The Impact of Record Linkage on Learning from Feature Partitioned Data

Richard Nock [1] **Stephen Hardy** [2] **Wilko Henecka** [2] **Hamish Ivey-Law** [3] **Jakub Nabaglo** [3] **Giorgio Patrini** [4] **Guillaume Smith** [2] **Brian Thorne** [5]

Abstract

There has been recently a significant boost to machine learning with distributed data, in particular with the success of federated learning. A common and very challenging setting is that of *vertical* or *feature partitioned* data, when multiple data providers hold different features about common entities. In general, training needs to be preceded by record linkage (RL), a step that finds the correspondence between the observations of the datasets. RL is prone to mistakes in the real world. Despite the importance of the problem, there has been so far no formal assessment of the way in which RL errors impact learning models. Work in the area either use heuristics or assume that the optimal RL is known in advance. In this paper, we provide the first assessment of the problem for supervised learning. For wide sets of losses, we provide technical conditions under which the classifier learned after noisy RL converges (with the data size) to the best classifier that would be learned from mistake-free RL. This yields new insights on the way the pipeline RL + ML operates, from the role of large margin classification on dampening the impact of RL mistakes to clues on how to further optimize RL as a preprocessing step to ML. Experiments on a large UCI benchmark validate those formal observations.

1. Introduction

The past few years have seen a very steep increase of the use of Machine Learning (ML) in the context of Federated Learning (Kairouz et al., 2021), a setting characterized by decentralized data over peers or clients and privacy constraints for training. Experimental and theoretical challenges abound, some of which are relevant beyond the privacy realm of federated learning. One such problem is:

"when is a global trained model better ?"

In the context of learning from distributed data, approaches can be segmented in terms of (a) whether the data is vertically or horizontally partitioned and (b) the family of models being learned. The majority of previous work on (secure) distributed learning considers a horizontal data partitioning in which data providers record the same features for different entities or observations, yet real-world cases now abound where the distribution of data rather fits to the vertical partition (or feature partitioned data) setting (Gu et al., 2020a). In this case, data providers record *different* features for the same entities / observations. This setting is more challenging than the horizontal one (Gascón et al., 2017), since it requires finding the *correspondence between rows* across providers to create examples that span all features, and this needs to be done before learning. The broad family of related techniques are called *Record Linkage* (RL, or entity matching, entity resolution, Christen (2012)). Error-free record linkage is often not available in the real-world for two main reasons. Firstly, unintentional noise may affect the quality of the linkage. Case studies report that exact matching can be very damaging when identifiers are not stable and error-prone: 25% of true matches would have been missed by exact matching in a census operation (Schnell, 2013; Winkler, 2009). The second reason is privacy: altering records with calibrated noise is a way to achieve specific privacy requirements (Christen, 2016; Kairouz et al., 2021).

While linking vertically partitioned data provides more features that should ultimately result in better ML models, mistakes during record linkage can impair data and thereby negatively affect ML models. Precisely quantifying the impact of record linkage on ML models is a non-trivial and important open problem (Kairouz et al., 2021, Section 3), all the more as RL algorithms now abound (Christophides et al., 2020): abstracting and understanding their impact at a high level on the RL + ML pipeline is crucial. Existing approaches are either heuristic (Kang et al., 2020) or make the assumption that the solution to RL is known *a priori* (Gascón et al., 2017; Gu et al., 2020a;b), which would often be violated in the real world (Hernández & Stolfo, 1998).

Our formal contribution provides a detailed coverage of this problem for linear models, that are a key component to federated learning (Gu et al., 2020a). Figure 1 presents a

[1]Google Research (Brain team) [2]Ambiata [3]The Australian National University [4]Sensity [5]HardByte. Correspondence to: Richard Nock <richardnock@google.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

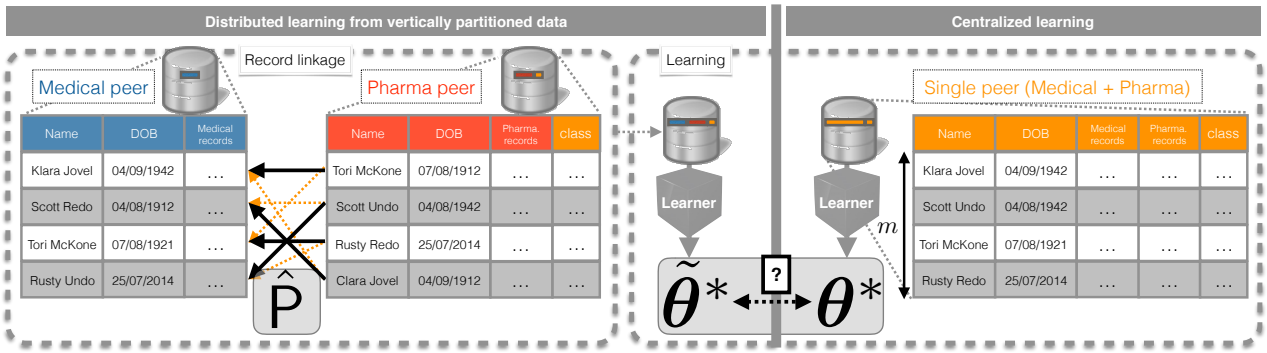


Figure 1: In a centralized learning setting (right pane), a single peer holds medical and pharmaceutical records and then learns a classifier $\boldsymbol{\theta}^*$ from this data to predict a class. In a distributed vertically partitioned learning setting (left pane), data features are split among peers (here, a medical and a pharmaceutical peer). At least one peer has the class feature. There exists an ideal mapping which allows us to reconstruct the same database as in the centralized setting (dashed orange arrows), but because some entries in the peers' databases are noisy, our linkage (**black** arrows) is different. The mismatch between the ideal linkage and ours can be represented by an unknown permutation matrix $\hat{\mathsf{P}}$. As a downstream consequence, the classifiers learned in the centralized ($\boldsymbol{\theta}^*$) and distributed ($\tilde{\boldsymbol{\theta}}^*$) settings are different, so a question we address is: what are the conditions on $\hat{\mathsf{P}}$ that guarantee $\tilde{\boldsymbol{\theta}}^* \rightarrow_m \boldsymbol{\theta}^*$, where m is the number of examples (rows)? (Best viewed in color)

sample case with two peers, highlighting the key parameters of our analysis. $\boldsymbol{\theta}^*$ is the optimal centralized model and $\tilde{\boldsymbol{\theta}}^*$ is the model learned in the distributed setting after entity resolution. Given dataset size m (number of rows), we first establish conditions to have for some $\alpha > 0$,

$$\frac{\|\tilde{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_2}{\|\boldsymbol{\theta}^*\|_2} \quad = \quad O\left(\frac{1}{m^\alpha}\right) .$$

There are two main technical conditions to get there: one relies on the size of $\hat{\mathsf{P}}$ and its factorization as elementary permutations, the other on the regularisation of the loss considered. Our theory brings a number of hints on how the RL + ML pipeline operates. A crucial one for RL is that *not all RL errors are equal for ML* and one should try to control not just the overall RL errors, but also the errors mixing observations of different classes. Another is the way *large margin classification on $\boldsymbol{\theta}^*$ implies correct labeling on $\tilde{\boldsymbol{\theta}}^*$*, whereby large margins abolish all negative effect of RL mistakes on classification. We coin this result the immunity of large margin classification to RL mistakes, and believe it brings a very strong justification to learning over vertically partitioned data such as in the context of federated learning, since this setting pools more features for learning and thereby increases further margins.

Our experimental contribution includes testing simple RL algorithms especially accounting for the class information, displaying that margin immunity is indeed observed experimentally and testing RL parameters that our theory predicts to give insights on downstream ML models' errors. Our experiments do not just validate our theory, we believe they provide several key insights on how to optimize and evaluate RL in the context of a ML+RL pipeline, in a field marked by an extreme paucity of formal insights but clearly becoming prominent in the context of federated learning (and also relevant to centralized learning, joining datasets being not just a constraint for distributed learning).

The rest is organised as follows. Section § 2 gives definitions. § 3 presents our formal results, followed by experiments in § 4. § 5 discusses our results and their extension to more losses and models, and a last Section concludes our paper. To save space and for readability, all proofs and complete experiments are postponed to a supplement ('SM').

2. Definitions

Supervised learning, losses — Let $[n] = \{1, 2, ..., n\}$. In ordinary batch supervised learning setting, one is given a set of m examples $\hat{S} \doteq \{(\hat{\boldsymbol{x}}_i, y_i), i \in [m]\}$, where $\hat{\boldsymbol{x}}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ is an observation ($\mathcal{X}$ is called the domain) and $y_i \in \{-1, 1\}$ is a label, or class (the 'hat' notation shall be explained below). Our objective is to learn a linear classifier $\boldsymbol{\theta} \in \mathbb{R}^d$. $\boldsymbol{\theta}$ gives every $\boldsymbol{x} \in \mathcal{X}$ a label equal to $\text{sign}(\boldsymbol{\theta}^\top \boldsymbol{x})$. The goodness of fit of $\boldsymbol{\theta}$ on $\hat{S}$ is measured by a loss function, which we take as a Ridge-regularized Taylor loss. Each such loss ℓ_F is defined by $\ell_F(\hat{S}, \boldsymbol{\theta}; \Gamma) \doteq L + R$ with

$$L \doteq \mathbb{E}_i[F(y_i \boldsymbol{\theta}^\top \hat{\boldsymbol{x}}_i)] \ , \ R \doteq \boldsymbol{\theta}^\top \Gamma \boldsymbol{\theta} \ . \qquad (1)$$

Γ is symmetric positive semi-definite and $F(z) \doteq a + bz + cz^2$ for $a \in \mathbb{R}, b, c \in \mathbb{R}_*$. This definition is very general: we do not assume that F is convex ($c \geq 0$) nor even classifi-

cation calibrated ($b < 0$, Bartlett et al. (2006)). Any twice differentiable loss can locally be approximated by a Taylor loss, and Taylor losses have a longstanding history in federated learning: be it just the square loss Ridge-regularized, (Esperança et al., 2017; Gascón et al., 2017; Giacomelli et al., 2017; Nikolaenko et al., 2013) or a Taylor approximation of a loss (Aono et al., 2016; Djatmiko et al., 2017).

Learning from distributed data — In learning from distributed data (distributed learning for short), $\hat{S}$ is built from separate data-handling sources, called *peers*. In our vertical partition setting, we have two peers A and B, each of which has the description of the m examples on a *subset* of the d features. At least one peer (A by default) has the labels. Matching observations among datasets of A and B involves a step called *record linkage* (Christen, 2012).

Characterization of linkage mistakes — To formally analyze our problem, we assume that the observed dataset $\hat{S}$ is created from an *unknown* dataset $S \doteq \{(\boldsymbol{x}_i, y_i), i \in [m]\}$ whose observation features have been split between A and B. If we define $\mathsf{X} \in \mathbb{R}^{d \times m}$ as the matrix storing (columnwise) observations of S, then each row of X is held by A or B. Without loss of generality (wlog), we assume the 'or' to be exclusive: some features may be observed in both peers, such as "Name" and "DOB" in Fig 1, but either they are useless for ML (*e.g.* because they are unique identifiers) or only one copy would be kept for ML. There is thus both an ideal X and an estimated observation matrix $\hat{\mathsf{X}}$ giving the observations of $\hat{S}$ and built from RL. To understand how the differences between $\hat{\mathsf{X}}$ and X impact learning, we need to drill down into the formalization of $\hat{\mathsf{X}}$. Both matrices can be represented by block matrices on the features of A (X_A) and B (X_B):

$$\mathsf{X} \doteq \left[\begin{array}{c} \mathsf{X}_\mathsf{A} \\ \hline \mathsf{X}_\mathsf{B} \end{array}\right] \quad , \quad \hat{\mathsf{X}} \doteq \left[\begin{array}{c} \mathsf{X}_\mathsf{A} \\ \hline \hat{\mathsf{X}}_\mathsf{B} \doteq \mathsf{X}_\mathsf{B}\hat{\mathsf{P}} \end{array}\right] , \qquad (2)$$

where $\hat{\mathsf{P}} \in \{0,1\}^{m \times m}$ is an unknown permutation matrix capturing the mistakes of linkage. Wlog, the features of A are not affected by linkage. Any permutation matrix can be factored as a product of *elementary* permutation matrices, each swapping two rows/columns of the identity I_m (Bierens, 2004). So, let:

$$\hat{\mathsf{P}} = \prod_{t=1}^{T} \mathsf{P}_t ; \hat{\mathsf{X}}_t \doteq \left[\begin{array}{c} \mathsf{X}_\mathsf{A} \\ \hline \hat{\mathsf{X}}_{t\mathsf{B}} \end{array}\right] , \hat{\mathsf{X}}_{t\mathsf{B}} \doteq \mathsf{X}_\mathsf{B} \prod_{j=1}^{t} \mathsf{P}_j , \forall t \in [T]. \quad (3)$$

Each $\mathsf{P}_\cdot$ denotes an elementary permutation matrix, and T, the *size* of $\hat{\mathsf{P}}$, is unknown. The corresponding sequence $\hat{\mathsf{X}}_0, \hat{\mathsf{X}}_1, ..., \hat{\mathsf{X}}_T$ constructs $\hat{\mathsf{X}}_T = \hat{\mathsf{X}}$ from $\hat{\mathsf{X}}_0 = \mathsf{X}$.

3. A theory for the RL + ML pipeline

Useful parameters of $\hat{\mathsf{P}}$ — We start by the characterisation of the useful parameters of $\hat{\mathsf{P}}$. By switching the B-part of two observations in $\hat{\mathsf{X}}_{t-1}$ to create $\hat{\mathsf{X}}_t$, elementary permutation P_t involves at most four distinct observations from X if we take into account their A-parts as well. Denote as $\boldsymbol{a}, \boldsymbol{b}$ (resp. $\boldsymbol{a}', \boldsymbol{b}'$) the two observations from X whose A-part (resp. B-part) is involved in P_t, where we do not put the t index on observations for readability. For the first elementary permutation involved ($t = 1$) we have $(\boldsymbol{a}, \boldsymbol{b}) = (\boldsymbol{a}', \boldsymbol{b}')$. Figure 2 (Pane **(A)**, left) put those names in context for an elementary permutation P_2. Notice that, as formalized in (2), only the B-part is affected by the permutation. An important parameter of $\hat{\mathsf{P}}$ for our theory is its size, T. Another one is the mistakes each elementary permutation causes wrt X. We model them as follows: if such an elementary permutation happens, it is because the A-parts involved are not too dissimilar from each other, and similarly for the two B-parts involved. This suggests a simple way to quantify the incorrectness of each elementary permutation ($\boldsymbol{w}_\mathsf{F}$ for $\boldsymbol{w} \in \mathbb{R}^d$ denotes subvector of $\boldsymbol{w}$ with features of peer $\mathsf{F} \in \{\mathsf{A}, \mathsf{B}\}$).

Definition 1 *We say that P_t is (ε, τ)-**inexact** for some $\varepsilon, \tau \geq 0, \varepsilon \leq 1$ iff for any $\boldsymbol{w} \in \mathbb{R}^d$,*

$$|(\boldsymbol{a} - \boldsymbol{b})_\mathsf{A}^\top \boldsymbol{w}_\mathsf{A}| \leq \varepsilon \cdot |\boldsymbol{a}^\top \boldsymbol{w}| \vee |\boldsymbol{b}^\top \boldsymbol{w}| + \tau \cdot \|\boldsymbol{w}\|_2 \quad (4)$$

$$|(\boldsymbol{a}' - \boldsymbol{b}')_\mathsf{B}^\top \boldsymbol{w}_\mathsf{B}| \leq \varepsilon \cdot |\boldsymbol{a}'^\top \boldsymbol{w}| \vee |\boldsymbol{b}'^\top \boldsymbol{w}| + \tau \cdot \|\boldsymbol{w}\|_2 (5)$$

where $\vee \doteq \max$. We say that $\hat{\mathsf{P}}$ is (ε, τ)-inexact iff each P_t is (ε, τ)-inexact, $\forall t \in [T]$.

We check that if $\hat{\mathsf{P}}$ is $(0, 0)$-inexact then permutations happen only between identical examples and so we have $\hat{\mathsf{P}} = \mathsf{I}$ wlog, *i.e.* RL makes no mistakes. Figure 2 (Pane **(A)**, left) illustrates Definition 1. Figure 2 (Pane **(A)**, right) illustrates the two types of mistakes that motivate the Definition:

(1) linkage mistakes with unnormalized measures like distances, happening because observations are similar. Suppose that there exists a small $u > 0$ such that $\|(\boldsymbol{a} - \boldsymbol{b})_\mathsf{A}\|_2 \leq u$ and $\|(\boldsymbol{a}' - \boldsymbol{b}')_\mathsf{B}\|_2 \leq u$. In this case, since Cauchy-Schwartz inequality and the fact that norms cannot increase by orthogonal projection bring for example $|(\boldsymbol{a} - \boldsymbol{b})_\mathsf{A}^\top \boldsymbol{w}_\mathsf{A}| \leq u\|\boldsymbol{w}_\mathsf{A}\|_2 \leq u\|\boldsymbol{w}\|_2$, we can easily conclude that P_t is $(0, u)$-inexact;

(2) linkage mistakes with normalized measures or that are 'blind' to norms. In this case, suppose that mistake happen with nearly collinear observations: $(\boldsymbol{a} - \boldsymbol{b})_\mathsf{A} = u \cdot \boldsymbol{a}_\mathsf{A} + \boldsymbol{v}$ and $\boldsymbol{a}_\mathsf{B} = \boldsymbol{0}$ for some small $|u| > 0$ and $\|\boldsymbol{v}\|_2 \leq |u|$. We get $|(\boldsymbol{a} - \boldsymbol{b})_\mathsf{A}^\top \boldsymbol{w}_\mathsf{A}| \leq |u| \cdot |\boldsymbol{a}_\mathsf{A}^\top \boldsymbol{w}_\mathsf{A}| + |\boldsymbol{v}^\top \boldsymbol{w}_\mathsf{A}| \leq |u| \cdot |\boldsymbol{a}^\top \boldsymbol{w}| + |u|\|\boldsymbol{w}\|_2 \leq |u| \cdot |\boldsymbol{a}^\top \boldsymbol{w}| \vee |\boldsymbol{b}^\top \boldsymbol{w}| + |u|\|\boldsymbol{w}\|_2$, so P_t is $(|u|, |u|)$-inexact without making assumptions on the norm of $\boldsymbol{a}_\mathsf{A}$.

It is arguably hard to end up with a theory for RL + ML that fully covers the zoo of RL techniques, yet ours does so for a

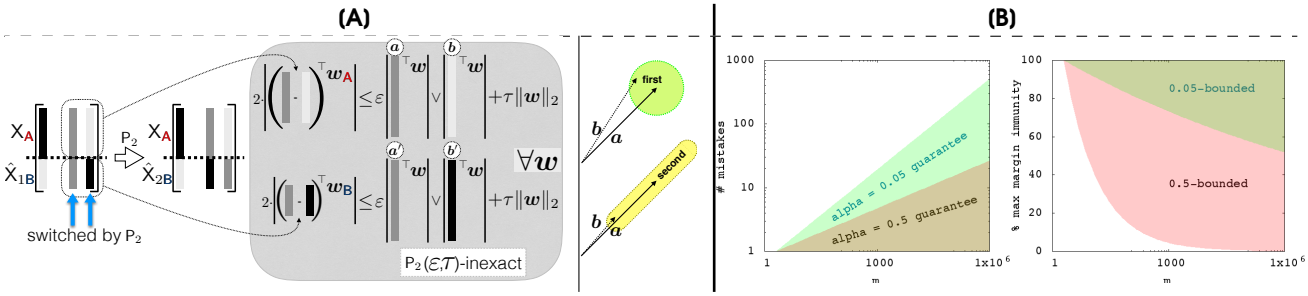


Figure 2: Pane **(A)**: *Left*: illustration of Definition 1 for elementary permutation P_2. Notations $\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{b}, \boldsymbol{b}'$ refer to the definition (in this case, $\boldsymbol{a} = \boldsymbol{a}'$). *Right*: Depiction of the two kinds of record linkage inaccuracies that motivated Definition 1, showing when a vector $\boldsymbol{b}$ can be considered 'similar' to a vector $\boldsymbol{a}$. Pane **(B)**: *Left*: number of mistakes authorized for $\hat{\mathsf{P}}$ to be guaranteed α-bounded for two values of α. *Right*: proportion (range) of the maximal margin $\delta_{\boldsymbol{\theta}}$ on $\boldsymbol{\theta}^*$ which guarantees immunity to RL mistakes, for two values of α, assuming $\rho = 0$ (See text for details).

substantial part via Definition 1: if the input of RL lives in a vector space, such as for numeric data or the latent space of deep RL techniques, using a distance for matching directly corresponds to **(1)** (Mudgal et al., 2018), while angle-based measures are relevant to **(2)** (Getoor & Machanavajjhala, 2013; Mudgal et al., 2018). Set measures can also directly be embedded via binary representations and *e.g.* the use of Hamming distance: Jaccard or Dice belong to **(2)** (Getoor & Machanavajjhala, 2013), while stripping off normalization would end up in **(1)**, etc. . String edit distances can be addressed via deep embeddings (Gómez et al., 2017).

There is a huge $\Omega(2^m)$ number of ways to build $\hat{\mathsf{P}}$ from subsets of elementary permutations, but only an $O(m^2)$ number of ways to do it with the smallest number of elementary permutations, which furthermore will always satisfy $T \leq m - 1$. The trick is simple: starting from X, we first pick an observation whose B-part in $\hat{\mathsf{X}}$ is different and perform the corresponding elementary permutation. We then remove this observation and repeat the process over the remaining observations until all mistakes are done. Denote $X_* \doteq \max_i \|\boldsymbol{x}_i\|_2$ the max column norm in X and

$$\xi(\hat{\mathsf{P}}) \quad \doteq \quad \min \left\{ \varepsilon + \frac{\tau}{X_*} : \hat{\mathsf{P}} \text{ is } (\varepsilon, \tau)\text{-inexact} \right\} \ . \quad (6)$$

Hereafter, we write ξ for short without reference to $\hat{\mathsf{P}}$ and suppose wlog that $\xi > 0$. We use a different normalisation of ε and τ in ξ to account for the different scales in their factors: indeed, $\varepsilon \cdot |\boldsymbol{a}^\top \boldsymbol{w}| \leq \varepsilon \cdot X_* \|\boldsymbol{w}\|_2$ while $\tau \cdot \|\boldsymbol{w}\|_2 = (\tau / X_*) \cdot X_* \|\boldsymbol{w}\|_2$. There is a trivial upperbound on ξ, as shown in the following Lemma.

Lemma 2 *$\forall t \in [T]$, P_t is $(0, 2X_*)$-inexact. So, $\xi \leq 2$.*

The proof stems from the fact that for example $(\boldsymbol{a} - \boldsymbol{b})_{\mathsf{A}}^\top \boldsymbol{w}_{\mathsf{A}} \leq (\|\boldsymbol{a}\|_2 + \|\boldsymbol{b}\|_2)\|\boldsymbol{w}\|_2 \leq 2X_* \|\boldsymbol{w}\|_2$ (and then choosing $\varepsilon = 0$). There is a quantity dependent on ξ that shall be important in our results:

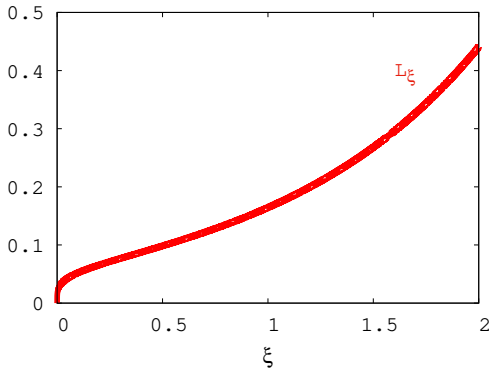


Figure 3: Plot of function L_ξ as a function of ξ in (7).

$$L_\xi \quad \doteq \quad \frac{\xi^{\frac{1}{4}}(\exp(\xi) + 1)}{16\sqrt{2}}. \quad (7)$$

We observe $L_\xi < 1/2$ for $\xi \in [0, 2]$, and L_ξ can be much smaller for smaller values of ξ, see Figure 3. Lemma 2 shows that ξ is never large, but small(er) values shall sometimes be desirable for the following assumption to hold.

Assumption 1 *$\hat{\mathsf{P}}$ is α-**bounded** for some $0 < \alpha \leq 1$ iff its size satisfies $T \leq \left(m / \xi^{1/4}\right)^{\frac{1-\alpha}{2}}$, where ξ is defined in* (6).

The larger α, the better it is for our theory to follow. If T is small enough, $\hat{\mathsf{P}}$ is guaranteed to be α-bounded for *some* $\alpha > 0$. We show in Figure 2, Pane **(B)** (left) the range of mistakes authorized, as a function of m, for $\hat{\mathsf{P}}$ to be α-bounded for two values of α.

We let $T_+ \leq T$ denote the number of *class mismatches* in elementary permutations, that is, the number of times

the class of observation $\boldsymbol{a}$ is different from the class of observation $\boldsymbol{b}$ (see for example Figure 2, Pane **(A)**, left). Let

$$\rho \doteq (T_+/T) \in [0,1] \tag{8}$$

denote the proportion of such elementary permutations.

Key parameters for our results — Quite remarkably, all our results depend on two parameters only, each characterizing a distinct unknown: the ideal classifier $\boldsymbol{\theta}^*$learned on S ($\delta_{\boldsymbol{\theta}}$) and the unknown permutation $\hat{\mathsf{P}}$ ($\delta_{\hat{\mathsf{P}},\ell}$):

$$\boxed{\delta_{\boldsymbol{\theta}} \doteq \|\boldsymbol{\theta}^*\|_2 X_* \quad , \quad \delta_{\hat{\mathsf{P}},\ell} \doteq \frac{\rho|b|L_\xi}{|c|}.}$$

Notice that $\delta_{\hat{\mathsf{P}},\ell}$ also aggregates loss parameters. These can globally be seen as penalties — the smaller they are, the less impact has $\hat{\mathsf{P}}$ on learning. The most important for learning appears to be $\delta_{\hat{\mathsf{P}},\ell}$ and when $\hat{\mathsf{P}}$ is 'good enough' that $\rho = 0$ (no linkage mistakes *between* classes), we have $\delta_{\hat{\mathsf{P}},\ell} = 0$, which appears to bring substantially better bounds. We let $\mu \doteq (1/m) \cdot \|\hat{\mathsf{X}}\|_F^2$ ($= \|\mathsf{X}\|_F^2$), where $\|.\|_F$ denotes Frobenius norm, and $\lambda^\circ(.)$ denote the smallest eigenvalue.

Assumption 2 *We say that the **data-loss calibration** assumption holds iff the two constraints are satisfied:*

(a) Maxnorm regularization: regularizor's Γ *satisfies* (b, c *are the Taylor loss parameters)*

$$0 \leq \frac{X_*^2}{c\mu + \lambda^\circ(\Gamma)} \leq \frac{1}{2} \cdot \min\left\{\frac{1}{|b|}, \frac{1}{4|c|}\right\}, \tag{9}$$

(b) Minimal size: $m \geq 4 \cdot \max\{1, \xi \cdot \max\{1, 2|c|/|b|\}\}$.

Condition (a) implies the full Taylor loss to be strictly convex and sufficiently regularized, with essentially $\lambda^\circ(\Gamma) = \Omega(X_*^2)$, but the hidden constant can be fairly small depending on the loss parameters (in particular if F is a proper loss in eq. (1), more in Section 5). Condition (b) just postulates that m is larger than a small constant. Note that both (a) and (b) can be checked for observed data, as for example X_* is bounded by twice the max observed norm.

Main result — We now show how $\tilde{\boldsymbol{\theta}}^*$ deviates from $\boldsymbol{\theta}^*$.

Theorem 3 *Suppose the data-loss calibration assumption holds. Then we have:*

$$\frac{\|\tilde{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_2}{\|\boldsymbol{\theta}^*\|_2} \leq \frac{\xi^{\frac{1}{4}}}{m} \cdot T^2 \cdot \left(\xi^{\frac{3}{4}} + \frac{\delta_{\hat{\mathsf{P}},\ell}}{\delta_{\boldsymbol{\theta}}}\right) . \tag{10}$$

If furthermore $\hat{\mathsf{P}}$ *is* α*-bounded, then*

$$\frac{\|\tilde{\boldsymbol{\theta}}^* - \boldsymbol{\theta}^*\|_2}{\|\boldsymbol{\theta}^*\|_2} \leq C(m) \cdot \left(\xi^{\frac{3}{4}} + \frac{\delta_{\hat{\mathsf{P}},\ell}}{\delta_{\boldsymbol{\theta}}}\right), \tag{11}$$

with $C(m) \doteq \left(\xi^{1/4}/m\right)^\alpha$.

(proof in SM, Section I.2) Theorem 3 calls for the following remarks:

- ▷ the drift between optimal classifiers *vanishes with* m *if* $\hat{\mathsf{P}}$ is 'small' enough to be α-bounded. More, if *e.g.* $T = o(\sqrt{m})$ then we get $C(m) \to_{+\infty} 0$ *regardless* of the algorithm chosen for record linkage. Otherwise, to get ξ small enough for $\hat{\mathsf{P}}$ to be α-bounded, this suggests to use features in RL that are also discriminative for ML;
- ▷ Hindrances to convergence can appear when $\delta_{\boldsymbol{\theta}} \ll \delta_{\hat{\mathsf{P}},\ell}$, such as when $\|\boldsymbol{\theta}^*\|_2$ is small. To prevent that, one can focus on decreasing linkage mistakes between classes (ρ), which results in decreasing $\delta_{\hat{\mathsf{P}},\ell}$. This suggests to design record linkage algorithms that use class information to link datasets. Notice that tweaking the loss to decrease $|b|/|c| = |(F'/F'')(0)|$ in $\delta_{\hat{\mathsf{P}},\ell}$ does not achieve the goal as this potentially decreases $\|\boldsymbol{\theta}^*\|_2$ as well.

The influence of large margin classification on $\boldsymbol{\theta}^*$— We now show that large margin classification on $\boldsymbol{\theta}^*$ *guarantees* right classification on $\tilde{\boldsymbol{\theta}}^*$ provided m is large enough, a phenomenon we define as *immunity* to record linkage.

Definition 4 *Fix* $\kappa > 0$. $\tilde{\boldsymbol{\theta}}^*$ *is **immune** to record linkage at margin* κ *iff* $\forall(\boldsymbol{x}, y), \left(y{\boldsymbol{\theta}^*}^\top \boldsymbol{x} > \kappa\right) \Rightarrow \left(y{\tilde{\boldsymbol{\theta}}^*}^\top \boldsymbol{x} > 0\right)$.

Hence, $(\boldsymbol{x}, y)$ receives the right class by both $\boldsymbol{\theta}^*$ and $\tilde{\boldsymbol{\theta}}^*$ — the corresponding margin may vary however.

Theorem 5 *If the data-loss calibration assumption holds then for any* $\kappa > 0$, $\tilde{\boldsymbol{\theta}}^*$ *is immune to record linkage at margin* κ *as long as:*

$$\frac{m}{\xi^{\frac{1}{4}} T^2} > \frac{\xi^{\frac{3}{4}}\delta_{\boldsymbol{\theta}} + \delta_{\hat{\mathsf{P}},\ell}}{\kappa} . \tag{12}$$

If furthermore $\hat{\mathsf{P}}$ *is* α*-bounded,* $\tilde{\boldsymbol{\theta}}^*$ *is immune to record linkage at margin* κ *if*

$$m > \xi^{\frac{1}{4}} \cdot \left(\frac{\xi^{\frac{3}{4}}\delta_{\boldsymbol{\theta}} + \delta_{\hat{\mathsf{P}},\ell}}{\kappa}\right)^{\frac{1}{\alpha}} . \tag{13}$$

(proof in SM, Section I.3) Theorem 5 is interesting for the relationships between m (data), ξ (permutation) and κ (margin) to achieve immunity 'at scale' by controlling permutation mistakes between classes. In particular,

- ▷ if $\delta_{\hat{\mathsf{P}},\ell} \ll \delta_{\boldsymbol{\theta}}$, such as when RL mistakes between classes (ρ) are small enough, the smallest possible margin κ at which immunity holds converges to zero at rate $1/m^\alpha$.

Setting	Labels: on peer A	on peer B	used (RL)
GREEDYRL	✗	✗	✗
GREEDYRL+$\overline{\text{C}}$	✓	✗	✓
GREEDYRL+$\tilde{\text{C}}$	✓	Noisy/✓	✓

Table 1: Baselines for RL with respect to using labels.

Indeed, the maximal optimal margin is bounded by $\delta_{\boldsymbol{\theta}}$ by Cauchy-Schwartz inequality. Theorem 5 says that if $\delta_{\hat{\mathsf{P}},\ell} \ll \delta_{\boldsymbol{\theta}}$, picking $\kappa \doteq \delta \cdot \delta_{\boldsymbol{\theta}}$ for $0 < \delta < 1$ brings immunity at margin κ if $\delta = \Omega(C(m))$ where $C(m)$ is defined in Theorem 3, so the lowest possible margin from which immunity holds indeed converges to zero at rate $1/m^{\alpha}$. We can indeed get $\delta_{\hat{\mathsf{P}},\ell} \ll \delta_{\boldsymbol{\theta}}$ if ρ sufficiently small since $\delta_{\hat{\mathsf{P}},\ell} \propto \rho$.

To illustrate this, Figure 2, Pane **(B)** (right) displays the proportion of the max margin $\delta_{\boldsymbol{\theta}}$ from which we are guaranteed immunity as a function of m, for two values of α if $\rho = 0$, using the conservative upperbound $\xi \leq 2$. As α increases, we are guaranteed immunity for a larger range. *E.g.* if $\alpha = 0.5, m = 10^6$, then all examples with margin $\geq 0.002 \cdot \delta_{\boldsymbol{\theta}}$ on $\boldsymbol{\theta}^*$ receive the right class on $\tilde{\boldsymbol{\theta}}^*$.

4. Experiments

Our experiments have been designed to test three key findings: (i) the influence of RL mistakes between classes on ML models, (ii) margin immunity to RL mistakes and (iii) how RL parameters ξ, T offer insights on both train and test ML errors. The related sections are mirrored in the SM.

Setting — We consider the realistic setting in which there exists a small set of features that is present in both peers A and B. We call them the *shared features* and use them for RL. This setting is realistic considering, for example, that many businesses or government bodies would share basic information about their customers (such as gender, postal code, age, contact number, etc.) (Patrini et al., 2016). We then put noise in those shared features as a slider to vary the hardness of the task. Noise injection follows standard modelling for noise in the field where given probability p, a value is replaced by a 'neighboring' value, where the neighboring relationship is determined by the attribute domain (Christen & Pudjijono, 2009). We use a similarity measure between observed shared vectors based on the cosine similarity, which is simple and standard among token-based RL approaches. Given observations from A and B, the cosine similarity $\widetilde{\cos}$ between them is $\widetilde{\cos}(.,.) \doteq 1 + \cos(.,.)$[1], where arguments are the subvectors of shared values. ML models obtained from AdaBoost (Schapire & Singer, 1999).

Domains — We used 15 UCI domains (Blake et al., 1998) from which we have generated our distributed data using the following process: given a set of shared features, split randomly the remaining features between A and B. The shared features of B are then noisified using the process sketched above. A always has access to the classes. Some UCI domains have features that would be a natural fit for shared features: for UCI creditcard, we have used sex, education, marriage, age as shared features. The complete list of domains, inclusive of shared attributes used and statistics, is given in SM (Table A1). For some UCI domains, we have considered two simulations, one in which the shared attributes are highly correlated with the class (H) and one in which they are not (L, Table 2). The proportion of linkage errors between classes for the class agnostic GREEDYRL (see below) goes up to $\approx 25\%$, which is very significant, and the proportion of shared features ranges from $\approx 3\%$ to $> 30\%$.

Algorithms and baselines for RL — Notation $\mathcal{I}$ shall refer to subsets of $[m]^2$ where first arguments are indexes in A and second arguments are indexes in B. Any solution to RL is a set $\mathcal{I}'$ of cardinal m. The algorithms we design use a simple and efficient subroutine we call GREEDY($\mathcal{I}$), where $\mathcal{I} \subseteq [m]^2$ is a set of couple of indexes from A $\times$ B: starting from $\mathcal{I}_{\text{g}} = \emptyset$, put couple $(i^*_{\mathsf{A}}, i^*_{\mathsf{B}}) \in \mathcal{I}$ whose indexed observations have largest $\widetilde{\cos}$ in $\mathcal{I}_{\text{g}}$, remove $i^*_{\mathsf{A}}, i^*_{\mathsf{B}}$ from $\mathcal{I}$, and repeat until $\mathcal{I} = \emptyset$. Finally, return $\mathcal{I}_{\text{g}}$. GREEDY has $1/2$-approximation of the optimum (Avis, 1983, Theorem 4). The baselines below are summarized in Table 1.

Our first algorithm for RL, GREEDYRL, is a simple use of GREEDY without class information: we let $\mathcal{I}_{\text{g}} \leftarrow$GREEDY($[m]^2$) and then perform RL using $\mathcal{I}_{\text{g}}$.

Our second algorithm, GREEDYRL+$\tilde{\text{C}}(p')$, corresponds to the case where both parties have knowledge of the class label, and the same proportion of positive examples, eventually noisified. This is a 'vertical-partition-amenable' version of the real world setting where parties would have prior knowledge of such a population-wide proportion, with then eventual individual mismatches (*e.g.* positive examples of A being negative in B and *vice versa*). We simulate *permutation noise* over labels in B: we randomly permute a random positive class and a random negative class for $p'm$ iterations in B, where p' is the noise parameter. We consider $p' \in \{0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.15, 0.2\}$. $p' = 0.2$ can be considered a fairly large noise. The ER algorithm is then straightforward: let $\mathcal{I}^+ \subseteq [m]^2$ (resp. $\mathcal{I}^-$) denote all couples of positive (resp. negative) examples indexes among peers. We then run $\mathcal{I}^+_{\text{g}} \leftarrow$GREEDY($\mathcal{I}^+$), $\mathcal{I}^-_{\text{g}} \leftarrow$GREEDY($\mathcal{I}^-$) and perform RL using $\mathcal{I}^+_{\text{g}}$ and $\mathcal{I}^-_{\text{g}}$.

Our third algorithm, GREEDYRL+$\overline{\text{C}}$, is a setting where B does *not* have classes. We proceed by a four-step approach in which B learns classes based on shared features:

▷ GREEDYRL+$\overline{\text{C}}$:

(i) Let $\mathcal{I}_{\text{g}} \leftarrow$ GREEDYRL; discard couples in $\mathcal{I}_{\text{g}}$ whose

[1] '+1' used to get a non-negative similarity measure.

Domain	Noise p	'Ideal'	GREEDYRL[as is | +$\overline{\text{C}}$ | +$\tilde{\text{C}}$]														
			as is	+$\overline{\text{C}}(k)$				+$\tilde{\text{C}}(p')$									
				1	2	5	10	0	0.01	0.02	0.03	0.04	0.05	0.10	0.15	0.20	
phishing$_H$	0.05	8.03	8.40	8.08	8.18	8.70	8.62	8.17	8.05	8.05	8.14	8.23	**8.95	**10.84	**12.85	**15.41	
	0.1	7.92	8.35	8.23	8.16	8.36	8.51	7.92	*7.76	8.01	8.21	8.72	8.76	**9.61	**12.86	**15.29	
	0.3	7.96	8.90	9.15	8.91	9.01	8.86	8.39	*8.09	*8.17	8.46	8.67	9.05	**11.49	**13.5	**15.46	
creditcard	0.05	23.26	23.26	23.26	23.26	23.26	23.26	23.26	23.26	23.26	23.26	23.26	23.26	23.26	23.26	23.26	
	0.1	23.26	41.87	40.66	41.46	*36.91	36.89	**23.26	**23.26	**23.26	**23.26	**23.26	*26.19	42.65	43.08	44.36	
	0.3	23.26	42.49	41.19	42.03	*38.82	*36.51	**23.26	**23.26	**23.26	*24.72	*25.01	*32.28	40.87	41.75	40.89	
firmteacher	0.05	12.45	17.57	18.03	18.23	17.75	18.00	**12.71	**12.68	**12.71	*13.06	**13.02	**13.35	*14.81	*15.90	17.38	
	0.1	12.39	21.03	21.06	21.29	21.51	21.54	**12.89	**12.71	**12.73	**12.72	**13.14	**13.36	**14.82	*16.98	*18.06	
	0.3	12.35	20.45	*21.12	*21.16	20.32	20.34	**12.54	**12.45	**12.42	**12.73	**12.81	**13.00	**14.54	**16.05	**17.44	
phishing$_L$	0.05	7.97	14.80	14.82	14.99	15.02	14.98	**7.91	**8.27	**8.44	**8.45	**8.61	**8.83	**9.94	**10.16	**11.18	
	0.1	7.89	11.11	11.11	11.11	11.11	11.11	**8.02	**7.92	**7.82	**7.82	**7.91	**8.11	**8.50	**9.32	*10.65	
	0.3	7.91	13.73	13.73	13.73	13.73	13.73	**8.29	**8.51	**8.47	**8.44	**8.60	**8.80	**9.16	**9.81	**10.54	

Table 2: Results (extract, test errors) comparing, for three values of the shared features noise (p), the approaches built on top of GREEDYRL to 'Ideal'. Full results over all UCI domains provided in SM, Table A2. Red denote results that are statistically outperformed by GREEDYRL; Green denote results of GREEDYRL[+$\overline{\text{C}}$ | +$\tilde{\text{C}}$] statistically better than greedyER. One star ($*$) indicated p-value in $(10^{-6}, 10^{-2}]$, two stars ($**$) indicated p-value $\leq 10^{-6}$ (best viewed in color).

similarity is below the median; assign labels in B for the remaining couples in $\mathcal{I}_g$ using labels in A;

(ii) complete labelling in B using a k-NN algorithm based its labels obtained from (i);

(iii) run GREEDYRL+$\tilde{\text{C}}$ on labeled data;

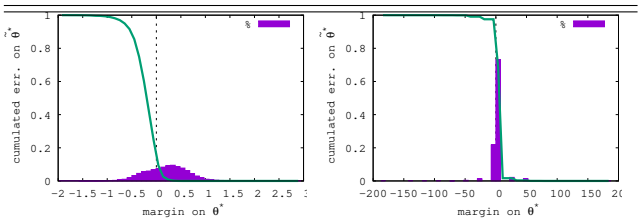
(iv) link the remaning observations using GREEDYRL.

Step (iv) is mandatory as in general we are not guaranteed that labels match in number between classes.

Our last baseline is just the ideal RL: since our vertical partition setting is simulated, we consider the perfect linkage that provides the true training sample S for ML. We call this baseline IDEAL to compare against the RL approaches.

The class is key to optimizing RL — Extracts of our results are displayed in Table 2 (Complete results in SM, Table A2). Several observations come to the fore. First, as GREEDYRL makes more linkage mistakes on observations from different classes, the more beneficial it is to use the class information for RL. On domains firmteacher, phishing, using the class information is almost always on par with or (significantly) better than GREEDYRL. Second, the improvement can be extremely significant as witnessed by domains creditcard or firmteacher, with almost 20 % improvement when using (even noisy) classes on creditcard, and still up to 6% improvement when using predicted classes (GREEDYRL+$\overline{\text{C}}$) on creditcard. This is good news because the shared features we used on creditcard — sex, education, marriage, age — are typically those that could be shared in a federated learning setting. Table 2 also displays that RL+ML can be competitive even against IDEAL. The phishing domain displays the two experiments using shared attributes that are highly correlated (H) or not (L) with class. Predictably, using class information brings the biggest edge when shared features are not correlated with class.

The approach that learns classes in GREEDYRL+$\overline{\text{C}}$ is simple but still manages to deliver significant improvements in some cases, typically high noise for shared features (creditcard) or shared features sufficiently correlated with class. Finally, we keep in mind that these results are obtained for simulations that include in general a small number of shared features (2.8 on average) and a shared feature noise that ranges up to p =30%, which would correspond to relatively challenging practical settings. This suggests that there would be for most domains good reasons to carry out tailored approaches to record linkage for learning with the ambition to challenge the unknown learner having access to the ideally linked data. This is no surprise: it is known that



winewhite GREEDYRL+$\overline{\text{C}}$(5) | creditcard GREEDYRL+$\overline{\text{C}}$(10)

Figure 4: Margin distribution on two domains with shared attribute noise $p = 0.3$. The magenta histogram displays the distribution of margins of $\boldsymbol{\theta}^*$ on training. The green curve is the cumulated relative error of $\tilde{\boldsymbol{\theta}}^*$ *above* some margin x relative to $\boldsymbol{\theta}^*$. For example, on winewhite, less than 20% of the errors on training happen on examples with positive margin, and approximately *no* error happens on examples with positive margin above 0.5 — in other words, *all* examples with margin above 0.5 on $\boldsymbol{\theta}^*$ receive the right class from $\tilde{\boldsymbol{\theta}}^*$ and so, following Definition 4, $\tilde{\boldsymbol{\theta}}^*$ happens to be immune to record linkage at margin 0.5. Since the maximal margin recorded for $\boldsymbol{\theta}^*$ is ≈ 3.0, we see in this example that immunity occurs for a comparatively small positive margin (best viewed in color, see text for details).

the sufficient statistics for the class for linear models is very simple for many losses (Patrini et al., 2014), so we should not expect perfect record linkage to be necessary for ML.

Observation of immunity of large margin classification to RL mistakes – To our knowledge, such a result has never been documented, even experimentally, but it would represent a significant support for federated learning since one can hope, by joining diverse databases, to increase not just the accuracy of classifiers but in fact the optimal margins of $\boldsymbol{\theta}^*$ over examples, thereby bringing immunity to the mistakes of record linkage for examples that would attain sufficiently large margins. But how 'large' a margin? On each domain, we have computed the margin distributions of $\boldsymbol{\theta}^*$ — approximated by the output of AdaBoost ran on the training sample S for twice the usual number of iterations, that is, 2000 (we do this for all cross validation folds). We then compute, for all examples, whether they are given the right class by $\tilde{\boldsymbol{\theta}}^*$. We finally compute the cumulative error distribution, in between 0 and 1, of $\tilde{\boldsymbol{\theta}}^*$. For any $x \in [\kappa_m, \kappa_M]$ (the interval of observed margins), the cumulative error on x is just the proportion of errors occurring for margins in the interval $[x, \kappa_M]$. When $x = \kappa_m$, this is just 1. Figure 4 provides two examples of curves obtained, which does not just validate immunity: on winewhite, it shows that it can happen for a quite small margin (≈ 0.5) with respect to the maximal margin ($\kappa_M \approx 3.0$), which reinforces the support for federated learning. On creditcard, we have $\kappa_M \approx 188$ while immunity happens at margin ≈ 100. Less than 1% of mistakes of $\tilde{\boldsymbol{\theta}}^*$ have margin larger than 30 on $\boldsymbol{\theta}^*$.

RL impact on ML via ξ or T – If our theory empirically stands, then Thm 3 yields reasonable causal dependences:

$$\boxed{\text{RL inaccuracy } \nearrow} \Rightarrow \boxed{\xi, T \nearrow} \Rightarrow \boxed{\text{ML performances } \searrow}$$

ξ and T thus offer simple scalar dials to linking upstream RL and downstream ML. To estimate those, we lowerbound T, the size of P, following the simple linear factorisation trick before (6). In this sequence, we compute an upperbound of ξ for $\varepsilon = 0$ in (6): letting $B \doteq \max_t \max\{\|\boldsymbol{a}_t - \boldsymbol{b}_t\|_2, \|\boldsymbol{a}'_t - \boldsymbol{b}'_t\|_2\}$ where $\boldsymbol{a}_t, \boldsymbol{b}_t, \boldsymbol{a}'_t, \boldsymbol{b}'_t$ are the four observations involved in P_t (Section 3), we then have $\xi \leq B$. Figure 5 provides two example plots (more plots in SM) clearly displaying such a picture: a 'cone' (dashed, left) showing the error range is displayed for ξ (left), the apex clearly located around the min ξ for small dots, while as dots get bigger (signalling more noise p' and thus a harder RL), one moves towards the 'north east' part of the plot, also showing worse ML errors on the y axis. Interestingly, such a description also holds on *test* errors (right). Estimating ξ, T from RL could yield insights on potential impacts on deployed ML models. Getting finely optimized parameters for a better picture than the one we observe would impose computing feasible $\hat{\mathsf{P}}$s under upperbounding constraints on ξ: from (6), optimizing the decomposition for T can indeed degrade ξ and *vice versa*.

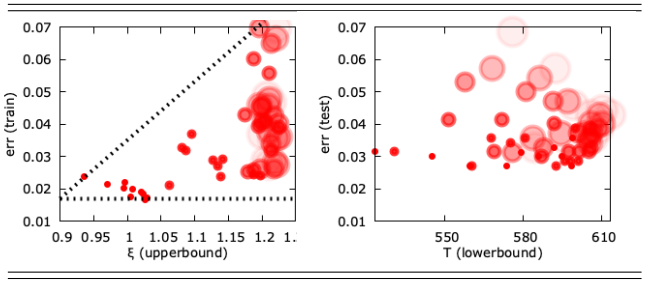


Figure 5: Breast-wisc: train (left) and test (right) errors of GREEDYRL+$\tilde{\text{C}}$ as a function of ξ (upperbound, left) and T (lowerbound, right). The bigger the disks, the larger p'.

5. Discussion

Beyond Taylor losses – Our results extend beyond Taylor losses, albeit in a more qualitative and asymptotic manner. For any F twice differentiable, any loss ℓ_F, eventually sufficiently regularized, can be locally approximated in a neighborhood of *any* local minimum by some Ridge regularized convex Taylor loss with a specific parameterization. Because the Taylor loss we analyze is in fact the leading terms of the Taylor expansion of the loss ℓ_F, F being sufficiently differentiable (Section 2), all results hold in a limit sense for ℓ_F as well, that is, as $m \to \infty$. Let $\mathcal{C}$ denote the set of local minima of $\ell_F(\hat{S}, \boldsymbol{\theta}; \gamma, \Gamma)$ — omitting dependences in $\hat{S}, \gamma, \Gamma$. a, b, c refer to the degree-2 polynomial decomposition after (1) and P is the related polynomial.

Theorem 6 *$\forall \lambda^\circ > 0$ and sample $\hat{S}$, there exists $\lambda^* > 0$ such that for any loss $\ell_F(\hat{S}, \boldsymbol{\theta}; \Gamma_F)$ satisfying $\lambda_1^\uparrow(\Gamma_F) \geq \lambda^*$ and any $\boldsymbol{\theta}^* \in \mathcal{C}$, there exists a convex Taylor loss $\ell_P(\hat{S}, \boldsymbol{\theta}; \Gamma_P)$ such that (i) $a = F(0), b = F'(0)$; (ii) $\arg\min_{\boldsymbol{\theta}} \ell_P(\hat{S}, \boldsymbol{\theta}; \Gamma_P) = \boldsymbol{\theta}^*$, and (iii) $\lambda_1^\uparrow(\Gamma_P) \geq \lambda^\circ$. Furthermore, if F is strictly convex, then $c > 0$.*

(Proof in SM, Subsection I.4) The proof shows that λ^* depends on the (finite) supremum of F'' in a certain interval. A relevant technical question is the strength of the regularization imposed (λ^*) — there would be little interest in an equivalence that would make the regularizer dominate the loss. The SM contains the proof that λ^* is small for a wide subset of major losses for supervised learning called *proper*, *i.e.* for which Bayes rule is optimal. To be complete with technical assumptions, we have also assumed wlog that the null vector is not optimal for the Taylor loss — it would also hold when $F'(0) \neq 0$, which is *e.g.* ensured for classification calibrated losses (Bartlett et al., 2006).

Beyond linear models – Working with linear models is technically convenient and a good fit for federated learning (Kairouz et al., 2021) but does not tell the full story of the application of our results. A deep model can be thought of as a model which replaces classification using $\boldsymbol{\theta}^\top \boldsymbol{x}$ by

$\boldsymbol{\theta}^\top \varphi(\boldsymbol{x})$ where $\varphi : \mathcal{X} \to \mathbb{R}^d$ is a learned 'deep' feature embedding: $\varphi \doteq h_n \circ h_{n-1} \circ ... \circ h_1$, where each h has Lipschitz constant L, an assumption now mainstream in robust learning (Cranko et al., 2019; 2021). Assuming $\mathcal{X}$ is normed (say $\|.\|_*$) and keeping the same definition for $X_* \doteq \max_i \|\boldsymbol{x}_i\|_*$, the changes due to the deep architecture on our results mainly appear in two different places: first in the key parameter $\delta_{\boldsymbol{\theta}} \doteq \|\boldsymbol{\theta}^*\|_2 L^n X_*$, second in the data-loss calibration assumption where (9) sees X_* replaced by $L^n X_*$. if $L \leq 1$ then the message of our paper stands without major changes. Otherwise, deep learning affects negatively m for the immunity of large margin classification (*e.g.* a factor $O(L^{n/\alpha})$ in the RHS of (13)), emphasizes a much stronger regularization on the loss (Assumption 2), and probably as a consequence gives slightly better bounds in terms of convergence $\tilde{\boldsymbol{\theta}}^* \to_m \boldsymbol{\theta}^*$ (Theorem 3). This is just a straight application of our results to deep learning. A devoted theory could be available with much better insights.

Consequences for federated learning – We emphasize the paucity of formal results for ML from vertical partitioned data. There has been no formal treatment of this question so far, the closest works being related to the restricted case where one peer would have the *complete* observations (and the other the class) (Unnikrishnan et al., 2015; Pananjady et al., 2017; Flammarion et al., 2016). In the context of siloed data for federated learning, our results could lead to RL algorithms designed for the RL + ML pipeline *and* privacy compliant. The strength of our results makes it reasonable to believe that a substantial weakening of the vertical partition setting to get to (1) is available at affordable formal expense for the pipeline record linkage-learning. Such a theory would probably bring new key parameters to the table (like ξ, T) whose formal analysis would be important to cover the understanding of the RL+ML pipeline.

6. Conclusion

This paper describes a global picture guaranteeing that the errors of an approximate RL algorithm do not snowball with those of learning linear models, in a framework compliant with the increasingly popular federated learning setting, but also relevant to centralized learning. The key parts rely on essential properties of the RL algorithm and, to a lesser extent, on the design (regularization) of the loss. Our setting offers simple rules on how to approach RL for supervised learning in different scenarii; at first, it formalizes the intuitive "peace of mind" case, operational already in the small data regime, where a handful number of RL mistakes cannot damage ML more than the generalization uncertainty inherent to ML (*e.g.* the slow rate regime, see Figure 2 **(B)**). Second, with our results on margin-based immunity, we formalise the idea that bringing new features that are *informative* relatively to the ones we already have can guarantee good classification. Third, if for some reason RL mistakes can be large, we formalise strategies to optimise RL as a preprocessing step to ML, *e.g.* by focusing on RL mistakes between classes or limiting the magnitude of mistakes (ξ). Last and importantly, all this happens with a substantial disconnect from the choice of the ML loss (§ 5).

We leave two important open question, (1) on the formal side, the extension of our results to the case where vertical partition does not hold anymore and some observations in one peer do not necessarily have a correspondence in the other, and (2) on the privacy and algorithmic side, as to how our algorithms can be translated to efficient algorithms in a *secure* federated learning environment where RL has to comply with privacy constraints. The strength of our results makes it reasonable to believe that a substantial weakening of the vertical partition setting to get to (1) is available at affordable formal expense for the pipeline record linkage-learning. This is crucial because this pipeline is pivotal to federated learning: to our knowledge, there is only *one* exception to this pipeline (Patrini et al., 2016). It was shown there how one can learn a model from sufficient statistics of the class instead of examples, many of which would not require record linkage to be considered. However, this approach suffers four shortcomings with respect to ours: (a) the results are developed for the square loss only, (b) building these sufficient statistics always require all peers to have the classes, (c) their theory does not give a quantitative account of the deviations to the ideal classifier that compares with ours and (d) experimentally, the approach does not compare to the ideal classifier, even when shared features are noise-free.

In all cases, our results are a very strong advocacy for federated learning, and signal the existence of non-trivial tradeoffs for RL to be optimized with the objective of learning from linked data. This can be interesting not just for our approach to RL but also when RL models are *trained* models themselves, providing scores that can be tweaked. These last observations are important because record linkage is an active field with a large number of different approaches, yet still achieving limited consensus regarding the functions to optimize during RL. We hope our theory helps shape agreement at least on some for specific usages of RL.

Acknowledgments

The authors would like to thank Kee Siong Ng, Max Ott, Hugh Durrant-Whyte and the anonymous reviewers for insightful comments. Work started while the authors were in the Confidential Computing project with Data61.

References

Aono, Y., Hayashi, T., Trieu Phong, L., and Wang, L. Scalable and secure logistic regression via homomorphic encryption. In *CODASPY*, 2016.

Avis, D. A survey of heuristics for the weighted matching problem. *Networks*, 13:475–493, 1983.

Bartlett, P., Jordan, M., and McAuliffe, J. D. Convexity, classification, and risk bounds. *J. of the Am. Stat. Assoc.*, 101:138–156, 2006.

Bierens, H.-J. *Introduction to the Mathematical and Statistical Foundations of Econometrics*. Cambridge University Press, 2004.

Blake, C. L., Keogh, E., and Merz, C. UCI repository of machine learning databases, 1998.

Christen, P. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

Christen, P. Privacy-preserving record linkage, 2016. ScaDS summer school on big data.

Christen, P. and Pudjijono, A. Accurate synthetic generation of realistic personal information. In *PAKDD*, pp. 507—-514, 2009.

Christophides, V., Efthymiou, V., Palpanas, T., Papadakis, G., and Stefanidis, K. End-to-end entity resolution for big data: A survey, 2020.

Cranko, Z., Menon, A.-K., Nock, R., Ong, C. S., Shi, Z., and Walder, C.-J. Monge blunts Bayes: Hardness results for adversarial training. In *36^{th} ICML*, pp. 1406–1415, 2019.

Cranko, Z., Shi, Z., Zhang, X., Nock, R., and Kornblith, S. Generalised Lipschitz regularisation equals distributional robustness. In *38^{th} ICML*, 2021.

Djatmiko, M., Hardy, S., Henecka, W., Ivey-Law, H., Ott, M., Patrini, G., Smith, G., Thorne, B., and Wu, D. Privacy-preserving entity resolution and logistic regression on encrypted data. In *ICML workshop on Private and Secure ML*, 2017.

Esperança, P.-M., Aslett, L.-J.-M., and Holmes, C.-C. Encrypted accelerated least squares regression. In *20^{th} AISTATS*, pp. 334–343, 2017.

Flammarion, N., Mao, C., and Rigollet, P. Optimal rates of statistical seriation. *CoRR*, abs/1607.02435, 2016.

Gascón, A., Schoppmann, P., Balle, B., Raykova, M., Doerner, J., Zahur, S., and Evans, D. Privacy-preserving distributed linear regression on high-dimensional data. *PoPET*, 2017.

Getoor, L. and Machanavajjhala, A. Entity resolution for big data. In *19^{th} KDD*, pp. 1527, 2013.

Giacomelli, I., Jha, S., Page, C.-D., and Yoon, K. Privacy-preserving ridge regression on distributed data. *IACR Cryptology ePrint Archive*, 2017:707, 2017.

Gómez, L., Rusiñol, M., and Karatzas, D. LSDE: Levenshtein space deep embedding for query-by-string word spotting. In *ICDAR'17*, pp. 499–504, 2017.

Gu, B., Dang, Z., Li, X., and Huang, H. Federated doubly stochastic kernel learning for vertically partitioned data. In *26^{st} KDD*, pp. 2483–2493, 2020a.

Gu, B., Xu, A., Huo, Z., Deng, C., and Huang, H. Privacy-preserving asynchronous federated learning algorithms for multi-party vertically collaborative learning. *CoRR*, abs/2008.06233, 2020b.

Hernández, M.-A. and Stolfo, S.-J. Real-world data is dirty: Data cleansing and the merge/purge problem. *DMKD*, 2: 9–37, 1998.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 2021.

Kang, Y., Liu, Y., and Chen, T. FedMVT: Semi-supervised vertical federated learning with multiview training. *CoRR*, abs/2008.10838, 2020.

Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., Deep, R., Arcaute, E., and Raghavendra, V. Deep learning for entity matching: A design space exploration. In Das, G., Jermaine, C. M., and Bernstein, P. A. (eds.), *SIGMOD'18*, pp. 19–34, 2018.

Nikolaenko, V., Weinsberg, U., Ioannidis, S., Joye, M., Boneh, D., and Taft, N. Privacy-preserving ridge regression on hundreds of millions of records. In *IEEE Symposium on Security and Privacy*, pp. 334–348, 2013.

Pananjady, A., Wainwright, M.-J., and Courtade, T.-A. Denoising linear models with permuted data. In *ISIT'17*, pp. 446–450, 2017.

Patrini, G., Nock, R., Rivera, P., and Caetano, T. (Almost) no label no cry. In *NIPS*27*, 2014.

Patrini, G., Nock, R., Hardy, S., and Caetano, T. Fast learning from distributed datasets without entity matching. In *IJCAI*, 2016.

Schapire, R. E. and Singer, Y. Improved boosting algorithms using confidence-rated predictions. *MLJ*, 37:297–336, 1999.

Schnell, R. Efficient private record linkage of very large datasets. In 59^{th} *World Statistics Congress*, 2013.

Unnikrishnan, J., Haghighatshoar, S., and Vetterli, M. Unlabeled sensing with random linear measurements. *CoRR*, abs/1512.00115, 2015.

Winkler, W.-E. Record linkage. In *Handbook of Statistics*, pp. 351–380. Elsevier, 2009.