Lecture 6

Vanishing Gradient Problem

\[ f(x) = A(w_1 \cdot A(w_{k-1} \cdot A(w_{k-2} \cdot A(\ldots \cdot A(w_1 \cdot x)) \ldots)) \]

\[ \frac{df}{dw_k} = G_k \cdot x_k \]

\[ x_k = A(w_{k-1} \cdot x_{k-1}) \]

\[ G_{k-1} = G_k \cdot v_k \cdot A'(w_{k-1} \cdot x_{k-1}) \]

In early days of NN's, common to use sigmoid \( \sigma \) for \( A \)

\[ A(x) = \frac{e^x}{1 + e^x} \]

weights getting stuck far from optimun

decreasing or suppressed gradients esp. for initial layers

as network gets deeper:

\[ A'(x) \approx 0.25 \]
Sol'n: use a better activation function. For example, ReLU: \( \max(0, x) \). This enabled much deeper networks.
Back to binary classification

Key result in binary classification: Neyman–Pearson lemma: optimal binary classifier is the likelihood ratio

\[ R(x) = \frac{P(x|S)}{P(x|B)} \]

Proof:

Consider BC loss

\[ L = \sum_{i \in S} \log f(x) + \sum_{i \in B} \log (1-f(x)) \]

\[ = \int dx \left( P(x|S) \log f(x) + P(x|B) \log (1-f(x)) \right) \]

\[ f \rightarrow f + \delta f, \text{ set } \delta f \text{ to zero} \]

\[ 0 = \frac{P(x|S)}{f} - \frac{P(x|B)}{1-f} \rightarrow f = \frac{P(x|S)}{P(x|S) + P(x|B)} \]

(\( f(x) = P(S|x) \)?)

Bayes' theorem: \( P(S|x) P(x) = P(x|S) P(S) \)

At fixed tpr, \( R(x) \) achieves lowest possible tpr for every tpr.

\[ R(x) \leq R \]

\[ \frac{R(x)}{R(x)+1} \]

\[ f > c \implies R > c' \]

\[ \frac{R}{f} \]
f is monotonic in R \implies \text{equivalent classifiers} \checkmark
\text{LR minimizes BCE loss} \checkmark

\text{Proof \#2: (directly showing that R gives best possible ROC curve)}

Consider any other classifier \( f(x) \). \( f(x) > c \rightarrow \) defines a region \( C \) in \( x \) space

\[
\text{Now consider a cut on } R(x) \geq a \uparrow \text{ same } \text{TPR}. \text{ This defines another region } f
\]

\[ e_{f}(f) = e_{f}(R) = \int_{C} p(x|S). \]
We want to compare $\varepsilon_B(f) \neq \varepsilon_B(R)$, (background efficiencies on fprz)

$$\int_C P(xIB) - \int_P P(xIB).$$

$$\varepsilon_B(R) = \int_A P(xIB) = \int_A P(xIB) + \int_{ANC} P(xIB) + \int_{ANC^\perp} P(xIB) = \int_C P(xIB) = \int_{A^\perp} P(xIB) + \int_{ANC} P(xIB)$$

$$\geq \int_{R \cap \{x\}} - \int_{R \setminus \{x\}} P(x15) + \int_{A^\perp} P(x15)$$

$$< \varepsilon_B(f) + \frac{1}{a} \left[ \int_{ANC} P(x15) - \int_{ANC^\perp} P(x15) \right]$$

$$= \varepsilon_B(f) + \frac{1}{a} \left[ \int_{ANC} P(x15) - \int_{ANC^\perp} P(x15) \right] = \varepsilon_B(f) + \frac{1}{a} \left[ \int_{ANC} P(x15) - \int_{ANC^\perp} P(x15) \right] = \varepsilon_B(f) \checkmark$$
NP lemma is extremely useful!

- optimal classifier exists, given \( H \) \( \Rightarrow \) provides strategy for hypothesis testing, especially for new physics

- modern ML: "likelihood ratio trick"

Assume: NN can learn the optimal classifier \( \Rightarrow \frac{P(y|s)}{P(y|x)} \)

"likelihood free inference" \( \Rightarrow \) gives access to likelihood ratio without knowing individual likelihoods!

Applications:

- generation: GANs
- anomaly detection: "weak supervision", CWoLa
- phase space reweighting, e.g., simulation vs data
**Multiclass Classification**

- Important & straightforward generalization of binary classification

- Objective: learn class probabilities
  \[ p(C_k|x) \quad k = 1, \ldots, N_{\text{class}} \]

  \[ \sum_k p(C_k|x) = 1. \]

  *Ex: MNIST*  
  classify 0-9 from each other.

- Loss fn? MLE binary:
  \[ \log p(y|x) \quad \sum_i \log p(y|x) \quad i \neq y \]

  multiclass:
  \[ L = - \sum_k \sum_{i \in C_k} \log p(C_k|x) \]

  "categorical cross entropy"
Think of $\mathbf{\rho}(C_k | x)$ as a vector $\begin{pmatrix} \rho(C_1 | x) \\ \vdots \\ \rho(C_k | x) \\ \vdots \\ \rho(C_N | x) \end{pmatrix}$.

Introduce truth labels $y_k = 1$ for $k = 1$, $y_{\text{other}} = 0$ for $k = 1, \ldots, k = N$.

One-hot encoding.

$$L = \sum_{i \in \text{data}} \log \mathbf{\rho}(x_i)$$

Log acts elementwise over vector $\mathbf{\rho}$. Truth label for data pt. $i$.
Metrics:
- Accuracy
  \[ \text{fraction of } \hat{y}(x) = \text{true class} \]
- One-vs-rest binary metrics
- Confusion matrix

\[
\begin{array}{cccc}
\text{true} & C_1 & C_2 & C_3 & \ldots \\
C_1 & & & & \\
C_2 & & & & \\
C_3 & & & & \\
\vdots & & & & \\
\end{array}
\]