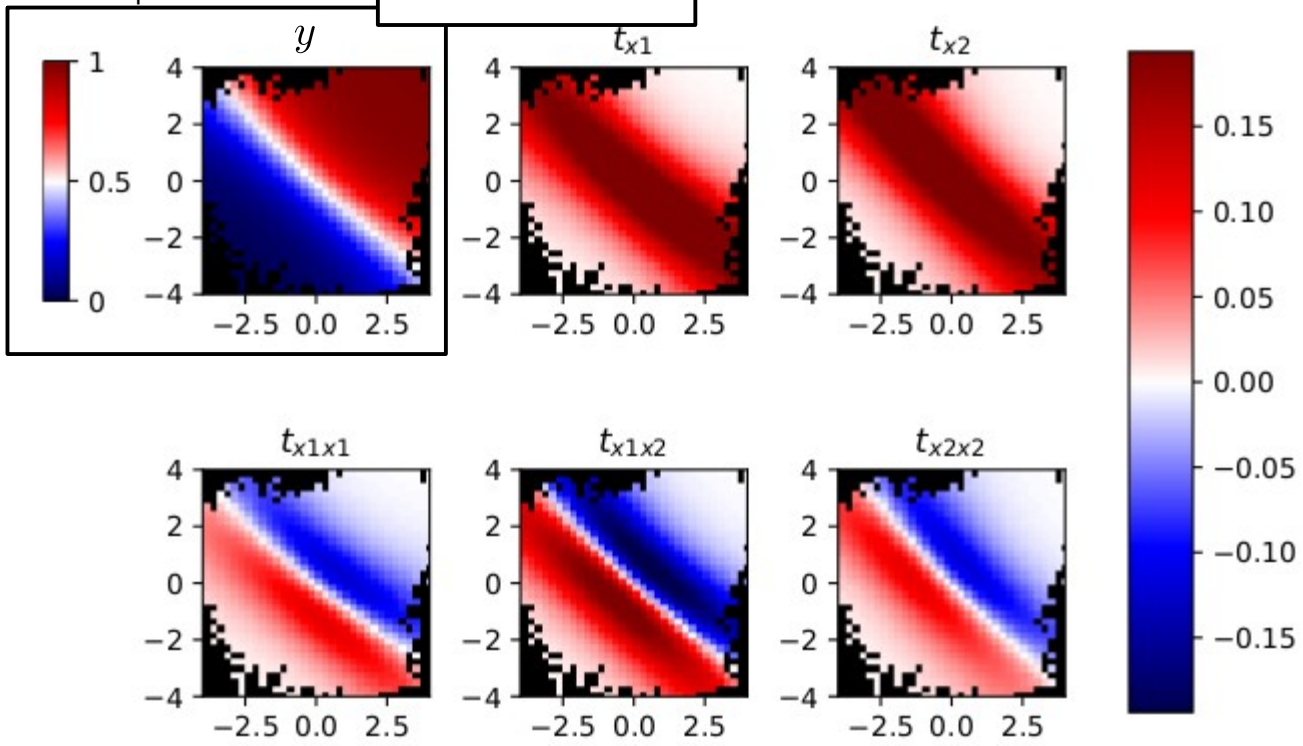


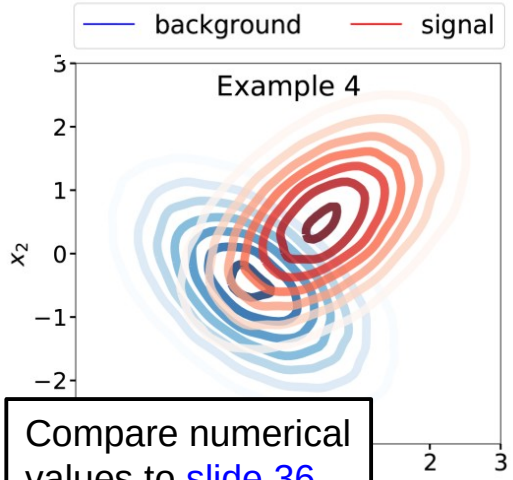
Watch the NN learn

- Learned that S and BG are separated in feature space. Difference in correlations missed until epoch ~1000!

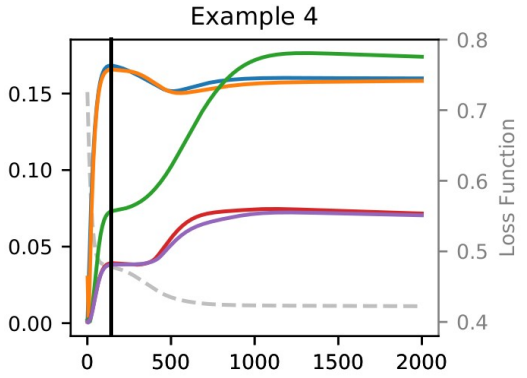
- After epoch 100:

Compare figures to slide 38.





Compare numerical values to slide 36.

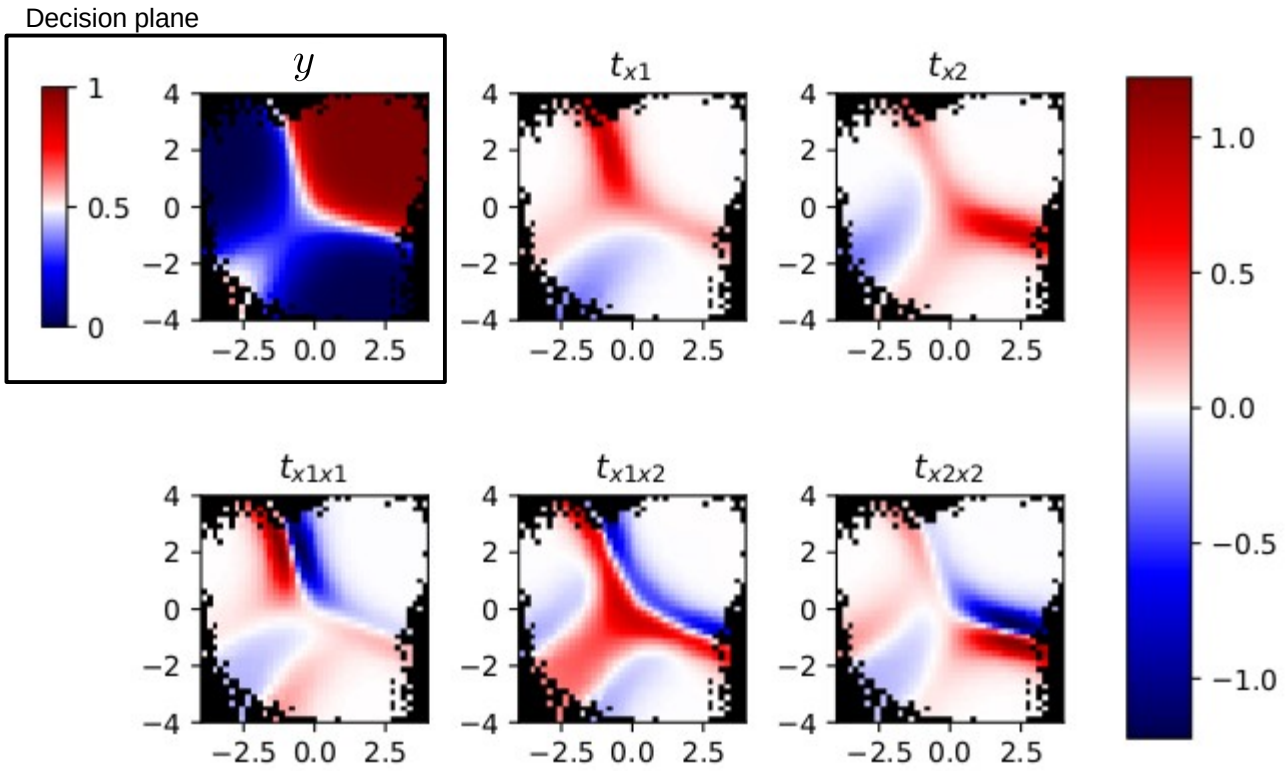


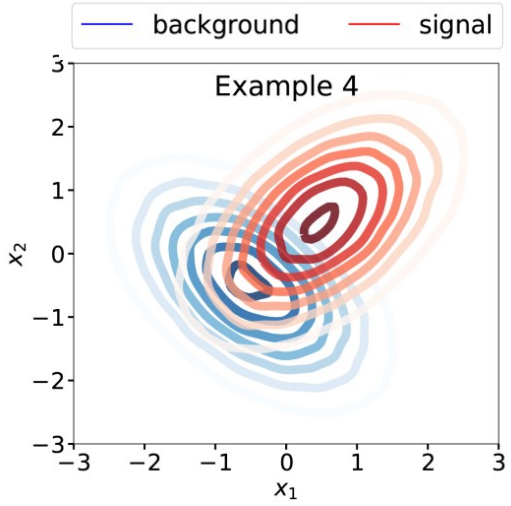
Early stopping as used for paper was reached after 350 epochs.

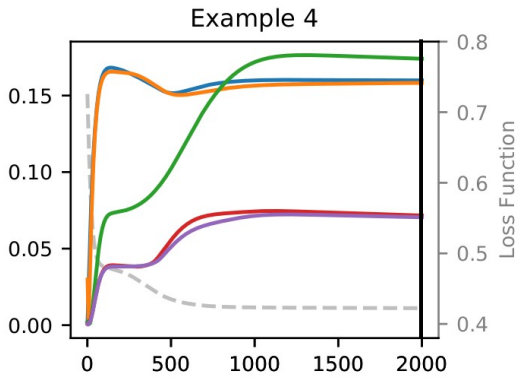
Watch the NN learn

- Learned that S and BG are separated in feature space. Difference in correlations missed until epoch ~1000!

- After epoch 2000:







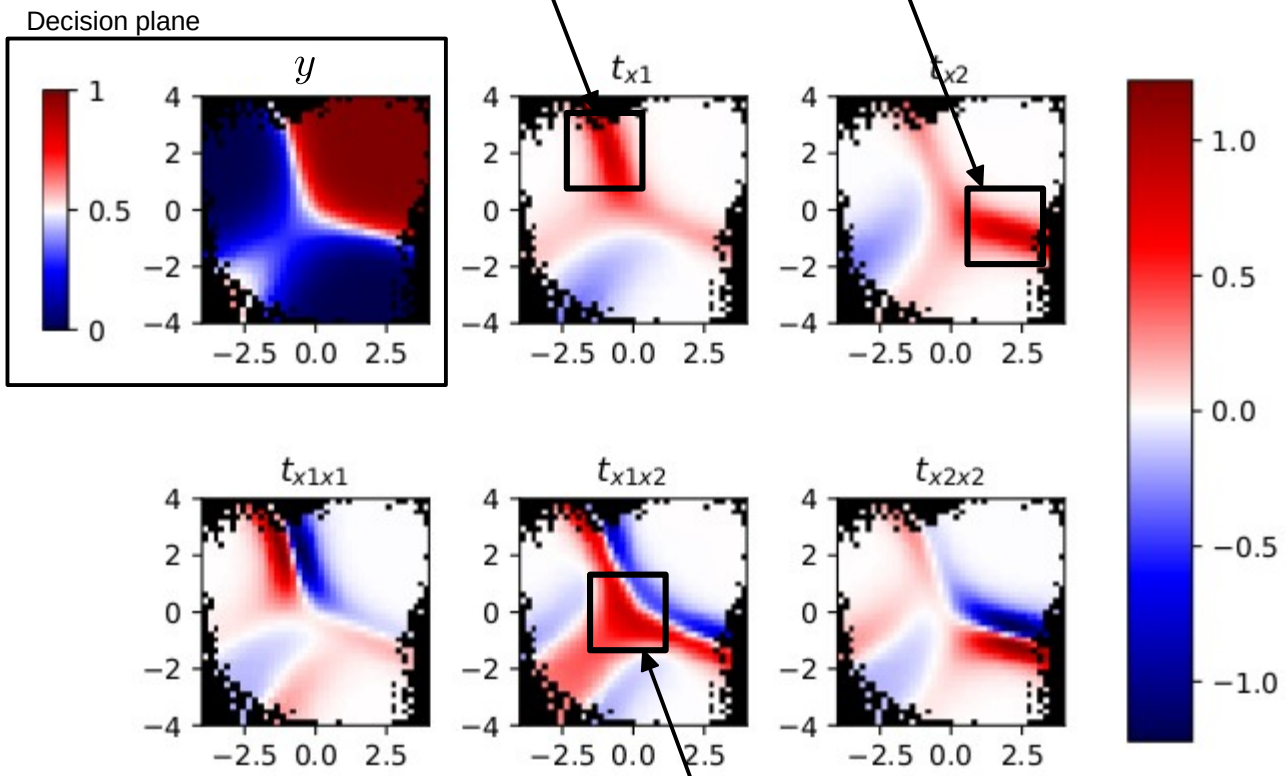
Priv. Doz. Dr. Roger Wolf
http://ekpwww.physik.uni-karlsruhe.de/~rwolf/

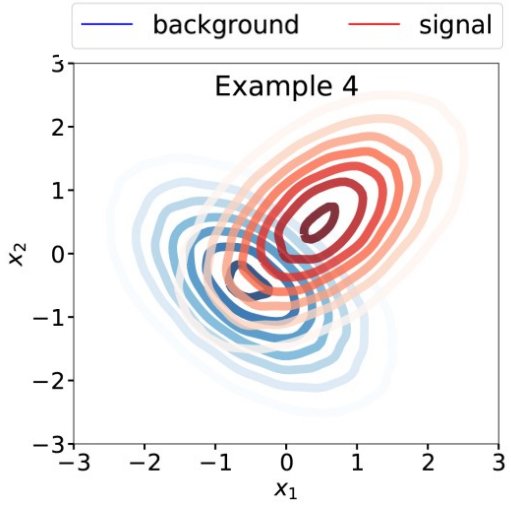
Watch the NN learn

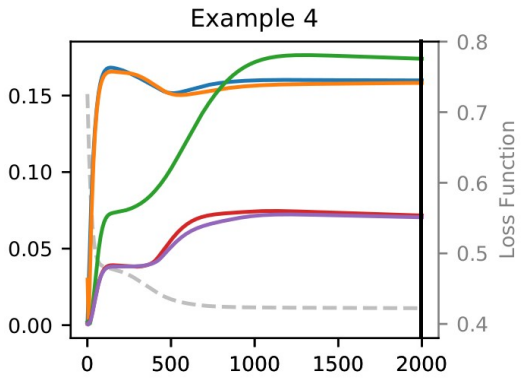
Position in x_1 is important for NN decision.

Position in x_2 is important for NN decision.

- After epoch 2000:





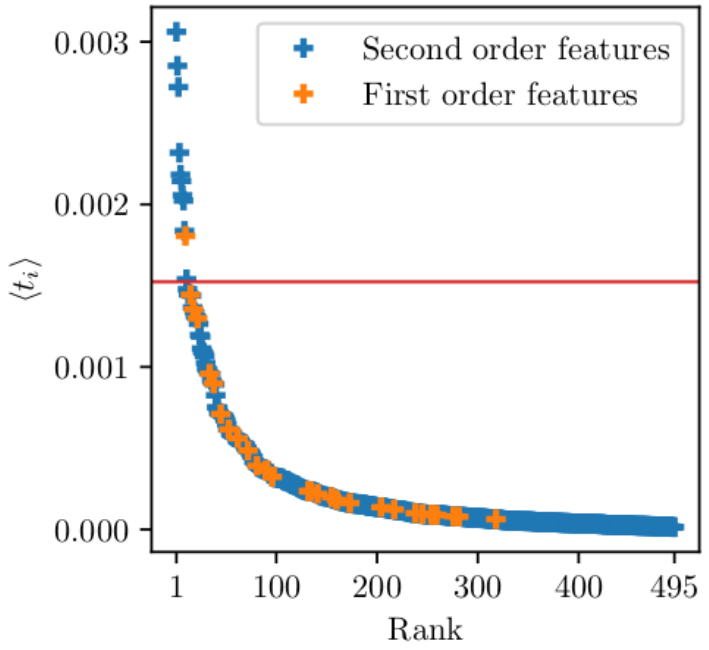


Correlation between x_1 and x_2 is important for NN decision.

Priv. Doz. Dr. Roger Wolf
http://ekpwww.physik.uni-karlsruhe.de/~rwolf/

More realistic physics task

- ATLAS data from the ML challenge of 2014 [1].
- $\langle t_\alpha \rangle$ used here for unambiguous importance ranking of *generalized features* (see right).



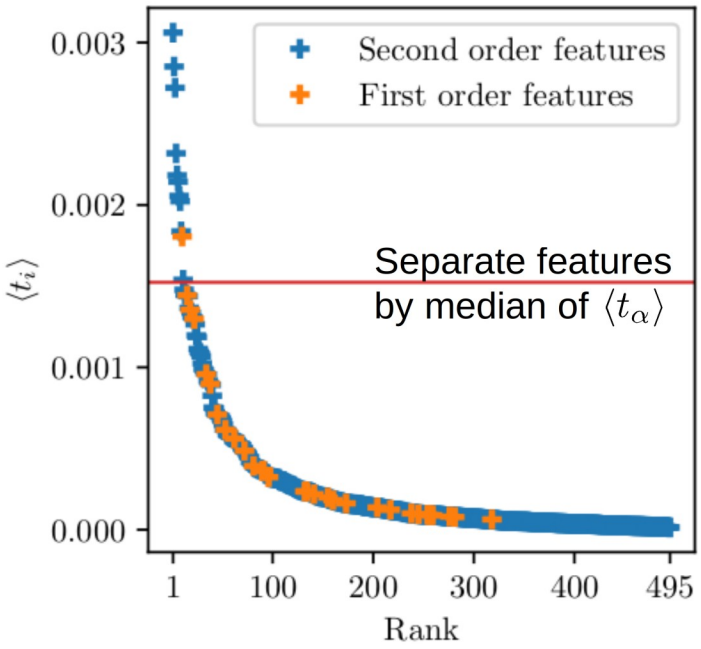
Rank	Variables		$\langle t_i \rangle \times 10^{-3}$
1	DER_mass_vis	DER_pt_ratio_lep_tau	3.061
2	DER_deltar_tau_lep	DER_mass_vis	2.852
3	DER_mass_vis	PRI_lep_pt	2.722
4	DER_deltar_tau_lep	DER_pt_ratio_lep_tau	2.318
5	DER_pt_ratio_lep_tau	PRI_lep_pt	2.182
6	DER_mass_vis	DER_mass_vis	2.144
7	DER_mass_MMC	DER_mass_vis	2.056
8	DER_deltar_tau_lep	PRI_lep_pt	2.023
9	DER_mass_jet_jet	DER_mass_vis	1.837
10	DER_mass_vis		1.806
11	DER_mass_MMC	DER_pt_ratio_lep_tau	1.539
12	DER_mass_transverse_met_lep	DER_mass_vis	1.478
13	DER_mass_jet_jet	DER_pt_ratio_lep_tau	1.447
14	DER_deltar_tau_lep	DER_mass_MMC	1.446
15	DER_pt_ratio_lep_tau		1.443
16	DER_mass_MMC	PRI_lep_pt	1.438
17	DER_deltar_tau_lep	DER_mass_jet_jet	1.366
18	DER_deltar_tau_lep		1.355
19	DER_mass_jet_jet	PRI_lep_pt	1.337
20	DER_mass_MMC	DER_mass_MMC	1.312
	...		
476	PRI_tau_eta	PRI_tau_phi	0.020
477	PRI_jet_leading_pt	PRI_met_phi	0.019
478	PRI_jet_leading_eta	PRI_jet_subleading_eta	0.019
479	PRI_jet_leading_eta	PRI_lep_phi	0.019
480	PRI_jet_subleading_phi	PRI_lep_phi	0.019
481	DER_sum_pt	PRI_tau_phi	0.019
482	DER_sum_pt	PRI_met_phi	0.019
483	PRI_jet_num	PRI_met_phi	0.018
484	DER_prodeta_jet_jet	PRI_met_phi	0.018
485	PRI_lep_eta	PRI_met_phi	0.018
486	DER_pt_tot	PRI_met_phi	0.018
487	PRI_jet_subleading_phi	PRI_met_phi	0.018
488	DER_sum_pt	PRI_tau_eta	0.018
489	PRI_lep_eta	PRI_tau_phi	0.018
490	PRI_jet_num	PRI_lep_phi	0.017
491	PRI_jet_leading_eta	PRI_lep_eta	0.017
492	PRI_jet_leading_eta	PRI_met_phi	0.016
493	PRI_met_phi	PRI_tau_eta	0.015
494	PRI_lep_phi	PRI_tau_eta	0.015
495	PRI_jet_leading_eta	PRI_tau_phi	0.014

PhD Stefan Wunsch 2021

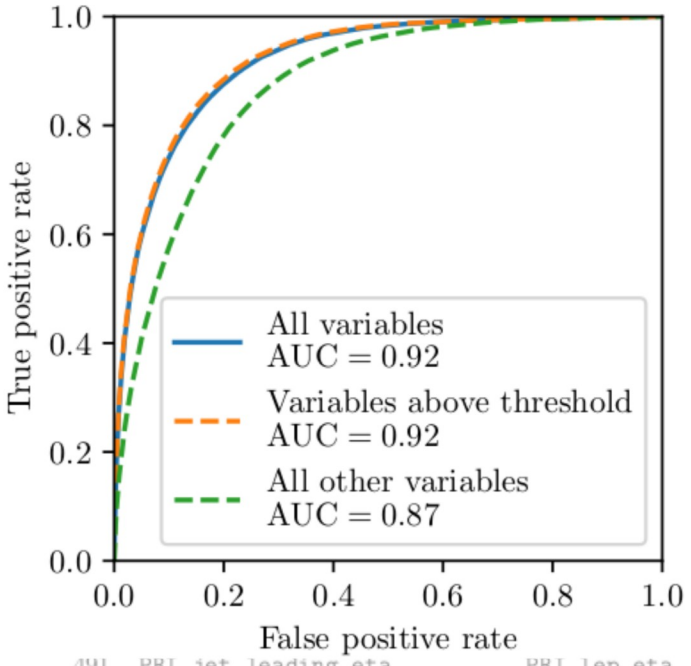
[1] Learning to discover: the Higgs boson machine learning challenge

More realistic physics task

- ATLAS data from the ML challenge of 2014 [1].
- $\langle t_\alpha \rangle$ used here for unambiguous importance ranking of *generalized features* (see right).



Rank	Variables		$\langle t_i \rangle \times 10^{-3}$
1	DER_mass_vis	DER_pt_ratio_lep_tau	3.061
2	DER_deltar_tau_lep	DER_mass_vis	2.852
3	DER_mass_vis	PRI_lep_pt	2.722
4	DER_deltar_tau_lep	DER_pt_ratio_lep_tau	2.318
5	DER_pt_ratio_lep_tau	PRI_lep_pt	2.182
6	DER_mass_vis	DER_mass_vis	2.144
7	DER_mass_MMC	DER_mass_vis	2.056
8	DER_deltar_tau_lep	PRI_lep_pt	2.023
9	DER_mass_jet_jet	DER_mass_vis	1.837
10	DER_mass_vis		1.806
11	DER_mass_MMC	DER_pt_ratio_lep_tau	1.539
12	DER_mass_transverse_met_lep	DER_mass_vis	1.478
13	DER_mass_jet_jet	DER_pt_ratio_lep_tau	1.447
14	DER_deltar_tau_lep	DER_mass_MMC	1.446
15	DER_pt_ratio_lep_tau		1.443
			1.438
		et	1.366
			1.355
			1.337
			1.312
			0.020
			0.019
		ding_eta	0.019
			0.019
			0.019
			0.019
			0.019
			0.018
			0.018
			0.018
			0.018
			0.018
			0.018
			0.018
			0.017
491	PRI_jet_leading_eta	PRI_lep_eta	0.017
492	PRI_jet_leading_eta	PRI_met_phi	0.016
493	PRI_met_phi	PRI_tau_eta	0.015
494	PRI_lep_phi	PRI_tau_eta	0.015
495	PRI_jet_leading_eta	PRI_tau_phi	0.014



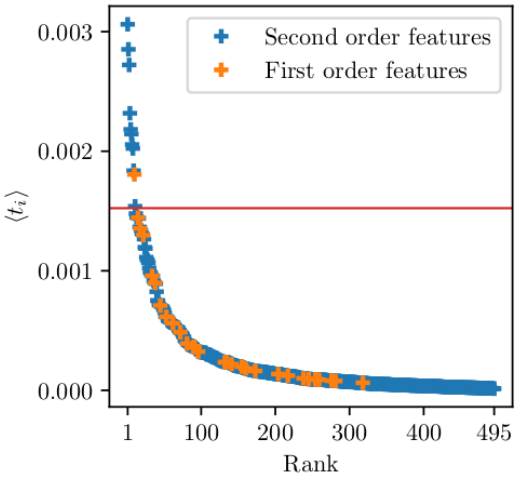
PhD Stefan Wunsch 2021

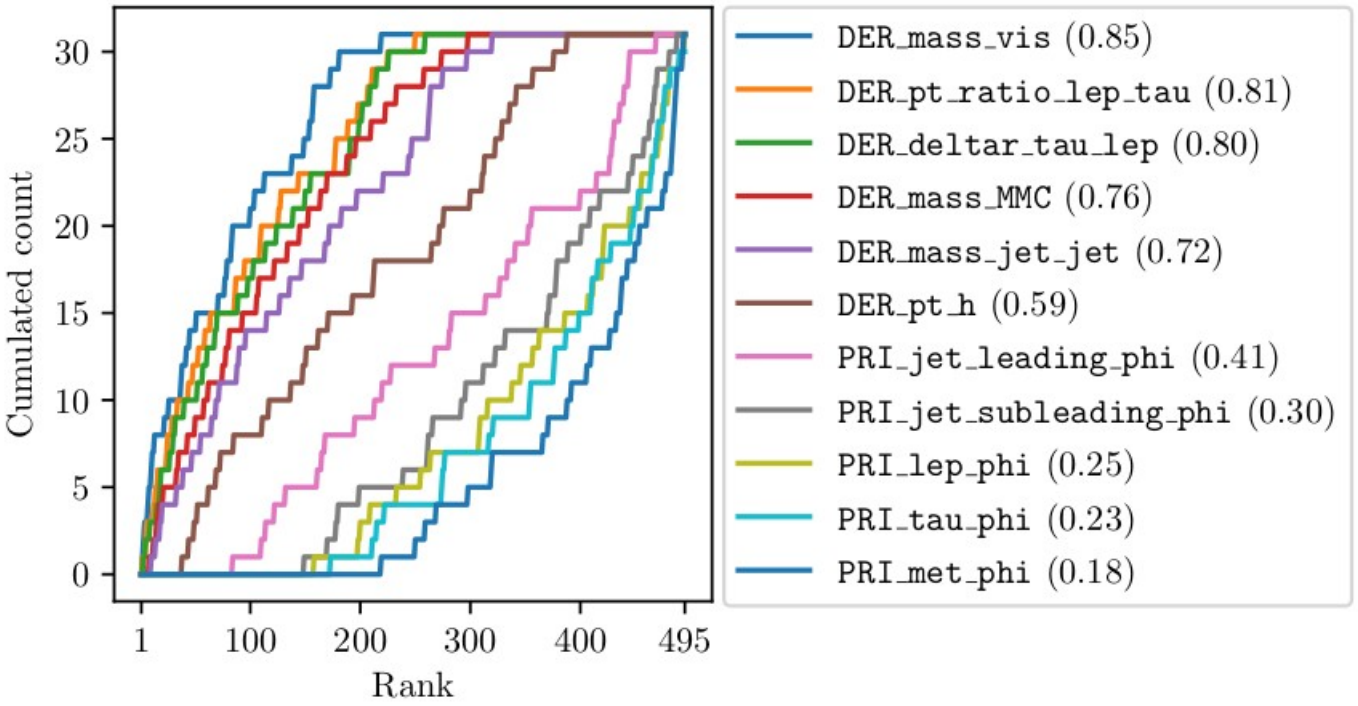
[1] Learning to discover: the Higgs boson machine learning challenge

Contract back to ranking of single features

- Check appearance of single features when scanning through 1. and 2. order feature ranks.

- Values in parentheses indicate AUC for each corresponding observable in the plot on the left .





PhD Stefan Wunsch 2021

[1] Learning to discover: the Higgs boson machine learning challenge

Conclusions from Higgs boson ML challenge

- Without knowing anything about physics the NN has:

Conclusions from Higgs boson ML challenge

- Without knowing anything about physics the NN has:
  - Identified mvis and MMC mass as important.

Conclusions from Higgs boson ML challenge

- Without knowing anything about physics the NN has:
  - Identified mvis and MMC mass as important.
  - Identified that both observables are peaking for S in contrast BGs. It also rates the position of the peak high.

Conclusions from Higgs boson ML challenge

- Without knowing anything about physics the NN has:
  - Identified mvis and MMC mass as important.
  - Identified that both observables are peaking for S in contrast BGs. It also rates the position of the peak high.
  - We observe that for the NN 2. order features, i.e., correlations across features are generally more important for the decision than 1. order features.

Conclusions from Higgs boson ML challenge

- Without knowing anything about physics the NN has:
  - Identified mvis and MMC mass as important.
  - Identified that both observables are peaking for S in contrast BGs. It also rates the position of the peak high.
  - We observe that for the NN 2. order features, i.e., correlations across features are generally more important for the decision than 1. order features.
- This convinces us that the NN does indeed **identify the general physics features** that we would also choose from our educated physics intuition.

Conclusions from Higgs boson ML challenge

- Without knowing anything about physics the NN has:
  - Identified mvis and MMC mass as important.
  - Identified that both observables are peaking for S in contrast BGs. It also rates the position of the peak high.
  - We observe that for the NN 2. order features, i.e., correlations across features are generally more important for the decision than 1. order features.
- This convinces us that the NN does indeed **identify the general physics features** that we would also choose from our educated physics intuition.
- We can even tell apart what features it identifies **when** in the training and from **where** in the sampled phase space of the training sample it picks up on it.

Lessons from this lecture

- ML is exciting! It is very quickly evolving. Somtimes people move quicker than their feet can carry them.

Lessons from this lecture

- ML is exciting! It is very quickly evolving. Somtimes people move quicker than their feet can carry them.

- Believe in statistics, believe in yourself as a physicist, **believe in your common sense**!

Lessons from this lecture

- ML is exciting! It is very quickly evolving. Somtimes people move quicker than their feet can carry them.

- Believe in statistics, believe in yourself as a physicist, **believe in your common sense**!

- Don‘t think you are stupid – if you don‘t understand a reasoning its very likely bullsh…

Lessons from this lecture

- ML is exciting! It is very quickly evolving. Somtimes people move quicker than their feet can carry them.
- Believe in statistics, believe in yourself as a physicist, **believe in your common sense**!
- Don‘t think you are stupid – if you don‘t understand a reasoning its very likely bullsh…
- Do not fight complex problems with even more complex tools. If you don‘t understand a complex phenomenon simplify it! How can you hope that making a problem even more complex can help your understanding?

Lessons from this lecture

- ML is exciting! It is very quickly evolving. Somtimes people move quicker than their feet can carry them.
- Believe in statistics, believe in yourself as a physicist, **believe in your common sense**!
- Don‘t think you are stupid – if you don‘t understand a reasoning its very likely bullsh…
- Do not fight complex problems with even more complex tools. If you don‘t understand a complex phenomenon simplify it! How can you hope that making a problem even more complex can help your understanding?
- In matters of understanding – **Never being satisfied will make you ever grow**.

Backup

Example to set up

- For an example how to set up and apply Taylor coefficients on a simple benchmark task you can checkout this Jupyter notebook (password: *gradient*).

- Bachelor:

  Illustration of the neural network learning process during training
  (**Bachelor**, KIT, Greta Heine, 2019).

  Illustration of neural network learning with uncertainties
  (**Bachelor**, KIT, Christian Winter, 2020).

- Master:

  Application of multivariate analysis techniques to an analysis of Higgs boson decays to τ leptons
  (**Master**, KIT, Marcus Schmitt, 2016).

  A novel strategy for the standard model H→ττ analysis with emphasize on minimizing systematic uncertainties in presence of modern multivariate methods
  (**Master**, KIT Stefan Wunsch, 2017).

  Standard model H→ττ analysis with a neural network trained on a mix of simulation and data samples
  (**Master**, KIT, Moritz Scham, 2020).

  Studies of the usage of neural networks in particle physic analyses
  (**Master**, KIT, Simon Jörger, 2020).

- Veröffentlichungen:

  Measurement of Higgs boson production and decay to the ττ final state
  (CMS-PAS-HIG-18-032).

  Identifying the relevant dependencies of the neural network response on characteristics of the input space
  (CSBS 2 (2018), 1).

  Reducing the dependence of the neural network function to systematic uncertainties in the input space
  (CSBS 4 (2020), 1).

  Optimal statistical inference in the presence of systematic uncertainties using neural network optimization base d on binned Poisson likelihoods with nuisance parameters
  (CSBS 5 (2021) 4).