TEMPLATE FOR STUDENT PROJECT FOR CS-E4740 FEDERATED LEARNING

N.N.

Aalto University, Espoo, Finland

ABSTRACT

This paper studies federated learning methods for high-precision weather forecasting.

Index Terms— Federated Learning, Networks, Personalized Machine Learning, Trustworthy AI

1. INTRODUCTION

- Explain the background (real-life scenario) of your ML application (see [1, Ch. 2]).
- Summarize the relevant literature (state-of-the art).
- Briefly outline the structure of this paper.

2. PROBLEM FORMULATION

- Discuss the source of the data used in your project.
- Formulate your application as an instance of GTV minimization (GTVMin) [2, Sec. 7]:
  - Discuss your choice/construction of the empirical graph whose nodes carry local datasets and local models.
  - Provide a precise definition of the local datasets, its data points and their features and labels.

3. METHODS

- Clearly state the number of data points in each local dataset.
- Explain your feature selection process, which might involve advanced ML methods ("representation learning").
- Describe your choice of local models and the measure for their variation across edges in the empirical graph.
- Describe the FL algorithms you have used to train the local models [2, Sec. 9]
- Describe and explain (why?) your choice of loss functions, e.g., logistic loss.
- Explain the process of model validation, e.e., how did you split the data into training set, validation set and test set.

4. RESULTS

- Compare and discuss the training error and validation error obtained for each node of the empirical graph
- What is the final chosen method? A federated learning (FL) method consists of a choice for the empirical graph, local models and measure for their variation as well as a FL (=distributed optimization) algorithm.
- What is the test set error (for each node in the empirical graph) of the final chosen method?

5. CONCLUSION

- Provide a succinct summary of your findings.
- Are the results suggesting that the problem is solved satisfactorily, or might there be room for improvement?
- Ponder about possible limitations of the considered methods and how they can be further improved (see [1, Sec. 6.6.]).

6. REFERENCES

[1] A. Jung, *Machine Learning: The Basics*, Springer Singapore, 1 edition, Feb. 2022.

[2] A. Jung, "Federated Learning," Lecture notes of the course CS-E4740, Aalto University, 2023.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New Jersey, 2 edition, 2006.

[4] P. Halmos, *Naive set theory*, Springer-Verlag, 1974.

[5] W. Rudin, *Principles of Mathematical Analysis*, McGraw-Hill, New York, 3 edition, 1976.

[6] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*, The MIT Press, Cambridge, Massachusetts, 2006.

[7] A. Jung and P.H.J. Nardelli, "An information-theoretic approach to personalized explainable machine learning," *IEEE Sig. Proc. Lett.*, vol. 27, pp. 825–829, 2020.

[8] A. Jung, "Explainable empiricial risk minimization," *submitted to IEEE Sig. Proc. Letters (preprint: https://arxiv.org/pdf/2009.01492.pdf)*, 2020.

[9] J. Chen, L. Song, M.J. Wainwright, and M.I. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation," in *Proc. 35th Int. Conf. on Mach. Learning*, Stockholm, Sweden, 2018.

[10] D.N. Gujarati and D.C. Porter, *Basic Econometrics*, McGraw Hill, 2009.

[11] Y. Dodge, *The Oxford Dictionary of Statistical Terms*, Oxford University Press, 2003.

[12] B.S. Everitt, *Cambridge Dictionary of Statistics*, Cambridge University Press, 2002.

[13] R. Tyrrell Rockafellar, *Network Flows and Monotropic Optimization*, Athena Scientific, Jul. 1998.

[14] C. Molnar, *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*, [online] Available: https://christophm.github.io/interpretable-ml-book/., 2019.

[15] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, Springer, New York, 2nd edition, 1998.

[16] D.P. Bertsekas and J.N. Tsitsiklis, *Introduction to Probability*, Athena Scientific, 2 edition, 2008.

[17] R.M. Gray, *Probability, Random Processes, and Ergodic Properties*, Springer, New York, 2 edition, 2009.

[18] Patrick Billingsley, *Probability and Measure*, Wiley, New York, 3 edition, 1995.

[19] Paul R. Halmos, *Measure Theory*, Springer, New York, 1974.

Glossary

computational aspects By computational aspects of a ML method, we mainly refer to the computational resources required for its implementation. For example, if a ML method uses iterative optimization techniques to solve empirical risk minimization (ERM), then its computational aspects include (i) how many arithmetic operations are needed to implement a single iteration (gradient step) and (ii) how many iterations are needed to obtain useful model parameters. One important example for an iterative optimization technique is gradient descent (GD). 4

data A (indexed) set of data points. 5

data point A data point is any object that conveys information [3]. Data points might be students, radio signals, trees, forests, images, RVs, real numbers or proteins. We characterize data points using two types of properties. One type of property is referred to as a feature. Features are properties of a data point that can be measured or computed in an automated fashion. Another type of property is referred to as labels. The label of a data point represents a higher-level facts or quantities of interest. In contrast to features, determining the label of a data point typically requires human experts (domain experts). Roughly speaking, ML aims at predicting the label of a data point based solely on its features. 1–6

dataset With a slight abuse of notation we use the terms "dataset" or "set of data points" to refer to an indexed list of data points $\mathbf{z}^{(1)}, \ldots,$. Thus, there is a first data point $\mathbf{z}^{(1)}$, a second data point $\mathbf{z}^{(2)}$ and so on. Strictly speaking a dataset is a list and not a set [4]. By using indexed lists of data points we avoid some of the challenges arising in concept of an abstract set. 2, 4

differentiable A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is differentiable if it has a gradient $\nabla f(\mathbf{x})$ everywhere (for every $\mathbf{x} \in \mathbb{R}^d$) [5]. 3

empirical graph Empirical graphs represent collections of local datasets and corresponding local models [6]. An empirical graph is an undirected weighted empirical graph whose nodes carry local datasets and models. FL methods learn a local hypothesis $h^{(i)}$, for each node $i \in \mathcal{V}$, such that it incurs small loss on the local datasets. 1, 4

empirical risk The empirical risk of a given hypothesis on a given set of data points is the average loss of the hypothesis computed over all data points in that set. 3, 5

empirical risk minimization Empirical risk minimization is the optimization problem of finding the hypothesis with minimum average loss (or empirical risk) on a given set of data points (the training set). Many ML methods are special cases of empirical risk. 2, 5, 6

Euclidean space The Euclidean space $\mathbb{R}^d$ of dimension d refers to the space of all vectors $\mathbf{x} = \big(x_1, \ldots, x_d\big)$, with real-valued entries $x_1, \ldots, x_d \in \mathbb{R}$, whose geometry is defined by the inner product $\mathbf{x}^T\mathbf{x}' = \sum_{j=1}^{d} x_j x'_j$ between any two vectors $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ [5]. 3, 5

explainability We can define the (subjective) explainability of a ML method as the level of predictability (or lack of uncertainty) in its predictions delivered to a specific human user. Quantitative measures for the (subjective) explainability can be obtained via probabilistic models for the data fed into the ML method. In particular, we can then identify explainability of a ML method via the uncertainty or entropy of the predictions it delivers [7, 8, 9]. 4

feature A feature of a data point is one of its properties that can be measured or computed in an automated fashion. For example, if a data point is a bitmap image, then we could use the red-green-blue intensities of its pixels as features. Some widely used synonyms for the term feature are "covariate","explanatory variable", "independent variable", "input (variable)", "predictor (variable)" or "regressor" [10, 11, 12]. However, this book makes consequent use of the term features for low-level properties of data points that can be measured easily. 1–5

feature space The feature space of a given ML application or method is constituted by all potential values that the feature vector of a data point can take on. Within this book the most frequently used choice for the feature space is the Euclidean space $\mathbb{R}^d$ with dimension d being the number of individual features of a data point. 3

federated learning (FL) Federated learning is an umbrella term for ML methods that train models in a collaborative fashion using decentralized data and computation. 1, 2

generalized total variation Generalized total variation measures the changes of vector-valued node attributes of a graph. 3

gradient For a real-valued function $f : \mathbb{R}^d \rightarrow \mathbb{R} : \mathbf{w} \mapsto f(\mathbf{w})$, a vector $\mathbf{a}$ such that $\lim_{\mathbf{w} \rightarrow \mathbf{w}'} \frac{f(\mathbf{w}) - \big(f(\mathbf{w}') + \mathbf{a}^T(\mathbf{w}-\mathbf{w}')\big)}{\|\mathbf{w}-\mathbf{w}'\|} = 0$ is referred to as the gradient of f at $\mathbf{w}'$. If such a vector exists it is denoted $\nabla f(\mathbf{w}')$ or $\nabla f(\mathbf{w})\big|_{\mathbf{w}'}$. 2, 3

gradient descent (GD) Gradient descent is an iterative method for finding the minimum of a differentiable function $f(\mathbf{w})$. 2

gradient step Given a differentiable real-valued function $f(\mathbf{w})$ and a vector $\mathbf{w}'$, the gradient step updates $\mathbf{w}'$ by adding the scaled negative gradient $\nabla f(\mathbf{w}')$, $\mathbf{w}' \mapsto \mathbf{w}' - \alpha \nabla f(\mathbf{w}')$. 2

graph A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a pair that consists of a node set $\mathcal{V}$ and an edge set $\mathcal{E}$. In general, a graph is specified by a map that assigns to each edge $e \in \mathcal{E}$ a pair of nodes [13]. One important family of graphs (simple undirected graphs) is obtained by identifying each edge $e \in \mathcal{E}$ with two different nodes $\{i, i'\}$. Weighted graphs also specify numeric weights A_e for each edge $e \in \mathcal{E}$. 2

GTV minimization GTV minimization is an instance of regularized empirical risk minimization (RERM) using the generalized total variation (GTV) of local paramter vectors as regularization term. 1

hypothesis A map (or function) $h : \mathcal{X} \rightarrow \mathcal{Y}$ from the feature space $\mathcal{X}$ to the label space $\mathcal{Y}$. Given a data point with features $\mathbf{x}$ we use a hypothesis map h to estimate (or approximate) the label y using the predicted label $\hat{y} = h(\mathbf{x})$. ML is about learning (or finding) a hypothesis map h such that $y \approx h(\mathbf{x})$ for any data point. 2, 4–6

hypothesis space Every practical ML method uses a specific hypothesis space (or model) $\mathcal{H}$. The hypothesis space of a ML method is a subset of all possible maps from the feature space to label space. The design choice of the hypothesis space should take into account available computational resources and statistical aspects. If the computational infrastructure allows for efficient matrix operations and we expect a linear relation between features and label, a resonable first candidate for the hypothesis space is a linear model. 3–6

i.i.d. It can be useful to interpret data points $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}$ as realizations of independent and identically distributed RVs with a common probability distribution. If these RVs are continuous, their joint probability density function (pdf) is $p\big(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}\big) = \prod_{r=1}^{m} p\big(\mathbf{z}^{(r)}\big)$ with $p(\mathbf{z})$ being the common marginal pdf of the underlying RVs. 4, 5

i.i.d. assumption The i.i.d. assumption interprets data points of a dataset as the realizations of independent and identically distributed (i.i.d.) RVs. 5

interpretability A synomym for explainability [14]. 4

label A higher level fact or quantity of interest associated with a data point. If a data point is an image, its label might be the fact that it shows a cat (or not). Some widely used synonyms for the term label are ”response variable”, ”output variable” or ”target” [10, 11, 12]. 1–5

label space Consider a ML application that involves data points characterized by features and labels. The label space is constituted by all potential values that the label of a data point can take on. Regression methods, aiming at predicting numeric labels, often use the label space $\mathcal{Y} = \mathbb{R}$. Binary classification methods use a label space that consists of two different elements, e.g., $\mathcal{Y} = \{-1, 1\}$, $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \{$“cat image”, ”no cat image”$\}$ 3, 4

linear model This book uses the term linear model in a very specific sense. In particular, a linear model is a hypothesis space which consists of all linear maps,

$$\mathcal{H}^{(d)} := \left\{ h(\mathbf{x}) = \mathbf{w}^T \mathbf{x} : \mathbf{w} \in \mathbb{R}^d \right\}. \tag{1}$$

Note that (1) defines an entire family of hypothesis spaces, which is parametrized by the number d of features that are linearly combined to form the prediction $h(\mathbf{x})$. The design choice of d is guided by computational aspects (smaller d means less computation), statistical aspects (increasing d might reduce prediction error) and interpretability (a linear model using few carefully chosen features might be considered interpretable). 3

local dataset The concept of a local dataset is in-between the concept of a data point and a dataset. Similar to the basic notion of a dataset, also a local dataset consists of several individual data points which are characterized by features and label. In constrast to a single dataset used in basic ML methods, a local dataset is also related to other local datasets via different notions of similarities. These similarities might arising from probabilistic models or communcation infrastructure and are encoded in the edges of an empirical graph. 1, 2, 4

local model Consider a collections of local datasets that are assigned to the nodes of an empirical graph. A local model $\mathcal{H}^{(i)}$ is a hypothesis space that is assigned to a node $i \in \mathcal{V}$. Different nodes might be assigned different hypothesis spaces, i.e., in general $\mathcal{H}^{(i)} \neq \mathcal{H}^{(i')}$ for different nodes $i, i' \in \mathcal{V}$. 1, 2

logistic loss Consider a data point that is characterized by the features $\mathbf{x}$ and a binary label $y \in \{-1, 1\}$. We use a real-valued hypothesis h to predict the label y solely from the features $\mathbf{x}$. The logistic loss incurred by a specific hypothesis h is defined as

$$L\left((\mathbf{x}, y), h\right) := \log(1 + \exp(-y h(\mathbf{x}))). \tag{2}$$

. Carefully note that the expression (2) for the logistic loss applies only for the label space $\mathcal{Y} = \{-1, 1\}$ and using the thresholding rule (**??**). 1

loss With a slight abuse of language, we use the term loss either for loss function itself or for its value for a specific pair of data point and hypothesis. 2–5

loss function A loss function is a map

$$L : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \rightarrow \mathbb{R}_+ : \big((\mathbf{x}, y), h\big) \mapsto L\left((\mathbf{x}, y), h\right)$$

which assigns a pair consisting of a data point, with features $\mathbf{x}$ and label y, and a hypothesis $h \in \mathcal{H}$ the non-negative real number $L\left((\mathbf{x}, y), h\right)$. The loss value $L\left((\mathbf{x}, y), h\right)$ quantifies the discrepancy between the true label y and the predicted label $h(\mathbf{x})$. Smaller (closer to zero) values $L\left((\mathbf{x}, y), h\right)$ mean a smaller discrepancy between predicted label and true label of a data point. Figure 1 depicts a loss function for a given data point, with features $\mathbf{x}$ and label y, as a function of the hypothesis $h \in \mathcal{H}$. 1, 4, 5

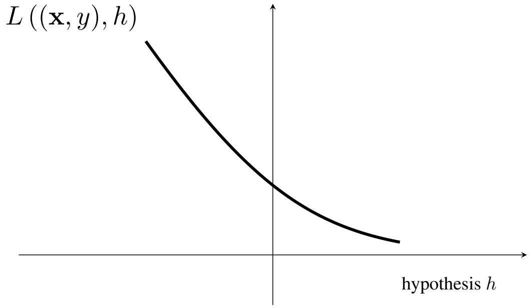

Fig. 1: Some loss function $L\left((\mathbf{x}, y), h\right)$ for a fixed data point, with feature vector $\mathbf{x}$ and label y, and varying hypothesis h. ML methods try to find (learn) a hypothesis that incurs minimum loss.

model We use the term model as a synonym for hypothesis space 2, 3

model parameters Model parameters are numbers that select a hypothesis map out of a hypothesis space. 2

prediction A prediction is an estimate or approximation for some quantity of interest. ML revolves around learning or finding a hypothesis map h that reads in the features $\mathbf{x}$ of a data point and delivers a prediction $\widehat{y} := h(\mathbf{x})$ for its label y. 4, 5

probabilistic model A probabilistic model interprets data points as realizations of RVs with a joint probability distribution.This joint probability distribution typically involves parameters which have to be manually chosen (=design choice) or learnt via statistical inference methods [15]. 3–5

probability density function (pdf) The probability density function (pdf) $p(x)$ of a real-valued RV $x \in \mathbb{R}$ is a particular representation of its probability distribution. If the pdf exists, it can be used to compute the probability that x takes on a value from a (measurable) set $\mathcal{B} \subseteq \mathbb{R}$ via $p(x \in \mathcal{B}) = \int_{\mathcal{B}} p(x')dx'$ [16, Ch. 3]. The pdf of a vector-valued RV $\mathbf{x} \in \mathbb{R}^d$ (if it exists) allows to compute the probability that $\mathbf{x}$ falls into a (measurable) region $\mathcal{R}$ via $p(\mathbf{x} \in \mathcal{R}) = \int_{\mathcal{R}} p(\mathbf{x}')dx'_1 \ldots dx'_d$ [16, Ch. 3]. 3

probability distribution The data generated in some ML applications can be reasonably well modelled as realizations of a RV. The overall statistical properties (or intrinsic structure) of such data are then governed by the probability distribution of this RV. We use the term probability distribution in a highly informal manner and mean the collection of probabilities assigned to different values or value ranges of a RV. The probability distribution of a binary RV $y \in \{0, 1\}$ is fully specified by the probabilities $p(y = 0)$ and $p(y = 1)\big(= 1 - p(y = 0)\big)$. The probability distribution of a real-valued RV $x \in \mathbb{R}$ might be specified by a probability density function $p(x)$ such that $p(x \in [a, b]) \approx p(a)|b - a|$. In the most general case, a probability distribution is defined by a probability measure [17, 18]. 3, 5

random variable (RV) A random variable is a mapping from a probability space $\mathcal{P}$ to a value space [18]. The probability space, whose element are elementary events, is equipped with a probability measure that assigns a probability to subsets of $\mathcal{P}$. A binary random variable maps elementary events to a set containing two different value, such as $\{-1, 1\}$ or $\{\text{cat}, \text{no cat}\}$. A real-valued random variable maps elementary events to real numbers $\mathbb{R}$. A vector-valued random variable maps elementary events to the Euclidean space $\mathbb{R}^d$. Probability theory uses the concept of measurable spaces to rigorously define and study the properties of (large) collections of random variables [17, 18]. 2–5

realization Consider a RV x which maps each element (outcome, or elementary event) $\omega \in \mathcal{P}$ of a probability space $\mathcal{P}$ to an element a of a measurable space $\mathcal{N}$ [18, 5, 19]. A realization of x is any element $a' \in \mathcal{N}$ such that there is an element $\omega' \in \mathcal{P}$ with $x(\omega') = a'$. 5

regularization Regularization techniques modify the ERM principle such that the learnt hypothesis performs well (generalizes) beyond the training set. One specific implementation of regularization is to add a penalty or regularization term to the objective function of ERM (which is the average loss on the training set). This regularization term can be interpreted as an estimate for the increase in the expected loss (risk) compared to the average loss on the training set. 3, 5

regularized empirical risk minimization Synoym for structural risk minimization (SRM). 3

risk Consider a hypothesis h that is used to predict the label y of a data point based on its features $\mathbf{x}$. We measure the quality of a particular prediction using a loss function $L\left((\mathbf{x}, y), h\right)$. If we interpret data points as the realizations of i.i.d. RVs, also the $L\left((\mathbf{x}, y), h\right)$ becomes the realization of a RV. Using such an i.i.d. assumption allows to define the risk of a hypothesis as the expected loss $\mathbb{E}\big\{L\left((\mathbf{x}, y), h\right)\big\}$. Note that the risk of h depends on both, the specific choice for the loss function and the probability distribution of the data points. 5

statistical aspects By statistical aspects of a ML method, we refer to (properties of) the probability distribution of its ouput given a probabilistic model for the data fed into the method. 4

structural risk minimization Structural risk minimization is the problem of finding the hypothesis that optimally balances the average loss (or empirical risk) on a training set with a regularization term. The regularization term penalizes a hypothesis that is not robust against (small) perturbations of the data points in the training set. 5

test set A set of data points that have neither been used in a training set to learn parameters of a model nor in a validation set to choose between different models (by comparing validation errors). 1

training error The average loss of a hypothesis when predicting the labels of data points in a training set. We sometimes refer by training error also the minimum average loss incurred on the training set by any hypothesis out of a hypothesis space. 1, 5

training set A set of data points that is used in ERM to learn a hypothesis $\hat{h}$. The average loss of $\hat{h}$ on the training set is referred to as the training error. The comparison between training error and validation error of $\hat{h}$ allows to diagnose ML methods and informs how to improve them (e.g., using a different hypothesis space or collecting more data points). 1, 3, 5, 6

validation Consider a hypothesis $\hat{h}$ that has been learn via ERM on some training set $\mathcal{D}$. Validation refers to the practice of trying out a hypothesis $\hat{h}$ on a validation set that consists of data points that are not contained in the training set $\mathcal{D}$. 1

validation error Consider a hypothesis $\widehat{h}$ which is obtained by ERM on a training set. The average loss of $\widehat{h}$ on a validation set, which is different from the training set, is referred to as the validation error. 1, 5

validation set A set of data points that has not been used as training set in ERM to train a hypothesis $\widehat{h}$. The average loss of $\widehat{h}$ on the validation set is referred to as the validation error and used to diagnose the ML method. The comparison between training and validation error informs adaptations of the ML method (such as using a different hypothesis space). 1, 5, 6