Exercise 7

Federated Learning

Reliable and Trustworthy Artificial Intelligence
ETH Zurich

Problem 1 (Analytical Gradient Inversion)**.** In this question, we are considering a simple 2-layer neural network with a hidden layer $o_2 \in \mathbb{R}^m$ that takes as input $x \in \mathbb{R}^n$ and produces a single binary classification output $p \in \mathbb{R}$:

$$\begin{aligned} z_1 &= W_1 \cdot x + b_1 \\ o_2 &= \mathrm{ReLU}(z_1) \\ z_2 &= w_2 \cdot o_2 + b_2 \\ p &= \frac{1}{1 + e^{-z_2}}. \end{aligned} \tag{1}$$

We train it using federated learning, where each client k is using a binary cross entropy error function applied on its private data $(x_i, y_i) \sim \mathcal{D}_k$:

$$\mathcal{L}(x_i, y_i) = y_i \cdot \log(p(x_i)) + (1 - y_i) \cdot \log(1 - p(x_i)). \tag{2}$$

1. Use the chain rule to calculate the gradients sent to the server by client k on the data (x_i, y_i).
   Hint: The gradients are $\nabla_{w_2}\mathcal{L}(x_i, y_i)$, $\nabla_{b_2}\mathcal{L}(x_i, y_i)$, $\nabla_{W_1}\mathcal{L}(x_i, y_i)$ and $\nabla_{b_1}\mathcal{L}(x_i, y_i)$.

2. Use the calculated gradients to derive analytical formula for x_i in terms of the gradients calculated in the previous question. When is this formula valid?
   Hint: Look at the chain rule formulas for $\nabla_{W_1}\mathcal{L}(x_i, y_i)$ and $\nabla_{b_1}\mathcal{L}(x_i, y_i)$.

For further insight on analytical gradient inversion see [1, 2, 3].

Problem 2 (Optimization-based Gradient Inversion)**.** In this question, we will implement the optimization-based gradient inversion techniques we described in the lecture. The code you are asked to complete is provided in the form of a Jupyter notebook here.

References

[1] Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shiho Moriai, et al. "Privacy-preserving deep learning: Revisited and enhanced". In: *International Conference on Applications and Techniques in Information Security*. Springer. 2017, pp. 100–110.

[2] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. "Inverting Gradients–How easy is it to break privacy in federated learning?" In: *arXiv preprint arXiv:2003.14053* (2020).

[3] Junyi Zhu and Matthew Blaschko. "R-gap: Recursive gradient attack on privacy". In: *arXiv preprint arXiv:2010.07733* (2020).