

Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption

Stephen Hardy Wilko Henecka Hamish Ivey-Law Richard Nock
 Giorgio Patrini Guillaume Smith Brian Thorne

N1Analytics, Data61 ^{*}

firstname.lastname@data61.csiro.au
 g.patrini@uva.nl

Abstract

Consider two data providers, each maintaining private records of different feature sets about common entities. They aim to learn a linear model *jointly* in a federated setting, namely, data is local and a shared model is trained from locally computed updates. In contrast with most work on distributed learning, in this scenario (i) data is split *vertically*, *i.e.* by features, (ii) only one data provider knows the target variable and (iii) entities are *not* linked across the data providers. Hence, to the challenge of private learning, we add the potentially negative consequences of mistakes in entity resolution.

Our contribution is twofold. First, we describe a three-party end-to-end solution in two phases—privacy-preserving entity resolution and federated logistic regression over messages encrypted with an additively homomorphic scheme—, secure against a honest-but-curious adversary. The system allows learning without either exposing data in the clear or sharing which entities the data providers have in common. Our implementation is as accurate as a naive non-private solution that brings all data in one place, and scales to problems with millions of entities with hundreds of features. Second, we provide what is to our knowledge the first formal analysis of the impact of entity resolution’s mistakes on learning, with results on how optimal classifiers, empirical losses, margins and generalisation abilities are affected. Our results bring a clear and strong support for federated learning: under reasonable assumptions on the number and magnitude of entity resolution’s mistakes, it can be extremely beneficial to carry out federated learning in the setting where each peer’s data provides a significant uplift to the other.

^{*}All authors contributed equally. Richard Nock is jointly with the Australian National University & the University of Sydney. Giorgio Patrini is now at the University of Amsterdam.

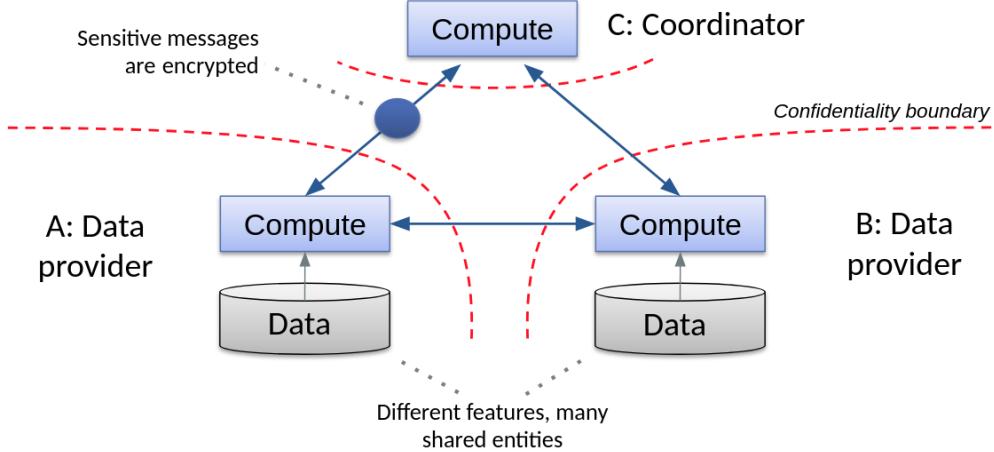


Figure 1: Relationships between the Coordinator, C, and the Data Providers, A and B.

1 Introduction

With the ever-expanding collection and use of data, there are increasing concerns about security and privacy of the data that is being collected and/or shared (Esperança et al. [2017] and references therein). These concerns are both on the part of the consumer, whose information is often used or traded with little consent, and on the part of the collector, who is often liable for protecting the collected data.

On the other hand, organisations are increasingly aware of the potential gain of combining their data assets, specifically in terms of increased statistical power for analytics and predictive tasks. For example, banks and insurance companies could collaborate for spotting fraudulent activities; hospitals and medical facilities could leverage the medical history of common patients in order to prevent chronic diseases and risks of future hospitalisation; and online businesses could learn from purchase patterns of common users so as to improve recommendations. There is little work on systems that can make use of distributed datasets for federated learning in a sufficiently secure environment, and even less any formal analysis of such a system, from the security and learning standpoint.

Contributions — In this paper, we provide **both**:

- First, we propose an end-to-end solution for the case where two organisations hold different data about (undisclosed) common entities. Two data providers, A and B, see only their own side of a vertically partitioned dataset. They aim to *jointly* perform logistic regression in the cross-feature space. Under the assumption that raw data *cannot* be exchanged, we present a secure protocol that is managed by a third party C, the coordinator, by employing privacy-preserving entity resolution and an additively homomorphic encryption scheme. This is our first contribution. Relationships between A, B and C are presented in Figure 1.
- Our second contribution is the first formal study of an often overlooked source of errors of this process: the impact of entity resolution errors on learning. Since A and B use different descriptive features, linking entities across databases is error prone [Christen, 2012]. Intuitively we might expect such errors to have negative impact on learning: for example,

wrong matches of a hospital database with pharmaceutical records with the objective to improve preventive treatments could be disastrous. Case studies report that exact matching can be very damaging when identifiers are not stable and error-prone: 25% true matches would have been missed by exact matching in a census operation [Schnell, 2013, Winkler, 2009]. We are not aware of results quantifying the impact of error-prone entity-resolution on learning. Such results would be highly desirable to (i) find the key components of entity resolution errors that impact learning the most and then (ii) devise improvements of entity-resolution algorithms in a machine learning context. We provide in this paper four main contributions from that standpoint.

- First, we provide a formal bound on the deviation of the optimal classifier when such errors occur. It shows for example that wrongly linking examples of *different classes* can be significantly more damaging than wrongly linking examples of the same class.
- Second, we show that under some reasonable assumptions, the classifier learned is *immune* to entity-resolution mistakes with respect to classification for *large margin* examples. More precisely, examples that receive a large margin classification by the optimal (unknown) classifier still receive the same class by the classifier we learn from mistake-prone entity-resolved data.
- Third, under the same assumptions, we bound the difference between the *empirical loss* of our classifier on the *true data* (*i.e.* built knowing the true entity-resolution) with respect to that of the optimal (unknown) classifier, and it shows a convergence of both losses at a rate of order $1/n^\alpha$, where n is the number of examples and $\alpha \in (0, 1]$ is an assumption-dependent constant. The bound displays interesting dependencies on three distinct penalties respectively depending on the optimal classifier, entity resolution and a sufficient statistics for the class in the true data.
- Fourth, under the additional assumption that entity resolution mistakes are small enough in number, we show that not even rates for *generalization* are notably affected by entity resolution. The same key penalties as for the empirical loss bounds appear to drive generalization bounds.

These contributions, we believe, represent *very strong* advocacies for federated learning when aggregating databases provides a significant uplift for classification accuracy.

The rest of this paper is organised as follows. Section 2 presents related work. Section 3 presents the security environment and primitives. Section 4 details our approach for privacy-preserving entity-resolution. Section 5 develops our approach for secure logistic regression. Section 7 investigates the formal properties of learning in a federated learning setting that relies on entity resolution. Section 8 presents and discusses experiments. We provide a conclusion in the last section. An appendix, starting page 25, provides all proofs and details on encryption, encoding, security evaluation and cryptographic longterm keys.

2 Related work

The scenario is federated learning [Konecný et al., 2016, McMahan et al., 2017], a distributed setting where data does not leave its premises and data providers protect their privacy against a central aggregator. Our interest is logistic regression on vertically partitioned data and, importantly, we consider additional privacy requirements and entity resolution with a given error rate.

Research in privacy-preserving machine learning is currently dominated by the approach of differential privacy [Dwork, 2008, Dwork and Roth, 2014]. In the context of machine learning, this amounts to ensuring—with high probability—that the output predictions are the same regardless of the presence or absence of an individual example in the dataset. This is usually achieved by adding properly calibrated noise to an algorithm or to the data itself [Chaudhuri and Monteleoni, 2009, Duchi et al., 2013]. While computationally efficient, these techniques invariably degrade the predictive performance of the model.

We opt for security provided by more traditional cryptographic guarantees such as homomorphic encryption, *e.g.* Paillier [Paillier, 1999] or Brakerski-Gentry-Vaikuntanathan cryptosystems [Gentry, 2009], secure multi-party computation, *e.g.* as garbled circuits [Yao, 1986] and secret sharing [Ben-Or et al., 1988]. By employing additively homomorphic encryption, we sit in this space. In contrast with differential privacy, instead of sacrificing predictive power, we trade security for computational cost—the real expense of working with encryption.

Work in the area can be classified in terms of whether the data is vertically or horizontally partitioned, the security framework, and the family of learning models. We limit ourselves to mention previous work with which we have common elements. The overwhelming majority of previous work on secure distributed learning considers a *horizontal* data partition. Solutions can take advantage of the separability of loss functions which decompose the loss by examples. Relevant approaches using partially homomorphic encryption are Xie et al. [2016], Aono et al. [2016].

A vertical data partition requires a more complex and expensive protocol, and therefore is less common. Wu et al. [2013] run logistic regression where one party holds the features and the other holds the labels. Duverle et al. [2015] use the Paillier encryption scheme to compute a variant of logistic regression which produces a p -value for each feature separately (rather than a logistic model as we do here). Their partition is such that one party holds the private key, the labels and a single categorical variable, while the other party holds all of the features. Gascón et al. [2017] perform linear regression on vertically partitioned data via a hybrid multi-party computation combining garbled circuits with a tailored protocol for the inner product. Recently, Mohassel and Zhang [2017] presented a system for privacy-preserving machine learning for linear regression, logistic regression and neural network training. They combine secret sharing, garbled circuits and oblivious transfer and rely on a setting with two un-trusted, but non-colluding servers.

None of the papers cited before consider the problem of entity resolution (or entity matching, record linkage, Christen [2012]). For example, Gascón et al. [2017] assume that the correspondence between rows in the datasets owned by different parties is *known a priori*. Such an assumption would not stand many real-world applications where identifiers are not stable and/or recorded with errors [Schnell, 2013, Winkler, 2009], making entity matching a prerequisite for working with vertically partitioned in most realistic scenarios. Finding efficient and privacy-compliant algorithms is a field in itself, *privacy-preserving entity resolution* [Hall and Fienberg, 2010, Christen, 2012, Vatsalan et al., 2013a].

Table 1 summarizes some comparisons between our setting and results to others. Perhaps

	us	Gascón et al. [2017]	Duverle et al. [2015]	Wu et al. [2013]	Esperança et al. [2017]	Xie et al. [2016]	Aono et al. [2016]	Konecný et al. [2016]
Vertical partition	✓	✓	✓	✓	✗	✗	✗	✗
Convergence	✓	✓	✗	✗	✓	✓	✗	✓
ER	✓	✗	✗	✗				

Table 1: Comparison of related approaches on federated learning, (from top row to bottom row) whether they rely on vertically partitioned data, analyze convergence and/or entity resolution (ER); the presence of ER is justified only in case of vertical partition.

surprisingly, there is an exception to that scheme: it was recently shown that entity resolution is *not necessary* for good approximations of the optimal model. Related methods exploit the fact that one can learn from sufficient statistics for the class instead of examples [Nock et al., 2015, Nock, 2016, Patrini, 2016], many of which do not require entity resolution to be computed, but rather a weak form of entity resolution between groups of observations that share common subset of features [Patrini et al., 2016b]. To our knowledge, Patrini et al. [2016b] is also the only work other than ours to study entity resolution and learning in a pipelined process, although the privacy guarantees are different. Crucially, it requires labels to be shared among all parties, which we *do not*, and also the theoretical guarantees are yet not as comprehensive as the ones we are going to deliver.

3 Security environment and primitives

Security model — We assume that the participants are *honest-but-curious*: (i) they follow the protocol without tampering with it in any way; (ii) they do not collude with one another; but (iii) they will nevertheless try to infer as much as possible from the information received from the other participants. The honest-but-curious assumption is reasonable in our context since A and B have an incentive to compute an accurate model. The third party, C, holds the private key used for decryption; however the only information C receives from A and B are encrypted model updates, which we do not consider private in our setup.

We assume that A and B’s data is secret, but that the schema (the number of features and the type of each) of each data provider is available to all parties. We assume that the agents communicate on pre-established secure channels. We work under additional privacy constraints:

1. Knowledge of common entities remain secret to A and B, as does the number of common entities.
2. No raw sensitive data leaves A or B before encryption.

On the other hand, a data provider can safely use its own unencrypted records *locally* anytime it is useful to do so.

Additively homomorphic encryption — We recall here the main properties of additively homomorphic encryption schemes such as Paillier [1999]. As a public key system, any party can encrypt their data with a known *public key* and perform computations with data encrypted by others with the same public key. To extract the plaintext, the result needs to be sent to the holder of the *private key*.

An additively homomorphic encryption scheme only provides arithmetic for elements of its plaintext space. In order to support algorithms over floating-point numbers, we must define an encoding scheme that maps floats to modular integers and which preserves the operations of addition and multiplication. The encoding system we use is similar to floating-point representation; a number is encoded as a pair consisting of an encrypted significand and a unencrypted exponent. Details of the encoding scheme and its limitations are given in Appendix II.

An additively homomorphic encryption scheme provides an operation that produces the encryption of the sum of two numbers, given only the encryptions of the numbers. Let the encryption of a number u be $\llbracket u \rrbracket$. For simplicity we overload the notation and we denote the operator with ‘+’ as well. For any plaintexts u and v we have:

$$\llbracket u \rrbracket + \llbracket v \rrbracket = \llbracket u + v \rrbracket. \quad (1)$$

Hence we can also multiply a ciphertext and a plaintext together by repeated addition:

$$v \cdot \llbracket u \rrbracket = \llbracket vu \rrbracket, \quad (2)$$

where v is *not* encrypted (it is not possible to multiply two ciphertexts). In short, we can compute sums and products of plaintexts and ciphertexts *without leaving the space of encrypted numbers*.

These operations can be extended to work with vectors and matrices component-wise. For example, we denote the inner product of two vectors of plaintexts \mathbf{u} and \mathbf{v} by $\mathbf{v}^\top \llbracket \mathbf{u} \rrbracket = \llbracket \mathbf{v}^\top \mathbf{u} \rrbracket$ and the component-wise product by $\mathbf{v} \circ \llbracket \mathbf{u} \rrbracket = \llbracket \mathbf{v} \circ \mathbf{u} \rrbracket$. Summation and matrix operations work similarly; see Appendix I for details. Hence, using an additively homomorphic encryption scheme we can implement useful linear algebra primitives for machine learning.

Doing arithmetic on encrypted numbers comes at a cost in memory and processing time. For example, with Paillier encryption scheme, the encryption of an encoded floating-point number (whether single or double precision) is $2m$ bits long, where m is typically at least 1024 [BlueKrypt, 2017] and the addition of two encrypted numbers is two to three orders of magnitude slower than the unencrypted equivalent. Nevertheless, as we will see later, with a carefully engineered implementation of the encryption scheme, a large proportion of real-world problems are tractable.

4 Privacy-preserving entity resolution

When a dataset is vertically partitioned across organisations the problem arises of how to identify corresponding entities. The ideal solution would be joining datasets by common unique IDs; however, across organisations this is rarely feasible. An approximate solution is given by techniques for *entity resolution* [Christen, 2012]; Figure 2 gives a pictorial representation. Solving this problem is a requirement for learning from the features of the two parties.

In a non-privacy-preserving context, matching can be based on shared *personal identifiers*, such as name, address, gender, and date of birth, that are effectively used as weak IDs. In our scenario however, weak identifiers are considered private to each party. Thus, in order to perform

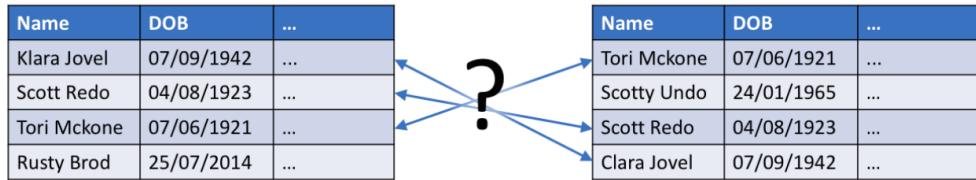


Figure 2: The problem of entity resolution.

Algorithm 1: Privacy-preserving entity matching

Data: personal identifiers for entities in D_A and D_B

Result: Permutations σ, τ and mask m

A and B

create a CLK for every entity and send them to C

C

obtain σ, τ and m by matching CLKs

encrypts $[m]$

sends σ and τ to A and B respectively and $[m]$ to both

privacy-preserving entity resolution we use an anonymous linking code called a *cryptographic longterm key* (CLK) introduced by Schnell et al. [2011].

The CLK is a Bloom filter encoding of multiple personal identifiers. The encoding method works by hashing n-gram sub-strings of selected identifiers to bit positions in the Bloom filter. A measure of similarity between CLKs is computed on the number of matching bits by the *Dice coefficient* [Schnell et al., 2011].¹

The protocol for privacy-preserving entity matching is shown in Algorithm 1. After CLKs of all entities from A and B are received, C matches them by computing the Dice coefficient for all possible pairs of CLKs, resulting in a number of comparisons equal to the product of the datasets sizes. The most similar pairs are selected as matches, in a greedy fashion. Faster computation is possible by *blocking* [Vatsalan et al., 2013b].

The outputs of entity matching are two permutations σ, τ and a mask m : σ, τ describe how A and B must rearrange their rows so as to be consistent with each other; m specifies whether a row corresponds to an entity available in both data providers, thus to be used for learning. The encrypted mask and its integration into the process of private learning is novel and part of our contribution. No assumption was made on their relative size. For the learning phase though, they must be the same. Hence the longer dataset is truncated, excluding only non-matched entities.²

¹While CLKs are robust to typographical errors, they are susceptible to cryptanalytic attacks if insecure parameters are used, or if too few identifiers are hashed [Kuzu et al., 2011, 2013, Niedermeyer et al., 2014].

²As a consequence, the owner of the longer dataset will learn that the truncated rows have no correspondence on the other dataset. This is a mild leak of information that does not violate the first security requirement in Section 3.

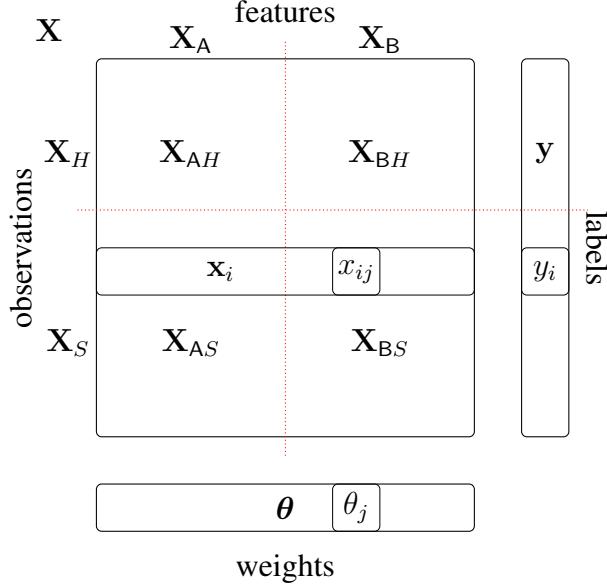


Figure 3: Overview of the notation for, and relationships between, the different variables in logistic regression.

More precisely, entity resolution of \mathbf{D}_A and \mathbf{D}_B relative to CLRs consists of two permutations σ and τ of the rows of \mathbf{D}_A and \mathbf{D}_B , and a mask m of length n whose elements i satisfy

$$m_i = \begin{cases} 1 & \text{if } \sigma(\mathbf{D}_A^{\text{CLR}})_i \sim \tau(\mathbf{D}_B^{\text{CLR}})_i, \text{ and} \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

The operator ‘ \sim ’ can be read as “the most likely match”, in the sense of Schnell et al. [2011]. Permutations and mask are randomized subject to the relation (3), where $m_i = 0$ means that there is no record in the other dataset which could match with a high enough probability.

In our scenario, whether a record in a data provider is a match is considered private; see requirement 1. in Section 3. For example, when linking to a medical dataset of patients, successful matching of an entity could reveal that a person in an unrelated database suffers the medical condition. In order to keep the mask confidential, C encrypts it with the Paillier scheme before sending it to A and B. Details on the use of mask are given in Section 6.

5 Logistic regression, Taylor approximation

We need to adapt logistic regression and stochastic gradient descent to work with an additively homomorphic encryption scheme and the masks. In this section the focus is on the “non-distributed” setting—all data is available in one place—, while Section 6 details the federated learning protocol.

With logistic regression we learn a linear model $\theta \in \mathbb{R}^d$ that maps examples $\mathbf{x} \in \mathbb{R}^d$ to a binary label $y \in \{-1, 1\}$. The learning sample S is a set of n example-label pairs (\mathbf{x}_i, y_i) from $i = 1, \dots, n$. Figure 3 presents the notation used in this section, showing in particular that the

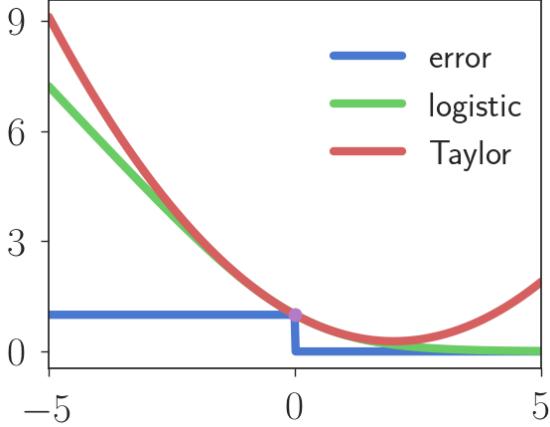


Figure 4: Loss profiles.

observation matrix is used row-wise. The average logistic loss computed on the training set is

$$\ell_S(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i \in S} \log(1 + e^{-y_i \boldsymbol{\theta}^\top \mathbf{x}_i}). \quad (4)$$

In turn, the stochastic gradients computed on a mini-batch $S' \subseteq S$ of size s' are

$$\nabla \ell_{S'}(\boldsymbol{\theta}) = \frac{1}{s'} \sum_{i \in S'} \left(\frac{1}{1 + e^{-y_i \boldsymbol{\theta}^\top \mathbf{x}_i}} - 1 \right) y_i \mathbf{x}_i. \quad (5)$$

Below we adapt equations (4) and (5) to accommodate the encrypted mask. Learning requires the computation of gradients only, not of the loss itself. Yet, to combat overfitting we monitor the loss function on a hold-out for early stopping, other than using ridge regularization. From now on, we denote ℓ_H as the loss value on a hold-out H of size h .

Taylor loss — In order to operate under the constraints imposed by an additively homomorphic encryption scheme, we need to consider approximations to the logistic loss and the gradient. To achieve this, we take a Taylor series expansion of $\log(1 + e^{-z})$ around $z = 0$:

$$\log(1 + e^{-z}) = \log 2 - \frac{1}{2}z + \frac{1}{8}z^2 - \frac{1}{192}z^4 + O(z^6). \quad (6)$$

The second order approximation of (4) evaluated on H is:

$$\ell_H(\boldsymbol{\theta}) \approx \frac{1}{h} \sum_{i \in H} \log 2 - \frac{1}{2} y_i \boldsymbol{\theta}^\top \mathbf{x}_i + \frac{1}{8} (\boldsymbol{\theta}^\top \mathbf{x}_i)^2, \quad (7)$$

where we have used the fact that $y_i^2 = 1, \forall i$. We call this function the *Taylor loss*. By differentiation, we write the gradient for a mini-batch S' as:

$$\nabla \ell_{S'}(\boldsymbol{\theta}) \approx \frac{1}{s'} \sum_{i \in S'} \left(\frac{1}{4} \boldsymbol{\theta}^\top \mathbf{x}_i - \frac{1}{2} y_i \right) \mathbf{x}_i. \quad (8)$$

The second-order is a compromise between precision and computational overhead. From (6), the third-order approximation equals the second, while fourth and fifth-orders are not fit for

minimization since their images take negative values. Minimizing the sixth-order Taylor loss is costly when working in the encrypted space, and some simple experiments are enough to show that the higher degree terms do not provide significant performance gains.

Additionally, the expansion is around 0, which implies a rough approximation of the logistic loss when $|\boldsymbol{\theta}^\top \mathbf{x}_i| \gg 0$. Nevertheless, square losses are commonly used for classification [Nock and Nielsen, 2009] and that is essentially what we have obtained. Experiments in Section 8 show that features standardization suffices for good performance. The loss is pictured in Figure 4. The logistic loss is bounded above by the Taylor loss at every point, so their values are not directly comparable and in addition their *minimizers* will differ.

Applying the encrypted mask — The encrypted mask can be incorporated into (4) by multiplying each term by $\llbracket m_i \rrbracket$. Combined with the Taylor loss, the *masked* gradient for a mini-batch S' is

$$\llbracket \nabla \ell_{S'}(\boldsymbol{\theta}) \rrbracket \approx \frac{1}{S'} \sum_{i \in S'} \llbracket m_i \rrbracket \left(\frac{1}{4} \boldsymbol{\theta}^\top \mathbf{x}_i - \frac{1}{2} y_i \right) \mathbf{x}_i, \quad (9)$$

and the *masked* logistic loss on H is

$$\llbracket \ell_H(\boldsymbol{\theta}) \rrbracket \approx \llbracket \nu \rrbracket - \frac{1}{2} \boldsymbol{\theta}^\top \llbracket \mu \rrbracket + \frac{1}{8h} \sum_{i \in H} \llbracket m_i \rrbracket (\boldsymbol{\theta}^\top \mathbf{x}_i)^2, \quad (10)$$

where $\llbracket \nu \rrbracket = ((\log 2)/h) \sum_{i \in H} \llbracket m_i \rrbracket$ and $\llbracket \mu \rrbracket = (1/h) \sum_{i \in H} \llbracket m_i \rrbracket y_i \mathbf{x}_i$. The constant term $\llbracket \nu \rrbracket$ is irrelevant for minimization since it is model-independent; henceforth we set it to 0.

6 Secure federated logistic regression

We detail now the second phase of our pipeline which amounts to run the federated logistic regression with SGD. We assume that the entity resolution protocol has been run, thus A and B have permuted their datasets accordingly, which now have the same number of rows n . The complete dataset is a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$. This matrix does not exist in one place but is composed of the columns of the datasets A and B held respectively by A and B; this gives the vertical partition:

$$\mathbf{X} = [\mathbf{X}_A \mid \mathbf{X}_B] \quad (11)$$

as seen in Figure 3. A also holds the label vector y . Let \mathbf{x} be a row of \mathbf{X} . Define \mathbf{x}_A to be the restriction of \mathbf{x} to the columns of A, and similarly for $\boldsymbol{\theta}$ in place of x and B in place of A. Then we can decompose $\boldsymbol{\theta}^\top \mathbf{x}$ as:

$$\boldsymbol{\theta}^\top \mathbf{x} = \boldsymbol{\theta}_A^\top \mathbf{x}_A + \boldsymbol{\theta}_B^\top \mathbf{x}_B. \quad (12)$$

Algorithm 2 computes the secure logistic regression and it is executed by C. It is a standard SGD where computations involving raw data are replaced with their secure variants. We describe computation of gradients in this section, while we defer to Appendix III for the initialization and the loss evaluation for stopping criterion.

Prior to learning, C generates a key pair for the chosen cryptosystem and shares the public key with A and B. Then, C sends the encrypted mask $\llbracket m \rrbracket$ to A and B. This allows the implementation of a protocol where C is oblivious to the hold-out split and the mini-batch sampling of the training set.

Algorithm 2: Secure logistic regression (run by C)

Data: Mask \mathbf{m} , learning rate η , regularisation Γ , hold-out size h , batch size s'

Result: Model θ

create an additively homomorphic encryption key pair

send the public key to A and B

encrypt \mathbf{m} with the public key, send $\llbracket \mathbf{m} \rrbracket$ to A and B

run Algorithm 4 (h)

$\theta \leftarrow \mathbf{0}, \ell_H \leftarrow \infty$

repeat

for every mini-batch S' **do**

$\nabla \ell_{S'}(\theta) \leftarrow$ Algorithm 3 (θ, t)

$\theta \leftarrow \theta - \eta (\nabla \ell_{S'}(\theta) + \Gamma \theta);$

$\ell_H(\theta) \leftarrow$ Algorithm 5 (θ)

if $\ell_H(\theta)$ has not decreased for a while **then break**

until max iterations

return θ

Algorithm 4 initializes the protocol. A common hold-out H is sampled by A and B. Additionally, they compute and cache $\llbracket \boldsymbol{\mu} \rrbracket$ in the logistic loss (10) computed on the hold-out.

Data has already been shuffled by the permutations result of entity matching (Section 4), hence we can access it by mini-batches sequentially. Any stochastic gradient algorithm can be used for optimisation. We choose SAG [Schmidt et al., 2013] for our experiments in Section 8; in this case, C keeps the memory of the previous gradients.

Early stopping is applied by monitoring the Taylor loss on the hold-out H by Algorithm 5 (Appendix III.2). We prefer the computation of the loss over the hold-out error, which would be too costly under any additively homomorphic encryption scheme.

Secure gradient — Algorithm 3 computes the secure gradient. It is called for every batch in Algorithm 2 and hence is the computational bottleneck. We demonstrate the correctness of Algorithm 3 as follows. Fix an $i \in S'$ and let $\mathbf{x} = (\mathbf{x}_A \mid \mathbf{x}_B)$ be the i th row of \mathbf{X} . Note that the i th component of $\llbracket \mathbf{m}_{S'} \circ \mathbf{u}' \rrbracket$ is $\llbracket m_i (\frac{1}{4} \boldsymbol{\theta}^\top \mathbf{x}_A - \frac{1}{2} y_i) \rrbracket$, so the i th component of $\llbracket \mathbf{w} \rrbracket$ is $\llbracket m_i (\frac{1}{4} \boldsymbol{\theta}^\top \mathbf{x} - \frac{1}{2} y_i) \rrbracket$ by (12). Then

$$\llbracket \mathbf{z} \rrbracket = \mathbf{X}_{BS'} \llbracket \mathbf{w} \rrbracket = \left[\left[\sum_{i \in H} m_i x_{ij} \left(\frac{1}{4} \boldsymbol{\theta}^\top \mathbf{x}_i - \frac{1}{2} y_i \right) \right] \right]_j \quad (13)$$

where j ranges over the columns of B (notation $[u_j]_j$ is a vector notation whose coordinates are defined by the set of values $\{u_j\}_j$); similarly A calculates the same for j ranging over columns of A

Algorithm 3: Secure gradient

Data: Model θ , batch size s'

Result: $\nabla \ell_{S'}(\theta)$ of an (undisclosed) mini-batch S'

C

send θ to A

A

select the next batch $S' \subset S$, $|S'| = s'$

$$\mathbf{u} = \frac{1}{4} \mathbf{X}_{AS'} \boldsymbol{\theta}_A$$

$$[\![\mathbf{u}']\!] = [\![\mathbf{m}]\!]_{S'} \circ (\mathbf{u} - \frac{1}{2} \mathbf{y}_{S'})$$

send θ , S' and $[\![\mathbf{u}']\!]$ to B

B

$$\mathbf{v} = \frac{1}{4} \mathbf{X}_{BS'} \boldsymbol{\theta}_B$$

$$[\![\mathbf{w}]\!] = [\![\mathbf{u}']\!] + [\![\mathbf{m}]\!]_{S'} \circ \mathbf{v}$$

$$[\![\mathbf{z}]\!] = \mathbf{X}_{BS'} [\![\mathbf{w}]\!]$$

send $[\![\mathbf{w}]\!]$ and $[\![\mathbf{z}]\!]$ to A

A

$$[\![\mathbf{z}']\!] = \mathbf{X}_{AS'} [\![\mathbf{w}]\!]$$

send $[\![\mathbf{z}']\!]$ and $[\![\mathbf{z}]\!]$ to C

C

obtain $[\![\nabla \ell_{S'}(\theta)]\!]$ by concatenating $[\![\mathbf{z}']\!]$ and $[\![\mathbf{z}]\!]$

obtain $\nabla \ell_{S'}(\theta)$ by decrypting with the private key

to obtain $[\![\mathbf{z}']\!]$. C concatenates these two and get $[\![\nabla \ell_{S'}(\theta)]\!]$ by (9). During Algorithm 3, the only information sent in clear is about the model θ and the mini-batch S' , both only shared between A and B. All other messages are encrypted and C only receives $\nabla \ell_{S'}(\theta)$. Section IV in the Appendix provides an additional security evaluation of our algorithms, including sources of potential leakage of information.

7 Theoretical assessment of the learning component

As we work on encrypted data, the convergence rate of our algorithm is an important point; since we consider entity resolution, it is crucial to investigate the impact of its errors on learning.

7.1 Convergence

The former question is in fact already answered: although computations are performed in the encrypted domain, the underlying arithmetic is equivalent and thus has no influence on the optimization of the Taylor loss. Our implementation of SAG is done on a second-order Taylor loss which is ridge regularized, so we have access to the strong convexity convergence [Schmidt et al., 2013, Theorem 1]. Let S be a learning sample, assume we learn from S (not a holdout H or subset) and let

$$\ell_S(\boldsymbol{\theta}; \gamma, \Gamma) = \ell_S(\boldsymbol{\theta}) + \gamma \boldsymbol{\theta}^\top \Gamma \boldsymbol{\theta} \quad (14)$$

denote the ridge regularized Taylor loss, with $\gamma > 0$, matrix $\Gamma \succ 0$ (positive semi-definite) symmetric, then we can expect convergence rates for $\ell_S(\boldsymbol{\theta}; \gamma, \Gamma) \rightarrow \min_{\boldsymbol{\theta}} \ell_S(\boldsymbol{\theta}; \gamma, \Gamma)$ at a rate approaching ρ^k for some $0 < \rho < 1$, k being the number of mini-batch updates in Algorithm 2.

7.2 Impact of entity resolution: parameters

This leaves us with the second problem, that of entity resolution, and in particular how wrong matches can affect $\min_{\boldsymbol{\theta}} \ell_S(\boldsymbol{\theta}; \gamma, \Gamma)$. To the best of our knowledge, the state of the art on formal analyses of how entity resolution affects learning is essentially a blank page, even in the vertical partition setting where both parties have access to the same set of n entities — the case we study. We let

$$\begin{aligned} \hat{\boldsymbol{\theta}}^* &= \arg \min_{\boldsymbol{\theta}} \ell_S(\boldsymbol{\theta}; \gamma, \Gamma) , \\ \boldsymbol{\theta}^* &= \arg \min_{\boldsymbol{\theta}} \ell_{S^*}(\boldsymbol{\theta}; \gamma, \Gamma) . \end{aligned}$$

S^* is the ideal dataset among all shared features, that is, reconstructed knowing the solution to entity resolution between A and B. S denotes our dataset produced via (mistake-prone) entity resolution. We shall deliver a number of results on how $\hat{\boldsymbol{\theta}}^*$ and $\boldsymbol{\theta}^*$ are related, but first focus in this section on defining and detailing the parameters and assumptions that will be key to obtaining our results.

7.2.1 Modelling entity resolution mistakes

In our setting, entity-resolution mistakes can be represented by *permutation* errors between A and B. Precisely, there exists an *unknown* permutation matrix, $P_* : [n] \rightarrow [n]$, such that instead of learning from the ideal \mathbf{X} as in (11), we learn from some

$$\hat{\mathbf{X}} = [\mathbf{X}_A | (\mathbf{X}_B^\top P_*)^\top] . \quad (15)$$

Without loss of generality, we assume that indices refer to columns in A and so permutation errors impact the indices in B. We recall that A holds the labels as well. Several parameters and assumptions will be key to our results. One such key parameter is the *size* T of P_* when factored as *elementary* permutations,

$$P_* = \prod_{j=1}^T P_j \quad (16)$$

(T unknown), where P_t (unknown) acts on some index $u_A(t), v_A(t) \in [n]$ in A . Such a factorization always holds, and it is not hard to see that there always exist a factorization with $T \leq n$.

Another key parameter is the number $T_+ \leq T$ of *class mismatch* permutations in the factorization, *i.e.* for which $y_{u_A(t)} \neq y_{v_A(t)}$. We let

$$\rho \doteq \frac{T_+}{T} \quad (17)$$

define the proportion of elementary permutations that act between classes.

We let $u_B(t)$ (resp. $v_B(t)$) denote the indices in $[n]$ of the rows in X_B that are in observation $u_A(t)$ (resp. $v_A(t)$) and that will be permuted by P_t . For example, if $u_B(t) = v_A(t), v_B(t) = u_A(t)$, then P_t correctly reconstructs observations $u_A(t)$ and $v_A(t)$.

7.2.2 Assumptions on permutations and data

We now proceed through our assumptions, that are covered in greater detail in the appendix, Section V.3. We make two categories of assumptions:

- P_* is bounded in *magnitude* and *size*,
- the data and learning problem parameters are accurately *calibrated*.

Bounding P_* in terms of magnitude — This is what we define as (ε, τ) -accuracy. Denote \hat{x}_{ti} as row i in \hat{X}_t , in which X_B is altered by the subsequence $\prod_{j=1}^t P_j$, and x_i as row i in X .

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

$$|(\hat{x}_{ti} - x_i)^\top w_B| \leq \varepsilon \cdot |x_i^\top w| + \tau \|w\|_2, \forall i \in [n], \quad (18)$$

$$|(x_{u_F(t)} - x_{v_F(t)})^\top w_F| \leq \varepsilon \cdot \max_{i \in \{u_F(t), v_F(t)\}} |x_i^\top w| + \tau \|w\|_2, \forall F \in \{A, B\}. \quad (19)$$

We say that P_* is (ε, τ) -accurate iff each P_t is (ε, τ) -accurate, $\forall t = 1, 2, \dots, T$.

If we consider that vectors $\hat{x}_{ti} - x_i, x_{u_F(t)} - x_{v_F(t)}$ quantify errors made by elementary permutation P_t , then (ε, τ) -accuracy postulates that errors along any direction are bounded by a fraction of the norm of original observations, plus a penalty that depends on the direction. We remark that in the context of the inequalities, τ is homogeneous to a norm, which is not the case for ε (which can be thought “unit-free”). For that reason, we define an important quantity that we shall use repeatedly, aggregating ε and a “unit-free” τ :

$$\xi \doteq \varepsilon + \frac{\tau}{X_*}, \quad (20)$$

where $X_* \doteq \max_i \|x_i\|_2$ is the max norm in (the columns of) X . Section V.3 in the appendix gives more context around Definition 1.

Bounding P_* in terms of size — The α -*boundedness* condition states that the decomposition in eq. (16) has a number of terms limited as a function of n .

Definition 2 We say that P_* is α -**bounded** for some $0 < \alpha \leq 1$ iff its size satisfies

$$T \leq \left(\frac{n}{\xi} \right)^{\frac{1-\alpha}{2}}. \quad (21)$$

The bounded permutation size assumption roughly means that $T = o(\sqrt{n})$ in the worst case. “Worst case” means that in favorable cases where we can fix ξ small, the assumption may be automatically verified even for α very close to 1 since it always holds that a permutation can be decomposed in elementary permutations with $T \leq n$. Note that to achieve a particular level of (ε, τ) -accuracy assumption, we may need more than the minimal factorisation, but it is more than reasonable to assume that we shall still have $T = O(n)$, which does not change the picture of the constraint imposed by α -boundedness.

Data and model calibration — We denote $\sigma(\mathcal{S})$ as the standard deviation of a set \mathcal{S} , and $\lambda_1^\dagger(\Gamma)$ the smallest eigenvalue of Γ , following Bhatia [1997]. Finally, we define the stretch of a vector.

Definition 3 The **stretch** of vector \mathbf{x} along direction \mathbf{w} with $\|\mathbf{w}\|_2 = 1$ is

$$\varpi(\mathbf{x}, \mathbf{w}) \doteq \|\mathbf{x}\|_2 |\cos(\mathbf{x}, \mathbf{w})|. \quad (22)$$

The stretch of \mathbf{x} is just the norm of the orthogonal projection along direction \mathbf{w} .

Definition 4 We say that the **data-model calibration** assumption holds iff the following two constraints are satisfied:

- Maxnorm-variance regularization: ridge regularization parameters γ, Γ are chosen so that

$$\frac{X_*^2}{\frac{(1-\varepsilon)^2}{8} \cdot \inf_{\mathbf{w}} \sigma^2(\{\varpi(\mathbf{x}_i, \mathbf{w})\}_{i=1}^n) + \gamma \lambda_1^\dagger(\Gamma)} \leq 1. \quad (23)$$

- Minimal data size: the dataset is not too small, namely,

$$n \geq 4\xi. \quad (24)$$

Remark that the data size roughly means that n is larger than a small constant, and the maxnorm-variance regularization means that the regularization parameters have to be homogeneous to a squared norm. Alternatively, having fixed the regularization parameters, we just need to recalibrate data by normalizing observations to control X_* .

Key parameters — We are now ready to deliver a series of results on various relationships between $\hat{\theta}^*$ and θ^* : deviations between the classifiers, their empirical Ridge-regularized Taylor losses, generalization abilities, etc. . Remarkably, all results depend on three distinct key parameters that we now define.

Definition 5 We define $\delta_m, \delta_\rho, \delta_\mu$ as follows:

$$\begin{aligned}\delta_m &\doteq \|\theta^*\|_2 X_*, \text{ a bound on the maximum margin for the optimal (unknown) classifier}^3; \\ \delta_\rho &\doteq \sqrt{\xi} \rho / 4, \text{ aggregates parameters } \varepsilon, \tau, \rho \text{ of } P_*; \\ \delta_\mu &\doteq \|\sum_i y_i \mathbf{x}_i\|_2 / (n X_*) (\in [0, 1]), \text{ the normalized mean-operator}^4.\end{aligned}$$

Notice that δ_m depends on the optimal classifier θ^* , δ_ρ depends on the permutation matrix P_* , while δ_μ depends on the true data S^* .

7.3 Relative error bounds for $\hat{\theta}^*$ vs θ^*

Our analysis gives the conditions on the unknown permutation, the learning problem and the data, for the following inequality to hold:

$$\|\hat{\theta}^* - \theta^*\|_2 \leq \frac{a}{n} \cdot \|\theta^*\|_2 + \frac{b}{n}, \quad (25)$$

where a and b are functions of relevant parameters. In other words, we display a parameter regime in which a reasonably robust entity resolution will *not* significantly affect the optimal classifier. We assume without loss of generality that $\|\theta^*\|_2 \neq 0$, a property guaranteed to hold if the mean operator is not the null vector [Patrini et al., 2014].

Theorem 6 Suppose P_* is (ε, τ) -accurate and the data-model calibration assumption holds. Then the following holds:

$$\frac{\|\hat{\theta}^* - \theta^*\|_2}{\|\theta^*\|_2} \leq \frac{\xi}{n} \cdot T^2 \cdot \left(1 + \frac{\sqrt{\xi}}{4\|\theta^*\|_2 X_*} \cdot \rho\right). \quad (26)$$

If furthermore, P_* is α -bounded, then we get

$$\frac{\|\hat{\theta}^* - \theta^*\|_2}{\|\theta^*\|_2} \leq C(n) \cdot \left(1 + \frac{\delta_\rho}{\delta_m}\right), \quad (27)$$

with $C(n) \doteq (\xi/n)^\alpha$.

Function $C(n)$ is going to have key roles in the results to follow. It will be useful to keep in mind that, as long as we can sample and link more data by keeping a finite upperbound on ξ , we shall observe:

$$C(n) \rightarrow_{+\infty} 0, \forall \alpha \in (0, 1]. \quad (28)$$

³Follows from Cauchy-Schwartz inequality.

⁴A sufficient statistics for the class [Patrini et al., 2014].

The proof of Theorem 6 is quite long and detailed in Sections V.1 – V.4 of the appendix. It uses a helper theorem which we think is of independent interest for the study of entity resolution in a learning setting. Informally, it is an exact expression for $\hat{\theta}^* - \theta^*$ which holds regardless of P_* . We anticipate that this result may be interesting to optimize entity resolution in the context of learning. Theorem 6 essentially shows that $\|\hat{\theta}^* - \theta^*\|_2 / \|\theta^*\|_2 \rightarrow_n 0$ and is all the faster to converge as "classification gain beats permutation penalty" (δ_m/δ_ρ large). Ultimately, we see that learning can withstand bigger permutations ($\alpha \searrow$), provided it is accompanied by a sufficient decrease in the proportion of elementary permutations that act between classes ($\rho \searrow$): efficient entity resolution algorithms for federated learning should thus focus on trade-offs of this kind.

The key problem that remains is how does the drift $\hat{\theta}^* - \theta^*$ impact learning. For that objective, we start with two results with different flavours: first, an *immunity* of optimal large margin classification to the errors of entity resolution, and second, a bound for the difference of the Taylor losses of the optimal classifier (θ^*) and our classifier $\hat{\theta}^*$ *on the true data*, which shows strong convergence properties for $\ell_{S^*}(\hat{\theta}^*; \gamma, \Gamma)$ towards $\ell_{S^*}(\theta^*; \gamma, \Gamma)$.

7.4 Immunity of optimal large margin classification to P_*

We show that under the conditions of Theorem 6, large margin classification by the optimal classifier (θ^*) is immune to the impact of P_* on learning, in the sense that the related examples will be given the *same* class by $\hat{\theta}^*$ — the corresponding margin, however, may vary. We formalize the definition now.

Definition 7 Fix $\kappa > 0$. We say that $\hat{\theta}^*$ is **immune to P_* at margin κ** iff for any example (\mathbf{x}, y) , if $y(\theta^*)^\top \mathbf{x} > \kappa$, then $y(\hat{\theta}^*)^\top \mathbf{x} > 0$.

We can now formalize the immunity property.

Theorem 8 Suppose P_* is (ε, τ) -accurate and α -bounded, and the data-model calibration assumption holds. For any $\kappa > 0$, $\hat{\theta}^*$ is immune to P_* at margin κ if

$$n > \xi \cdot \left(\frac{\delta_m + \delta_\rho}{\kappa} \right)^{\frac{1}{\alpha}}. \quad (29)$$

(proof in Appendix, Section V.5) Eq. (29) is interesting for the relationships between n (data), ξ (permutation) and κ (margin) to achieve immunity. Consider a permutation P_* for which $\rho = 0$, that is, P_* factorizes as cycles that act within one class each. Since the maximal optimal margin within S is bounded by δ_m by Cauchy-Schwartz inequality, so we can let $\kappa \doteq \delta \cdot \delta_m$ where $0 < \delta < 1$ is the margin immunity parameter. In this case, (29) simplifies to the following constraint on δ to satisfy immunity at margin κ :

$$\delta > C(n), \quad (30)$$

where $C(n)$ is defined in Theorem 6. Roughly, increasing the domain size (n) without degrading the effects of P_* (ξ) can allow to significantly reduce the minimal immune margin.

7.5 Bound on the difference of Taylor losses for θ^* vs $\hat{\theta}^*$ on the true data

We essentially show that under the assumptions of Theorem 8, it holds that $\ell_{S^*}(\hat{\theta}^*; \gamma, \Gamma) - \ell_{S^*}(\theta^*; \gamma, \Gamma) = o(1)$, and the convergence is governed by $C(n)$. By means of words, the loss of our classifier (built over entity-resolved data) on the true data converges to that of the optimal classifier at a rate roughly proportional to $1/n^\alpha$.

Theorem 9 Suppose P_* is (ε, τ) -accurate and α -bounded, and the data-model calibration assumption holds. Then it holds that:

$$\ell_{S^*}(\hat{\theta}^*; \gamma, \Gamma) - \ell_{S^*}(\theta^*; \gamma, \Gamma) \leq \bar{\delta}_{m,\rho} (\delta_\mu + 6\bar{\delta}_{m,\rho}) \cdot C(n) , \quad (31)$$

where $\bar{\delta}_{m,\rho} \doteq (\delta_m + \delta_\rho)/2$ and $C(n)$ is defined in Theorem 6.

(proof in Appendix, Section V.6) We remark the difference with Theorem 8 that the bound also depends on a normalized sufficient statistics for the class of the true data (δ_μ).

7.6 Generalization abilities for classifiers learned from E/R'ed data

Suppose that ideal sample S^* is obtained i.i.d. from some unknown distribution \mathcal{D} , before it is "split" between A and B, and then reconstructed to form our training sample S . What is the generalization ability of classifier $\hat{\theta}^*$? This question is non-trivial because it entails the impact of entity resolution (E/R) on generalization, and not just on training, that is, we want to upperbound $\Pr_{(\mathbf{x}, y) \sim \mathcal{D}}[y(\hat{\theta}^*)^\top \mathbf{x} \leq 0]$ with high probability given that the data we have access to may not exactly reflect sampling from \mathcal{D} . The following Theorem provides such a bound.

Theorem 10 With probability at least $1 - \delta$ over the sampling of S^* according to \mathcal{D}^n , if the permutation P_* that links S and S^* is (ε, τ) -accurate and α -bounded and the data-model calibration assumption holds, then it holds that

$$\Pr_{(\mathbf{x}, y) \sim \mathcal{D}}[y(\hat{\theta}^*)^\top \mathbf{x} \leq 0] \leq \ell_{S^*}(\theta^*; \gamma, \Gamma) + \frac{2LX_*\theta_*}{\sqrt{n}} + \sqrt{\frac{\ln(2/\delta)}{2n}} + U(n) , \quad (32)$$

with L the Lipschitz constant of the Ridge-regularized Taylor loss, θ_* an upperbound on $\|\theta^*\|_2$ and

$$U(n) \doteq \bar{\delta}_{m,\rho} \cdot (\delta_{\mu_0} + 6\bar{\delta}_{m,\rho} + (4L/\sqrt{n})) \cdot C(n) . \quad (33)$$

$\bar{\delta}_{m,\rho}$ is defined in Theorem 9 and $C(n)$ is defined in Theorem 6.

(proof in Appendix, Section V.7) What is quite remarkable with that property is that we immediately have with probability at least $1 - \delta$, from Bartlett and Mendelson [2002]:

$$\Pr_{(\mathbf{x}, y) \sim \mathcal{D}}[y(\theta^*)^\top \mathbf{x} \leq 0] \leq \ell_{S^*}(\theta^*; \gamma, \Gamma) + \frac{2LX_*\theta_*}{\sqrt{n}} + \sqrt{\frac{\ln(2/\delta)}{2n}} , \quad (34)$$

i.e. the corresponding inequality in which we remove $U(n)$ in eq. (32) actually holds for substituting $\hat{\theta}^*$ by θ^* in the left hand side. In short, entity resolution affects generalization *only* through penalty $U(n)$. In addition, if P_* is "small" enough so that $\alpha \geq 1/2$, then we keep the slow-rate convergence ($O(1/\sqrt{n})$) of the E/R-free case. Otherwise, entity-resolution may impact generalization by slowing down this convergence.

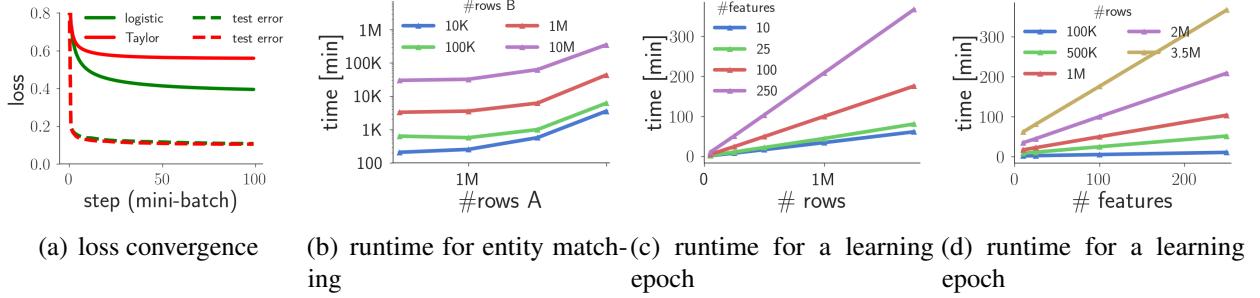


Figure 5: (a) Learning curve for Taylor vs. logistic loss (straight lines) and their test error (dotted); (b) runtime of entity matching with respect to the size of the two datasets; runtime of one learning epoch (all mini-batches + hold-out loss evaluation) with respect to number of examples (c) and features (d).

8 Experiments

We test our system experimentally and show that: i) the Taylor loss converges at a similar rate to the logistic loss in practice; ii) their accuracy is comparable at test time; iii) privacy-preserving entity resolution scales to the tens of millions of rows per data provider in a matter of hours; iv) federated logistic regression scales to millions of rows in the order of hours per epoch; v) the end-to-end system shows learning performance *on par* with learning on the perfectly linked datasets, the non-private setting.

Taylor vs. logistic — We have sketched the convergence rate of the Taylor loss in Section 7. Here we are interested in its experimental behaviour; see Figure 5(a). We run vanilla SGD on MNIST (odd vs. even digits) with constant learning rate $\eta = 0.05$ and $\Gamma = 10^{-2}I$ ridge regularization. The Taylor loss shows a very similar rate of convergence to the logistic loss. In both cases, features are standardized. Notice that the two losses converge to different minima *and* minimizers. What is important is to verify that at test time the two minimizers result in similar accuracy, which is indeed the case in Figure 5(a). In Table 2 we show that this is no coincidence. For several UCI datasets [Bache and Lichman, 2013], we standardize features and run SGD with the same hyper-parameters as above. Quantitative performance (accuracy and AUC) of the two losses differ by at most 1.8 points.

Scalability — For the rest of this Section, empirical results will be based on a “quasi-real scenario”, an augmented version of Kaggle [2011]’s *Give me some credit* dataset. The original problem involves classifying whether a customer will default on a loan. We call this benchmark dataset PACS (PI Augmented Credit Scoring): realistic PIs, potentially missing and with typos, are generated for the entities by Febrl [Christen, 2008]. The data is then split vertically in two halves of the 12 original features. The size of the dataset is about 160K examples.

In order to test the scalability of each component of our pipeline, we upsample the dataset and augment the feature sets by feature products, and record the runtime of our implementations of entity resolution and logistic regression solve by SAG. For these experiments we run each party of the protocol on a separate AWS machine with 16 cores and 60GiB of memory. We use the Paillier encryption scheme as our additively homomorphic scheme with a key size of 1024 bits, from the implementation of python-paillier [2017].

First, for benchmarking the entity resolution algorithm, we make a grid of datasets with sizes

dataset	#rows	#feat.	logistic		Taylor (delta)	
			acc	AUC	acc	AUC
<i>iris</i>	100	3	98.0	100	+0.0	+0.0
<i>diabetes</i>	400	10	69.0	79.3	+0.0	+0.5
<i>bostonhousing</i>	400	13	76.4	98.2	+0.0	+0.1
<i>breastcancer</i>	400	30	94.7	98.4	+1.7	+0.3
<i>digits</i>	1.5K	64	92.3	97.0	+1.0	+0.4
<i>simulated</i>	8K	100	68.8	76.0	-0.1	+0.0
<i>mnist</i>	60K	784	89.2	95.8	+0.1	-0.1
<i>20newsgroups</i>	10K	5K	51.7	66.8	-0.2	-1.8
<i>covertype</i>	500K	54	72.2	79.1	-0.1	+0.1

Table 2: Taylor vs. logistic loss on UCI datasets. When the original problem is multi-class or regression, we cast it into binary classification.

spanning on a log-scale from $1K$ to $10M$. The entity matching only acts on the PI attributes. In Figure 5(b) we measure the runtime of the combined construction of CLKs, network communication to C and matching two datasets. The coordinating party uses a cluster of 8 machines for matching. As expected, entity resolution scales quadratically in the number of rows, due to the number of comparisons that are needed for brute-force matching. Due to the embarrassingly parallel workload the C component can handle matches with datasets up to $10M$ rows in size in a matter of hours.

Second, we measured the runtime of learning. We generated multiple versions of the dataset varying the number of rows and features; while rows are simply copied multiple times, we create new features by multiplying the original ones. We are not interested in convergence time here but only on a scalability measure on an epoch basis; the recorded runtime accounts for updating the model for all mini-batches and one hold-out loss computation. See Figures 5(c) and 5(d). Runtime grows linearly in both scales. This is the bottleneck of the whole system, due to the expensive encrypted operations and the communication of large encrypted numbers between the parties. We estimated that encryption amounts to a slow down of about two orders of magnitude.

The communication costs of one epoch consists of

$$\begin{aligned} \text{cost}_{\nabla} &\leq (2n + 2\lceil n/s' \rceil d)ct \\ \text{cost}_{\ell} &= (h_s + 2)ct \end{aligned}$$

where ct describes the size of one cipher text. Thus, for $s' = d = 100$, $n = 1$ M and a cipher text size of 256 bytes, the overall communication costs for one epoch a just under 1 GB.

Benchmark — Finally, we test the whole system on three version of PACS, where we control the percent of shared entities between A and B, in the range of $\{100, 66, 33\}\%$. The same test set is used for evaluating performance; labels in the test set are unbalanced (93/7%) while the training set is artificially balanced. Table 3 shows our system performs exactly *on par* with logistic regression run on perfectly linked datasets (the non-private setting), with respect to accuracy, AUC and f1-score. We also record the percent for wrongly linked entities by our entity matching algorithm: around 1% of the entities is linked as a mistake. The main take-away message is that those mistakes are not detrimental for learning; hence, we support the claim of Theorem 6 on a practical ground. (Notice also how the fraction of shared entities does not seem to matter, arguably due to the fact that we

dataset	logistic regression			our (delta)			matching errors (%)
	acc	AUC	f1	acc	AUC	f1	
PACS-33	88.5	80.4	36.9	+0.1	+0.0	+0.1	1.0
PACS-66	88.6	80.3	36.9	-0.1	+0.0	+0.1	0.9
PACS-100	88.6	80.3	37.2	+0.1	-0.0	+0.0	0.8

Table 3: Our system vs. logistic regression on perfectly linked data on PACS. We also record the percent of wrong matches by our entity resolution algorithm.

learn a low dimensional linear model which cannot capitalize the large training set.)

9 Conclusion

We have shown how to learn a linear classifier in a privacy-preserving federated fashion when data is vertically partitioned. We addressed the problem end-to-end, by pipelining entity resolution and distributed logistic regression by using Paillier encryption. All records, including identifiers required for linkage and the linkage map itself, remain confidential from other parties. To the best of our knowledge, our system is the first scalable and accurate solution to this problem. Importantly, we do not introduce extrinsic noise for gaining privacy, which would hinder predictive performance; differential privacy could, however, be applied on top of our protocol when the model itself is deemed susceptible to malicious attacks. On the theory side, we provide the first analysis of the impact of entity resolution on learning, with a potential for further use for the design of entity matching methods specifically targeted to learning. Our results show that the picture of learning is not changed under reasonable assumptions on the magnitude and size of mistakes due to entity resolution — even rates for generalization can remain of the same order. More: errors introduced during entity resolution do not impact the examples with large margin classification by the optimal (unknown) classifier: they are given the same class by our classifier as well. Since federated learning can dramatically improve the accuracy of this optimal classifier (because it is built over a potentially much larger set of features), the message this observation carries is that it can be extremely beneficial to carry out federated learning in the setting where each peer’s data provides a significant uplift to the other. Our results signal the existence of non-trivial tradeoffs for entity-resolution to be optimized with the objective of learning from linked data. We hope such results will contribute to spur related research in the active and broad field of entity resolution.

References

- Y. Aono, T. Hayashi, L. Trieu Phong, and L. Wang. Scalable and secure logistic regression via homomorphic encryption. In *CODASPY*, 2016.
- K. Bache and M. Lichman. Uci machine learning repository, 2013. <http://archive.ics.uci.edu/ml>.
- P.-L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2002.

- M. Ben-Or, S. Goldwasser, and A. Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation. In *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*, 1988.
- R. Bhatia. *Matrix Analysis*. Springer, 1997.
- BlueKrypt, 2017. v 30.4 - February 23, 2017 www.keylength.com.
- K. Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In *NIPS*, 2009.
- P. Christen. Febrl: a freely available record linkage system with a graphical user interface. In *Proceedings of the second Australasian workshop on Health data and knowledge management-Volume 80*. Australian Computer Society, Inc., 2008.
- P. Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.
- I. Damgård, M. Geisler, and M. Krøigaard. A correction to “Efficient and Secure Comparison for On-Line Auctions”. Cryptology ePrint Archive, Report 2008/321, 2008. <http://eprint.iacr.org/2008/321>.
- J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *FOCS*, 2013.
- D. Duverle, S. Kawasaki, Y. Yamada, J. Sakuma, and K. Tsuda. Privacy-preserving statistical analysis by exact logistic regression. In *IEEE Security and Privacy Workshops (SPW)*, 2015.
- C. Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, 2008.
- C. Dwork and A. Roth. The algorithmic foudations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9:211–407, 2014.
- P. M. Esperança, L. J. M. Aslett, and C. C. Holmes. Encrypted accelerated least squares regression. In *AISTATS’17*, pages 334–343, 2017.
- A. Gascón, P. Schoppmann, B. Balle, M. Raykova, J. Doerner, S. Zahur, and D. Evans. Privacy-preserving distributed linear regression on high-dimensional data. *PoPET*, 2017.
- C. Gentry. Fully homomorphic encryption using ideal lattices. In *Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing*, 2009.
- R. Hall and S. E. Fienberg. Privacy-preserving record linkage. In *International conference on privacy in statistical databases*, 2010.
- Kaggle. Give me some credit, 2011. www.kaggle.com/c/GiveMeSomeCredit.
- S. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*21*, pages 793–800, 2008.

- J. Konecný, H. B. McMahan, D. Ramage, and P. Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *CoRR*, abs/1610.02527, 2016.
- M. Kuzu, M. Kantarcioglu, E. Durham, and B. Malin. A constraint satisfaction cryptanalysis of bloom filters in private record linkage. In *PETS*, Berlin, Heidelberg, 2011.
- M. Kuzu, M. Kantarcioglu, E. Durham, C. Toth, and B. Malin. A practical approach to achieve private medical record linkage in light of public resources. *Journal of the American Medical Information Association*, 2013.
- H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017.
- P. Mohassel and Y. Zhang. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 19–38, May 2017. doi: 10.1109/SP.2017.12.
- F. Niedermeyer, S. Steinmetzer, M. Kroll, and R. Schnell. Cryptanalysis of basic bloom filters used for privacy preserving record linkage. *Journal of Privacy and Confidentiality*, 2014.
- R. Nock. On regularizing rademacher observation losses. In *NIPS*, 2016.
- R. Nock and F. Nielsen. Bregman divergences and surrogates for learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.
- R. Nock, G. Patrini, and A. Friedman. Rademacher observations, private data, and boosting. In *ICML*, 2015.
- P. Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *International Conference on the Theory and Applications of Cryptographic Techniques*, 1999.
- G. Patrini. *Weakly supervised learning via statistical sufficiency*. PhD thesis, The Australian National University, 2016.
- G. Patrini, R. Nock, P. Rivera, and T. Caetano. (Almost) no label no cry. In *NIPS*27*, 2014.
- G. Patrini, F. Nielsen, R. Nock, and M. Carioni. Loss factorization, weakly supervised learning and label noise robustness. In *ICML*, 2016a.
- G. Patrini, R. Nock, S. Hardy, and T. Caetano. Fast learning from distributed datasets without entity matching. In *IJCAI*, 2016b.
- python-paillier, 2017. release v1.3, github.com/n1analytics/python-paillier.
- M. Schmidt, N. L. Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 2013.
- R. Schnell. Efficient private record linkage of very large datasets. In *59th World Statistics Congress*, 2013.

- R. Schnell, T. Bachteler, and J. Reiher. A novel error-tolerant anonymous linking code. Technical report, Paper No. WP-GRLC-2011-02, German Record Linkage Center Working Paper Series, 2011.
- D. Vatsalan, P. Christen, and V. S. Verykios. A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, 2013a.
- D. Vatsalan, P. Christen, and V. S. Verykios. Efficient two-party private blocking based on sorted nearest neighborhood clustering. In *CIKM '13 Proceedings of the 22nd ACM international Conference on Information and Knowledge Management*, 2013b.
- W. E. Winkler. Record linkage. In *Handbook of Statistics*, pages 351–380. Elsevier, 2009.
- S. Wu, T. Teruya, J. Kawamoto, J. Sakuma, and H. Kikuchi. Privacy-preservation for stochastic gradient descent application to secure logistic regression. In *27th Annual Conference of the Japanese Society for Artificial Intelligence*, 2013.
- W. Xie, Y. Wang, S. M. Boker, and D. E. Brown. Privlogit: Efficient privacy-preserving logistic regression by tailoring numerical optimizers. *CoRR*, 2016.
- A. Yao. How to generate and exchange secrets. *27th Annual Symposium on Foundations of Computer Science (SFCS)*, 1986.

Appendix

Theorems and Lemmata are numbered with letters (A, B, ...) to make a clear difference with the main file numbering.

Table of contents

Paillier encryption scheme	Pg 26
Encoding	Pg 27
Secure initialization and loss computation	Pg 29
Security evaluation	Pg 30
Proofs	Pg 32
→ Notations for proofs	Pg 32
→ Helper Theorem	Pg 36
→ Assumptions: details and discussion	Pg 41
→ Proof of Theorem 6	Pg 42
→ Proof of Theorem 8	Pg 54
→ Proof of Theorem 9	Pg 55
→ Proof of Theorem 10	Pg 56
Cryptographic longterm keys, entity resolution and learning	Pg 59

I Paillier encryption scheme

We describe in more detail the Paillier encryption scheme Paillier [1999], an asymmetric additively homomorphic encryption scheme. In what follows, \mathbb{Z}_t will denote the ring of integers $\{0, 1, \dots, t-1\}$, with addition and multiplication performed modulo the positive integer t , and \mathbb{Z}_t^* will denote the set of invertible elements of \mathbb{Z}_t (equivalently, elements that do not have a common divisor with t).

In the Paillier system the private key is a pair of randomly selected large prime numbers p and q , and the public key is their product $m = pq$. The plaintext space is the set \mathbb{Z}_m , the ciphertext space is $\mathbb{Z}_{m^2}^*$, and the encryption of an element $x \in \mathbb{Z}_m$ is given by:

$$x \mapsto g^x r_x^m \quad (35)$$

where g is a generator of $\mathbb{Z}_{m^2}^*$ (which can be taken to be $m+1$) and r_x is an element of \mathbb{Z}_m^* selected uniformly at random (the subscript x is to indicate that a new value should be selected for every input). Note that any r_x produces a valid ciphertext, so in particular many valid ciphertexts correspond to a given plaintext. Let the encryption of a number x be $\llbracket x \rrbracket$. It follows immediately from (35) that, for elements $x, y \in \mathbb{Z}_m$,

$$\llbracket x \rrbracket \llbracket y \rrbracket = g^{x+y} (r_x r_y)^m = \llbracket x + y \rrbracket, \quad (36)$$

whence we define the operator ‘ \oplus ’ by $\llbracket x \rrbracket \oplus \llbracket y \rrbracket = \llbracket x + y \rrbracket$. (In Section I we denoted this operator simply as ‘ $+$ ’.) Then, by linearity,

$$\bigoplus_{i=1}^y \llbracket x \rrbracket = \llbracket \sum_{i=1}^y x \rrbracket = \llbracket xy \rrbracket. \quad (37)$$

Note that in this latter case, y is not encrypted. Suppose a party is sent $\llbracket x \rrbracket$ and computes $\llbracket xy \rrbracket$ from $\llbracket x \rrbracket$ and y using (37). If the party who originally computed $\llbracket x \rrbracket$ subsequently obtains $\llbracket xy \rrbracket$, then they can verify whether a guess y' for y is correct by calculating $\bigoplus_{i=1}^{y'} \llbracket x \rrbracket$ and comparing it with $\llbracket xy \rrbracket$. In general recovering y here is equivalent to the discrete logarithm problem, but if y is known to come from a small set then this becomes simple to solve. We can protect y when “multiplying by y ” by adding $\llbracket 0 \rrbracket$, forcing the random multiplier r to change. Hence we define multiplication by:

$$\llbracket x \rrbracket y = \llbracket xy \rrbracket \oplus \llbracket 0 \rrbracket. \quad (38)$$

It is worth reiterating that all operations that occur *inside* the $\llbracket \cdot \rrbracket$ occur in the plaintext space \mathbb{Z}_m , hence modulo the public key m , while all operations that occur *outside* the $\llbracket \cdot \rrbracket$ occur in the ciphertext space \mathbb{Z}_{m^2} , hence modulo m^2 .

For a vector $\mathbf{x} = (x_i)_i$, the notation $\llbracket \mathbf{x} \rrbracket$ should be interpreted component-wise, that is $\llbracket \mathbf{x} \rrbracket = (\llbracket x_i \rrbracket)_i$. We also extend the definition of \oplus to operate component-wise on vectors, and of multiplication to perform scalar multiplication: $s \llbracket \mathbf{x} \rrbracket = \llbracket s\mathbf{x} \rrbracket = \llbracket s \rrbracket \mathbf{x}$. Additionally, we define:

$$\llbracket \mathbf{x} \rrbracket \circ \mathbf{y} = (\llbracket x_i \rrbracket y_i)_i = \llbracket \mathbf{x} \circ \mathbf{y} \rrbracket \quad (39)$$

to be the component-wise (or “Hadamard”) product with a plaintext vector and:

$$\llbracket \mathbf{x} \rrbracket \odot \mathbf{y} = \bigoplus_i \llbracket x_i \rrbracket y_i = \llbracket \mathbf{x}^\top \mathbf{y} \rrbracket \quad (40)$$

to be the “dot product” of an encrypted vector with a plaintext vector. (In Section I we omitted the symbol for this operator, as it would for a standard product.) We can also extend the definition of “ $\llbracket \cdot \rrbracket$ ” to express matrix multiplication: For matrices \mathbf{A} and \mathbf{B} with compatible dimensions,

$$\begin{aligned}\llbracket \mathbf{AB} \rrbracket &= \mathbf{A} \llbracket \mathbf{B} \rrbracket \\ &= (\mathbf{A}_i \odot \llbracket \mathbf{B}^j \rrbracket)_{i,j} = (\llbracket \mathbf{A}_i \rrbracket \odot \mathbf{B}^j)_{i,j} = \llbracket \mathbf{A} \rrbracket \mathbf{B}\end{aligned}\tag{41}$$

is the usual product of two matrices. In particular, for a vector \mathbf{x} , we can compute: $\llbracket \mathbf{Ax} \rrbracket = \llbracket \mathbf{A} \rrbracket \mathbf{x} = \mathbf{A} \llbracket \mathbf{x} \rrbracket$ or $\llbracket \mathbf{xA} \rrbracket = \llbracket \mathbf{x} \rrbracket \mathbf{A} = \mathbf{x} \llbracket \mathbf{A} \rrbracket$ depending on when the dimensions are compatible.

There is, of course, a decryption process corresponding to the encryption process of (35), but the details are out of scope here. Suffice it to say that, computationally speaking, the cost of a decryption is in the order of a single modular exponentiation modulo m .

II Encoding

The Paillier cryptosystem only provides arithmetic for elements of \mathbb{Z}_m . We give details on our encoding of floating-point numbers.

Fix an integer $\beta > 1$, which will serve as a *base*, and let q be any fraction. Any pair of integers (s, e) satisfying $q = s\beta^e$ is called a *base- β exponential representation* of q . The values s and e are called the *significand* and *exponent* respectively. Let $q = (s, e)$ and $r = (t, f)$ be the base- β exponential representations of two fractions and without loss of generality assume that $f \geq e$. Then $(s + t\beta^{f-e}, e)$ is a representation of $q + r$ which follows by using the equivalent representation $(t\beta^{f-e}, e)$ for (t, f) . A representation for qr is simply given by $(st, e + f)$.

In order to be compatible with the Paillier encryption system we will require the significand s to satisfy $0 \leq s < m$ where m is the public key. This limits the precision of the encoding to $\log_2 m$ bits, which is not a significant impediment as a reasonable choice of m is at least 1024 bits [BlueKrypt, 2017]. Let:

$$q = \sigma \varphi 2^\varepsilon,\tag{42}$$

be a unbiased IEEE 754 floating-point number, where

$$\sigma = \pm 1, \quad \varphi = 1 + \sum_{i=1}^d b_{d-i} 2^{-i},\tag{43}$$

and $b_j \in \{0, 1\}$ for all $j = 0, \dots, d - 1$; for single precision $d = 23$ and $-127 < \varepsilon < 128$, while for double precision $d = 52$ and $-1023 < \varepsilon < 1024$. (We do not treat subnormal numbers, so we can treat $\varepsilon = 0$ specially as representing $q = 0$.) An *encoding* of q relative to a Paillier key m is a base- β exponential representation computed as follows. If $q = 0$, set $s = e = 0$. Otherwise, let $e = \lfloor (\varepsilon - d + 1) \log_\beta 2 \rfloor$ and $s' = \varphi \beta^{-e}$; one can verify that this choice of e is the smallest such that $s' \geq 1$. Finally, set $s = s'$ if $\sigma = 1$ and $s = m - s'$ if $\sigma = -1$. Then (s, e) is a base- β exponential representation for q with $0 \leq s < m/2$ when $q \geq 0$ and $m/2 \leq s < m$ when $q < 0$. This consciously mimics the familiar two’s compliment encoding of integers. The advantage of this encoding (over, say, a fixed-point encoding scheme) is that we can encode a very large range of values.

Since the significand must be less than m , it can happen that the sum or product of two encodings is too big to represent correctly and *overflow* occurs. Specifically, for a fraction $q = (s, e)$, if $|q|/\beta^e \geq m/2$ then overflow occurs. In practice we start with fractions with 53 bits of precision (for a double precision IEEE 754 float) and the Paillier public key m allows for $\log_2(m/2) = 1023$ bits of precision so we can accumulate $\lfloor 1023/53 \rfloor = 19$ multiplications by 53-bit numbers before overflow might happen. Overflow cannot be detected in the scheme described here. One partial solution to detecting overflow is to reserve an interval between the positive and negative regions and to consider a number to have overflowed if its value is in that region. This method does not detect numbers that overflow beyond the reserved interval, and has the additional drawback of reducing the available range of encoded numbers.

We *define* the encryption of an encoded number $q = (s, e)$ to be $\llbracket q \rrbracket = (\llbracket s \rrbracket, e)$. Note in particular that we encrypt the significand but *we do not encrypt the exponent*, so some information about the order of magnitude of an encrypted number is public. There are protocols that calculate the maximum of two encrypted numbers (see, for example, [Damgård et al., 2008]) which could be used to calculate the maximum of encrypted exponents (needed during addition as seen above). However it would be prohibitively expensive to compute this maximum for every arithmetic operation, so instead we leave the exponents public and leak the order of magnitude of each number.

To see more precisely how much information is leaked, consider an encrypted number $\llbracket q \rrbracket = (\llbracket s \rrbracket, e)$ encoded with respect to the base β . The exponent satisfies $e = \lfloor (\varepsilon - d + 1) \log_\beta 2 \rfloor$ (with $d = 23$ or 52 as above), hence:

$$c \leq \varepsilon < c + \log_2 \beta \quad (44)$$

where $c = e \log_2 \beta + d - 1$ is public information. Thus the original exponent ε is known to be in a range of length $\log_2 \beta$, and hence we learn that:

$$2^c \leq q < 2^{c+1} \beta, \quad (45)$$

that is, q is within a factor of β of 2^c . (Note that $e = 0$ includes the special case of $q = 0$.) The extent of the leakage can be mitigated in several ways, for example by choosing a large enough base so that no significant detail can be recovered, or by fixing the exponent for all encoded numbers as in fixed-point encodings.

Using the arithmetic of encoded numbers defined above, and assuming $f \geq e$, we can extend the definitions of ‘ \oplus ’ and multiplication by:

$$(\llbracket s \rrbracket, e) \oplus (\llbracket t \rrbracket, f) = (\llbracket s \rrbracket \oplus \llbracket t \rrbracket \beta^{f-e}, e) \quad (46)$$

and

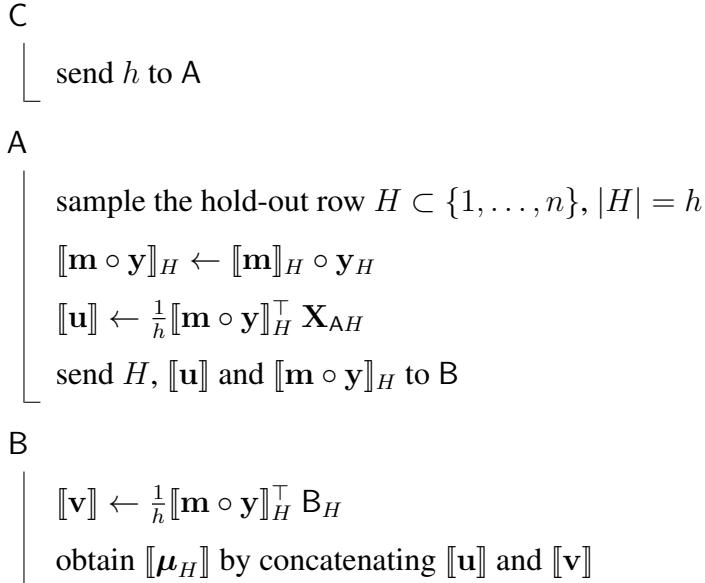
$$(\llbracket s \rrbracket, e)(t, f) = (\llbracket s \rrbracket t, e + f). \quad (47)$$

Note that the definition of ‘ \oplus ’ in (46) has the unfortunate consequence of requiring a multiplication of $\llbracket t \rrbracket$ whenever $f \neq e$. It is clear that addition does not change the leakage range (it is still β), while multiplication increases the range to β^2 , mitigating the leakage by a factor of β .

Algorithm 4: Secure loss initialization

Data: Hold-out size h

Result: Caching of $\llbracket \boldsymbol{\mu} \rrbracket$ of the (undisclosed) hold-out H



III Secure initialization and loss computation

III.1 Secure loss initialization

Algorithm 4 performs the secure loss initialization evaluation. Since $\boldsymbol{\mu}_H$ is independent from $\boldsymbol{\theta}$, we can cache for future computation of the logistic loss. To see why Algorithm 4 is correct, note that:

$$\llbracket \mathbf{u} \rrbracket = \frac{1}{h} \llbracket \mathbf{m} \circ \mathbf{y} \rrbracket_H^\top \mathbf{X}_{AH} = \frac{1}{h} \llbracket (\mathbf{m} \circ \mathbf{y})_H^\top \mathbf{X}_{AH} \rrbracket. \quad (48)$$

Similarly, $\llbracket \mathbf{v} \rrbracket = \frac{1}{h} \llbracket (\mathbf{m} \circ \mathbf{y})_H^\top \mathbf{X}_{BH} \rrbracket$, and so, after concatenating, B obtains $\llbracket \boldsymbol{\mu}_H \rrbracket = \frac{1}{h} \llbracket (\mathbf{m} \circ \mathbf{y})_H^\top \mathbf{X}_H \rrbracket$. B only receives the encrypted forms of $(\mathbf{m} \circ \mathbf{y})_H$ and $\frac{1}{h}(\mathbf{m} \circ \mathbf{y})_H^\top \mathbf{X}_{AH}$ from A and only holds the final result $\llbracket \boldsymbol{\mu}_H \rrbracket$ in encrypted form which is not sent to the C.

III.2 Secure logistic loss

Algorithm 5 computes the secure logistic loss. It is called in Algorithm 2, once per iteration of the outer loop.

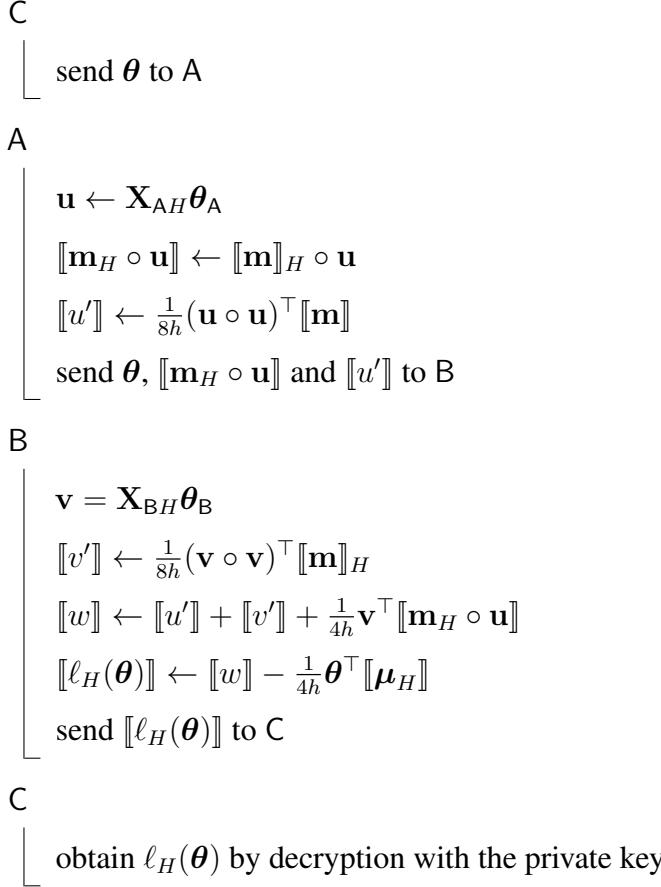
To show the correctness of Algorithm 5 first note that, for $i \in H$, the i th component of $\llbracket \mathbf{m}_H \circ \mathbf{u} \rrbracket$ is $\llbracket m_i \boldsymbol{\theta}_A^\top \mathbf{x}_{iA} \rrbracket$ and $u' = \frac{1}{8h} \sum_i m_i (\boldsymbol{\theta}_A^\top \mathbf{x}_{iA})^2$; similarly for \mathbf{v} and v' *mutatis mutandis*. Hence

$$\begin{aligned} w_0 &= u' + v' + \frac{1}{4h} ((\mathbf{m}_H \circ \mathbf{u})^\top \mathbf{v}) \\ &= \frac{1}{8h} \sum_{i \in H} m_i (\boldsymbol{\theta}_A^\top \mathbf{x}_i)^2, \end{aligned} \quad (49)$$

Algorithm 5: Secure logistic loss

Data: Model $\boldsymbol{\theta}$, requires $\llbracket \boldsymbol{\mu}_H \rrbracket$ and H cached by Algorithm 4

Result: $\ell_H(\boldsymbol{\theta})$ of the (undisclosed) hold-out



since

$$(\boldsymbol{\theta}^\top \mathbf{x}_i)^2 = (\boldsymbol{\theta}_A^\top \mathbf{x}_{iA})^2 + (\boldsymbol{\theta}_B^\top \mathbf{x}_{iB})^2 + 2(\boldsymbol{\theta}_A^\top \mathbf{x}_{iA})(\boldsymbol{\theta}_B^\top \mathbf{x}_{iB}) \quad (50)$$

for every row \mathbf{x} of \mathbf{X} by (12). Then $w = w_0 - \frac{1}{4} \boldsymbol{\theta}^\top \boldsymbol{\mu}_H \approx \ell(\boldsymbol{\theta})$ by (10), which can be computed since $\llbracket \boldsymbol{\mu}_H \rrbracket$ was previously computed by B in Algorithm 4.

C only sees the final result $\ell_H(\boldsymbol{\theta})$. A and B see the model $\boldsymbol{\theta}$, and B receives only the encrypted forms of the values $m_i \boldsymbol{\theta}_A^\top \mathbf{x}_A$ and $\sum_i m_i (\boldsymbol{\theta}_A^\top \mathbf{x}_A)^2$ (for $i \in H$) from A.

IV Security evaluation

It is important to clarify how the data is distributed and what is visible to the three parties. Table A4 gives a summary of which values each party will see during an execution of Algorithm 2. Most entries will be obvious, with the possible exception of the column for $\nabla \ell'_S(\boldsymbol{\theta})$. The values for $\nabla \ell'_S(\boldsymbol{\theta})$ can be derived from the successive values of $\boldsymbol{\theta}$, which are seen by all parties. If $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are successive values of the weight vector, then $\nabla \ell'_S(\boldsymbol{\theta}) = \eta^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}') - \gamma \boldsymbol{\theta}$ from line 2; the values

agent	A	B	\mathbf{m}	n	θ	H	S	S'	μ_H	ℓ_H	$\nabla \ell'_S$
A	✓			✓	✓	✓	✓	✓			✓
B		✓		✓	✓	✓	✓	✓		✓	✓
C			✓	✓	✓					✓	✓

Table A4: Summary of data visibility for each agent.

of η and γ are not private. Notice how μ_H is known by nobody, and it is accessible by B only in encrypted form.

Different protection mechanisms are in place to secure the communication and protect against information leakage among the three parties.

The matched entities are protected from A and B by the implementation of a private entity matching algorithm which creates a random permutation for A and for B, and a mask received only by C. During the private entity matching algorithm, the information from A and B are protected from C by the construction of the CLKs.

Additively homomorphic encryption protects the values in A’s data from B, and vice-versa. This does not hold for C, who owns the private key. Protection from C is achieved by sending only computed values which are not considered private (such as the gradient which is shared with all participants).

We finally note that, while the data is protected via different mechanisms, our work does not protect against the leakage of information through the model and its iterative computation.

Potential leakage of information — We analyze two sources of leakage:

- **Number of matches in a mini batch:** After the entity resolution, the data providers only receive an encrypted mask and a permutation. They do not learn the total number of matches between them. Thus, when the first data provider chooses the mini batches at random, the number of matches in each of the batches is unknown. However, this can lead to a potential leak of information, as the corresponding gradient will be revealed in the clear. If, for instance, this gradient is zero, then both data provider know, that they do not have any entities in common within this mini batch. Or a mini batch only contains one match, then the gradient reveals the corresponding label. This kind of leakage can be mitigated by carefully choosing the batch size. Let X be a random variable describing the number of matches in a mini batch, then X follows the hypergeometric distribution. Let $\text{CDF}_{\text{hyper}}(N, K, n, k)$ be its cumulative density function with N is the population size, K is the number of success states in the population, n is the number of draws, and k is the number of observed successes. Then the probability that a mini batch contains not more than k matches is:

$$P[X \leq k] = \text{CDF}_{\text{hyper}}(n, M, s', k).$$

M stands of the total number of matches. Note that the coordinator has all the necessary information available, and could therefore test if the probability for the chosen parameters is acceptably low or otherwise abort the protocol.

- **Public exponent in encoding of encrypted numbers:** The floating point inspired encoding scheme for encrypted numbers leaks information about the plaintext through a public exponent.

We quantify the leakage in Appendix II and want to point out that the leakage can be reduced by increasing the value for the base. However we do admit that this form of leakage is unnecessary, as we are confident that the algorithm will also work with a fixed-point encoding scheme. We plan to address this issue in a future version.

V Proofs

V.1 Notations

The proofs being quite heavy in notations, we define two new notations (on the observation matrix and training sample) that will be more convenient than in the main file.

(observation matrix) the observation matrix is now transposed — it has observations in columns. Thus, instead of writing as in eq. (11) (main file):

$$\mathbf{X} = [\mathbf{X}_A \mid \mathbf{X}_B] , \quad (51)$$

We let from now on, with a matrix notation,

$$\mathbf{X} \doteq \begin{bmatrix} \mathbf{X}_A \\ \mathbf{X}_B \end{bmatrix} \quad (52)$$

denote a block partition of \mathbf{X} , where $\mathbf{X}_A \in \mathbb{R}^{d_A \times m}$ and $\mathbf{X}_B \in \mathbb{R}^{d_B \times m}$ (so, $\mathbf{X}_A = \mathbf{X}_A^\top$, $\mathbf{X}_B = \mathbf{X}_B^\top$).

(training and ideal sample) Instead of S^* to denote the ideal sample, we let S denote the ideal sample, and the training sample we have access to (produced via error-prone entity resolution) is now \hat{S} instead of S .

We do not observe \mathbf{X} but rather an estimate through *approximately accurate linkage*,

$$\hat{\mathbf{X}} \doteq \begin{bmatrix} \mathbf{X}_A \\ \hat{\mathbf{X}}_B \end{bmatrix} , \quad (53)$$

where

$$\hat{\mathbf{X}}_B \doteq \mathbf{X}_B P_* , \quad (54)$$

where $P_* \in \{0, 1\}^{n \times n}$ is a permutation matrix. To refer to the features of A and B with words we shall call them the *anchor* (A) and *shuffle* (B) feature spaces.

Any permutation matrix can be factored as a product of elementary permutation matrices. So suppose that

$$P_* = \prod_{t=1}^T P_t , \quad (55)$$

where P_t is an elementary permutation matrix, where $T < n$ is unknown. \hat{X} can be progressively constructed from a sequence $\hat{X}_0, \hat{X}_1, \dots, \hat{X}_T$ where $\hat{X}_0 = X$ and for $t \geq 1$,

$$\hat{X}_t \doteq \begin{bmatrix} X_A \\ \hat{X}_{tB} \end{bmatrix} , \hat{X}_{tB} \doteq X_B \prod_{j=1}^t P_j . \quad (56)$$

We recall that by convention, permutations act on the shuffle set B of features, and labels are associated with A and therefore are not affected by the permutation (without loss of generality). Let

$$\hat{X}_t \doteq [\hat{x}_{t1} \ \hat{x}_{t2} \ \cdots \ \hat{x}_{tn}] , \quad (57)$$

denote the column vector decomposition of \hat{X}_t (with $\hat{x}_{0i} \doteq x_i$) and let \hat{S}_t be the training sample obtained from the t first permutations in the sequence ($\hat{S}_0 \doteq S$). Hence, $\hat{S}_t \doteq \{(\hat{x}_{ti}, y_i), i = 1, 2, \dots, n\}$.

Definition K *The mean operator associated to \hat{S}_t is $\mathbb{R}^d \ni \mu_t \doteq \mu(\hat{S}_t) \doteq \sum_i y_i \cdot \hat{x}_{ti}$.*

The mean operator is a sufficient statistics for the class in linear models [Patrini et al., 2014]. We can make at this point a remark that is going to be crucial in our results, and obvious from its definition: the mean operator is *invariant* to permutations made within classes, *i.e.* $\mu_T = \mu_0$ if P_* factorizes as two permutations, one affecting the positive class only, and the other one affecting the negative class only. Since the optimal classifier for the Taylor loss is a linear mapping of the mean operator, we see that the drift in its optimal classifier due to permutations will be due *only* to mistakes in the linear mapping when P_* factorizes in such a convenient way. Now, consider elementary permutation P_t . We call *pairs* for P_t the two column indexes affected by P_t . Namely,

- (a) we denote $u_A(t)$ and $v_A(t)$ the two column indexes of the observations in X affected by elementary permutation P_t ;
- (b) we denote $u_B(t)$ and $v_B(t)$ the two column indexes of the observations in X such that *after* P_t , the shuffle part of column $u_A(t)$ (resp. $v_A(t)$) is the shuffle part of $x_{u_B(t)}$ (resp. $x_{v_B(t)}$).

To summarize (b), we have the following Lemma.

Lemma L *The following holds for any $t \geq 1$:*

$$(\hat{x}_{tu_A(t)})_B = (x_{u_B(t)})_B , \quad (58)$$

$$(\hat{x}_{tv_A(t)})_B = (x_{v_B(t)})_B . \quad (59)$$

Figure 6 shows an example of our notations.

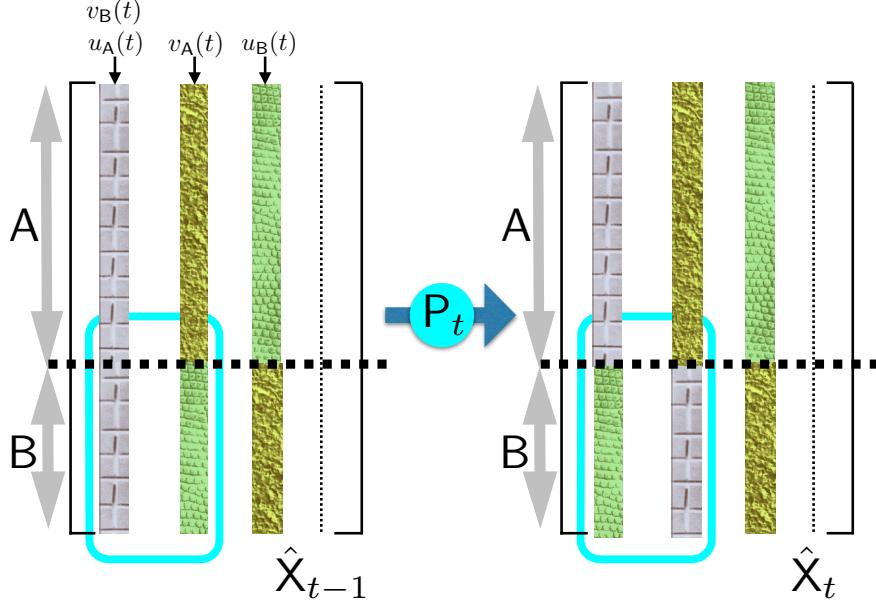


Figure 6: Permutation P_t applied to observation matrix \hat{X}_{t-1} and subsequent matrix \hat{X}_t , using notations $u_A(t)$, $u_B(t)$, $v_A(t)$ and $v_B(t)$. Anchor textures denote references for the textures in B (best viewed in color).

Example M Denote for short $\{0, 1\}^{n \times n} \ni \Theta_{u,v} \doteq \mathbf{1}_u \mathbf{1}_v^\top + \mathbf{1}_v \mathbf{1}_u^\top - \mathbf{1}_v \mathbf{1}_v^\top - \mathbf{1}_u \mathbf{1}_u^\top$ (symmetric) such that $\mathbf{1}_u$ is the u^{th} canonical basis vector of \mathbb{R}^n . For $t = 1$, it follows

$$u_B(1) = v_A(1) , \quad (60)$$

$$v_B(1) = u_A(1) . \quad (61)$$

Thus, it follows:

$$\begin{aligned} & \mathbf{X}_A \Theta_{u_A(1), v_A(1)} \hat{X}_{1B}^\top \\ &= (\mathbf{x}_{u_A(1)})_A (\mathbf{x}_{1v_A(1)})_B^\top + (\mathbf{x}_{v_A(1)})_A (\mathbf{x}_{1u_A(1)})_B^\top - (\mathbf{x}_{v_A(1)})_A (\mathbf{x}_{1v_A(1)})_B^\top - (\mathbf{x}_{u_A(1)})_A (\mathbf{x}_{1u_A(1)})_B^\top \\ &= (\mathbf{x}_{u_A(1)})_A (\mathbf{x}_{v_B(1)})_B^\top + (\mathbf{x}_{v_A(1)})_A (\mathbf{x}_{u_B(1)})_B^\top - (\mathbf{x}_{v_A(1)})_A (\mathbf{x}_{v_B(1)})_B^\top - (\mathbf{x}_{u_A(1)})_A (\mathbf{x}_{u_B(1)})_B^\top \quad (62) \\ &= (\mathbf{x}_{u_A(1)})_A (\mathbf{x}_{u_A(1)})_B^\top + (\mathbf{x}_{v_A(1)})_A (\mathbf{x}_{v_A(1)})_B^\top - (\mathbf{x}_{v_A(1)})_A (\mathbf{x}_{u_A(1)})_B^\top - (\mathbf{x}_{u_A(1)})_A (\mathbf{x}_{v_A(1)})_B^\top \quad (63) \\ &= (\mathbf{x}_{u_A(1)} - \mathbf{x}_{v_A(1)})_A (\mathbf{x}_{u_A(1)} - \mathbf{x}_{v_A(1)})_B^\top . \quad (64) \end{aligned}$$

In eq. (62), we have used eqs (58, 59) and in eq. (63), we have used eqs (60, 61).

Key matrices — The proof of our main helper Theorem is relatively heavy in linear algebra notations: for example, it involves T double applications of Sherman-Morrison's inversion Lemma. We now define a series of matrices and vectors that will be most useful to simplify notations and proofs. Letting $b \doteq 8n\gamma$ (where γ is the parameter of the Ridge regularization in our Taylor loss, see Lemma N below), we first define the matrix we will use most often:

$$\mathbf{v}_t \doteq \left(\hat{X}_t \hat{X}_t^\top + b \cdot \Gamma \right)^{-1} , t = 0, 1, \dots, T , \quad (65)$$

where Γ is the Ridge regularization parameter matrix in our Taylor loss, see Lemma N below. Another matrix U_t , quantifies precisely the local mistake made by each elementary permutation. To define it, we first let (for $t = 1, 2, \dots, T$):

$$\mathbf{a}_t \doteq (\mathbf{x}_{u_A(t)} - \mathbf{x}_{v_A(t)})_A , \quad (66)$$

$$\mathbf{b}_t \doteq (\mathbf{x}_{u_B(t)} - \mathbf{x}_{v_B(t)})_B . \quad (67)$$

Also, let (for $t = 1, 2, \dots, T$)

$$\mathbf{a}_t^+ \doteq \begin{bmatrix} (\mathbf{x}_{u_A(t)} - \mathbf{x}_{v_A(t)})_A \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^d , \quad (68)$$

$$\mathbf{b}_t^+ \doteq \begin{bmatrix} \mathbf{0} \\ (\mathbf{x}_{u_B(t)} - \mathbf{x}_{v_B(t)})_B \end{bmatrix} \in \mathbb{R}^d , \quad (69)$$

and finally (for $t = 1, 2, \dots, T$),

$$c_{0,t} \doteq \mathbf{a}_t^{+\top} \mathbf{v}_{t-1} \mathbf{a}_t^+ , \quad (70)$$

$$c_{1,t} \doteq \mathbf{a}_t^{+\top} \mathbf{v}_{t-1} \mathbf{b}_t^+ , \quad (71)$$

$$c_{2,t} \doteq \mathbf{b}_t^{+\top} \mathbf{v}_{t-1} \mathbf{b}_t^+ . \quad (72)$$

We now define U_t as the following block matrix:

$$U_t \doteq \frac{1}{(1 - c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \left[\begin{array}{c|c} c_{2,t} \cdot \mathbf{a}_t \mathbf{a}_t^\top & (1 - c_{1,t}) \cdot \mathbf{a}_t \mathbf{b}_t^\top \\ \hline (1 - c_{1,t}) \cdot \mathbf{b}_t \mathbf{a}_t^\top & c_{0,t} \cdot \mathbf{b}_t \mathbf{b}_t^\top \end{array} \right] , t = 1, 2, \dots, T . \quad (73)$$

U_t can be computed only when $(1 - c_{1,t})^2 \neq c_{0,t}c_{2,t}$. This shall be the subject of the *invertibility* assumption below. Hereafter, we suppose without loss of generality that $\mathbf{b}_t \neq \mathbf{0}$, since otherwise permutations would make no mistakes on the shuffle part.

There is one important thing to remark on U_t : it is defined from the indices $u_A(t)$ and $v_A(t)$ in A that are affected by P_t . Hence, U_1 collects the two first such indices (see Figure 6). We also define matrix Λ_t as follows:

$$\Lambda_t \doteq 2\mathbf{v}_t U_{t+1} , t = 0, 1, \dots, T-1 . \quad (74)$$

To finishup with matrices, we define a doubly indexed matrices that shall be crucial to our proofs, $H_{i,j}$ for $0 \leq j \leq i \leq T$:

$$H_{i,j} \doteq \begin{cases} \prod_{k=j}^{i-1} (I_d + \Lambda_k) & \text{if } 0 \leq j < i \\ I_d & \text{if } j = i \end{cases} . \quad (75)$$

Key vectors — we let

$$\boldsymbol{\varepsilon}_t \doteq \boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t , t = 0, 1, \dots, T-1 , \quad (76)$$

which is the difference between two successive mean operators, and

$$\boldsymbol{\lambda}_t \doteq 2\mathbf{v}_{t+1} \boldsymbol{\varepsilon}_t , t = 0, 1, \dots, T-1 . \quad (77)$$

Figure 7 summarizes our key notations in this Section. We are now ready to proceed through the proof of our key helper Theorem.

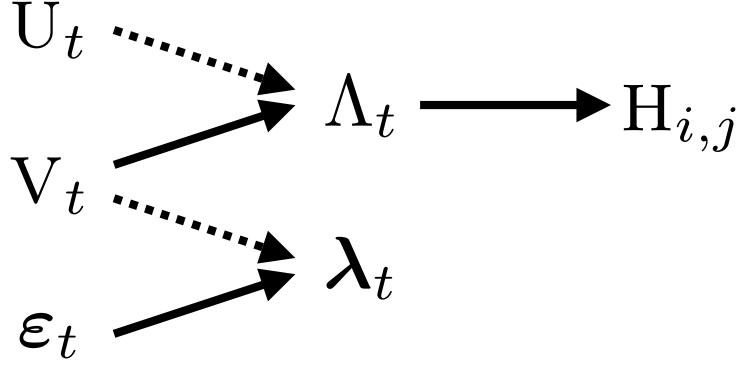


Figure 7: Summary of our key notations on matrices and vectors, and dependencies. The dashed arrow means indexes do not match (eq. (77)).

V.2 Helper Theorem

In this Section, we first show (Theorem P below) that under lightweight assumptions to ensure the existence of v_t , the difference between two successive optimal classifiers in the progressive computation of the overall permutation matrix that generates the errors is *exactly* given by:

$$\begin{aligned} \boldsymbol{\theta}_{t+1}^* - \boldsymbol{\theta}_t^* &= 2 \cdot v_t \mathbf{U}_{t+1} \boldsymbol{\theta}_t^* + 2 \cdot v_{t+1} \boldsymbol{\epsilon}_t \\ &= \Lambda_t \boldsymbol{\theta}_t^* + \boldsymbol{\lambda}_t, \quad \forall t \geq 0, \end{aligned} \quad (78)$$

where $\Lambda_t, \boldsymbol{\epsilon}_t, \boldsymbol{\lambda}_t$ are defined in eqs (76, 74, 77). This holds regardless of the permutation matrices in the sequence.

We start by the trivial solutions to the minimization of the Taylor loss for all $t = 1, 2, \dots, T$.

Lemma N *Let*

$$\ell_{S,\gamma}(\boldsymbol{\theta}) \doteq \log 2 - \frac{1}{n} \sum_i \left\{ \frac{1}{2} \cdot y_i \boldsymbol{\theta}^\top \mathbf{x}_i - \frac{1}{8} \cdot (\boldsymbol{\theta}^\top \mathbf{x}_i)^2 \right\} + \gamma \boldsymbol{\theta}^\top \Gamma \boldsymbol{\theta} \quad (79)$$

denote the γ -Ridge regularized Taylor loss for set S on classifier $\boldsymbol{\theta}$. Then

$$\begin{aligned} \boldsymbol{\theta}^*(\hat{S}) &\doteq \arg \min_{\boldsymbol{\theta}} \ell_{\hat{S},\gamma}(\boldsymbol{\theta}) \\ &= 2 \cdot \left(\hat{\mathbf{X}} \hat{\mathbf{X}}^\top + b \cdot \Gamma \right)^{-1} \boldsymbol{\mu}(\hat{S}), \end{aligned} \quad (80)$$

with $b \doteq 8n\gamma$. More generally, if we let $\boldsymbol{\mu}_t$ denote the mean operator for set S_t , then the optimal classifier for $\ell_{\hat{S}_t,\gamma}$ is given by

$$\boldsymbol{\theta}_t^* \doteq \boldsymbol{\theta}^*(\hat{S}_t) = 2 \cdot v_t \boldsymbol{\mu}_t. \quad (81)$$

(proof straightforward)

Lemma O *Suppose v_{t-1} exists. Then v_t exists iff the following holds true:*

$$\begin{cases} c_{1,t} \neq 1, \\ (1 - c_{1,t})^2 \neq c_{0,t}c_{2,t}. \end{cases} \quad (82)$$

Proof We know that \hat{X}_t is obtained from \hat{X}_{t-1} after permuting the shuffle part of observations at indexes $u_A(t)$ and $v_A(t)$ in $\hat{X}_{(t-1)B}$ by P_t (see Figure 6). So,

$$\begin{aligned}\hat{X}_{tB} &= \hat{X}_{(t-1)B} + \hat{X}_{(t-1)B}(P_t - I_n) \\ &= \hat{X}_{(t-1)B} + \hat{X}_{(t-1)B}(\mathbf{1}_{u_A(t)}\mathbf{1}_{v_A(t)}^\top + \mathbf{1}_{v_A(t)}\mathbf{1}_{u_A(t)}^\top - \mathbf{1}_{v_A(t)}\mathbf{1}_{v_A(t)}^\top - \mathbf{1}_{u_A(t)}\mathbf{1}_{u_A(t)}^\top),\end{aligned}\quad (83)$$

where $\mathbf{1}_u \in \mathbb{R}^n$ is the u^{th} canonical basis vector. We also have

$$\begin{aligned}\hat{X}_t \hat{X}_t^\top &= \left[\begin{array}{c|c} X_A X_A^\top & X_A \hat{X}_{tB}^\top \\ \hline \hat{X}_{tB} X_A^\top & \hat{X}_{tB} \hat{X}_{tB}^\top \end{array} \right] \\ &= \left[\begin{array}{c|c} X_A X_A^\top & X_A \hat{X}_{tB}^\top \\ \hline \hat{X}_{tB} X_A^\top & \hat{X}_{(t-1)B} P_t P_t^\top \hat{X}_{(t-1)B}^\top \end{array} \right] \\ &= \left[\begin{array}{c|c} X_A X_A^\top & X_A \hat{X}_{tB}^\top \\ \hline \hat{X}_{tB} X_A^\top & \hat{X}_{(t-1)B} \hat{X}_{(t-1)B}^\top \end{array} \right],\end{aligned}\quad (84)$$

because the inverse of a permutation matrix is its transpose. We recall that X_A does not change throughout permutations, only X_B does. Hence,

$$\begin{aligned}\hat{X}_t \hat{X}_t^\top &= \hat{X}_{t-1} \hat{X}_{t-1}^\top + \left[\begin{array}{c|c} 0 & X_A (\hat{X}_{tB} - X_{(t-1)B})^\top \\ \hline (\hat{X}_{tB} - X_{(t-1)B}) X_A^\top & 0 \end{array} \right] \\ &= \hat{X}_{t-1} \hat{X}_{t-1}^\top + \left[\begin{array}{c|c} 0 & X_A \Theta_{u_A(t), v_A(t)} \hat{X}_{(t-1)B}^\top \\ \hline \hat{X}_{(t-1)B} \Theta_{u_A(t), v_A(t)} X_A^\top & 0 \end{array} \right],\end{aligned}\quad (85)$$

with $\Theta_{u_A(t), v_A(t)} = \mathbf{1}_{u_A(t)}\mathbf{1}_{v_A(t)}^\top + \mathbf{1}_{v_A(t)}\mathbf{1}_{u_A(t)}^\top - \mathbf{1}_{v_A(t)}\mathbf{1}_{v_A(t)}^\top - \mathbf{1}_{u_A(t)}\mathbf{1}_{u_A(t)}^\top$ (symmetric, see eq. (83) and example M). Now, remark that

$$\begin{aligned}&X_A \Theta_{u_A(t), v_A(t)} \hat{X}_{(t-1)B}^\top \\ &= X_A (\mathbf{1}_{u_A(t)}\mathbf{1}_{v_A(t)}^\top + \mathbf{1}_{v_A(t)}\mathbf{1}_{u_A(t)}^\top - \mathbf{1}_{v_A(t)}\mathbf{1}_{v_A(t)}^\top - \mathbf{1}_{u_A(t)}\mathbf{1}_{u_A(t)}^\top) \hat{X}_{tB}^\top \\ &= (\mathbf{x}_{u_A(t)})_A (\mathbf{x}_{tv_A(t)})_B^\top + (\mathbf{x}_{v_A(t)})_A (\mathbf{x}_{tu_A(t)})_B^\top - (\mathbf{x}_{v_A(t)})_A (\mathbf{x}_{tv_A(t)})_B^\top - (\mathbf{x}_{u_A(t)})_A (\mathbf{x}_{tu_A(t)})_B^\top \\ &= (\mathbf{x}_{u_A(t)})_A (\mathbf{x}_{v_B(t)})_B^\top + (\mathbf{x}_{v_A(t)})_A (\mathbf{x}_{u_B(t)})_B^\top - (\mathbf{x}_{v_A(t)})_A (\mathbf{x}_{v_B(t)})_B^\top - (\mathbf{x}_{u_A(t)})_A (\mathbf{x}_{u_B(t)})_B^\top \quad (86) \\ &= -((\mathbf{x}_{u_A(t)})_A - (\mathbf{x}_{v_A(t)})_A)((\mathbf{x}_{u_B(t)})_B - (\mathbf{x}_{v_B(t)})_B)^\top \\ &= -(\mathbf{x}_{u_A(t)} - \mathbf{x}_{v_A(t)})_A (\mathbf{x}_{u_B(t)} - \mathbf{x}_{v_B(t)})_B^\top = -\mathbf{a}_t \mathbf{b}_t^\top.\end{aligned}\quad (87)$$

Eq. (86) holds because of Lemma L. We finally get

$$\hat{X}_t \hat{X}_t^\top = \hat{X}_{t-1} \hat{X}_{t-1}^\top - \mathbf{a}_t^+ \mathbf{b}_t^{+\top} - \mathbf{b}_t^+ \mathbf{a}_t^{+\top}, \quad (88)$$

and so we have

$$v_t = \left(v_{t-1}^{-1} - \mathbf{a}_t^+ \mathbf{b}_t^{+\top} - \mathbf{b}_t^+ \mathbf{a}_t^{+\top} \right)^{-1}. \quad (89)$$

We analyze when v_t can be computed. First notice that assuming v_{t-1} exists implies its inverse also exists, and so

$$\begin{aligned}\det(v_{t-1}^{-1} - \mathbf{a}_t^+ \mathbf{b}_t^{+\top}) &= \det(v_{t-1}^{-1}) \det(I_d - v_{t-1} \mathbf{a}_t^+ \mathbf{b}_t^{+\top}) \\ &= \det(v_{t-1}^{-1})(1 - \mathbf{b}_t^{+\top} v_{t-1} \mathbf{a}_t^+) \\ &= \det(v_{t-1}^{-1})(1 - c_{1,t}),\end{aligned}\quad (90)$$

where the last identity comes from Sylvester's determinant formula. So, if in addition $1 - c_{1,t} \neq 0$, then

$$\begin{aligned} & \det(v_{t-1}^{-1} - \mathbf{a}_t^+ \mathbf{b}_t^{+\top} - \mathbf{b}_t^+ \mathbf{a}_t^{+\top}) \\ &= \det(v_{t-1}^{-1} - \mathbf{a}_t^+ \mathbf{b}_t^{+\top}) \det(I_d - (v_{t-1}^{-1} - \mathbf{a}_t^+ \mathbf{b}_t^{+\top}) \mathbf{b}_t^+ \mathbf{a}_t^{+\top}) \\ &= \det(v_{t-1}^{-1})(1 - c_{1,t}) \det(I_d - (v_{t-1}^{-1} - \mathbf{a}_t^+ \mathbf{b}_t^{+\top})^{-1} \mathbf{b}_t^+ \mathbf{a}_t^{+\top}) \end{aligned} \quad (91)$$

$$= \det(v_{t-1}^{-1})(1 - c_{1,t}) \left(1 - \mathbf{a}_t^{+\top} (v_{t-1}^{-1} - \mathbf{a}_t^+ \mathbf{b}_t^{+\top})^{-1} \mathbf{b}_t^+ \right) \quad (92)$$

$$= \det(v_{t-1}^{-1})(1 - c_{1,t}) \left(1 - \mathbf{a}_t^{+\top} \left(v_{t-1} + \frac{1}{1 - \mathbf{b}_t^{+\top} v_{t-1} \mathbf{a}_t^+} \cdot v_{t-1} \mathbf{a}_t^+ \mathbf{b}_t^{+\top} v_{t-1} \right) \mathbf{b}_t^+ \right) \quad (93)$$

$$\begin{aligned} &= \det(v_{t-1}^{-1})(1 - c_{1,t}) \left(1 - c_{1,t} - \frac{c_{0,t} c_{2,t}}{1 - c_{1,t}} \right) \\ &= \frac{(1 - c_{1,t})^2 - c_{0,t} c_{2,t}}{\det(v_{t-1})}. \end{aligned} \quad (94)$$

Here, eq. (91) comes from eq. (90). Eq. (92) is another application of Sylvester's determinant formula. Eq. (92) is Sherman-Morrison formula. We immediately conclude on Lemma O. ■

If we now assume without loss of generality that v_0 exists — which boils down to taking $\gamma > 0, \Gamma \succ 0$ —, then we get the existence of the complete sequence of matrices v_t (and thus the existence of the sequence of optimal classifiers $\theta_0^*, \theta_1^*, \dots$) provided the following **invertibility** condition is satisfied.

(invertibility) For any $t \geq 1$, $(1 - c_{1,t})^2 \notin \{0, c_{0,t} c_{2,t}\}$.

Theorem P Suppose the invertibility assumption holds. Then we have:

$$\frac{1}{2} \cdot (\theta_{t+1}^* - \theta_t^*) = v_t u_{t+1} \theta_t^* + v_{t+1} \epsilon_t, \forall t \geq 0,$$

where ϵ_t is defined in eq. (76).

Proof We have from Lemma N, for any $t \geq 1$,

$$\begin{aligned} \frac{1}{2} \cdot (\theta_t^* - \theta_{t-1}^*) &= v_t \mu_t - v_{t-1} \mu_{t-1} \\ &= \Delta_{t-1} \mu_{t-1} + v_t \epsilon_{t-1}, \end{aligned} \quad (95)$$

with $\Delta_t = v_{t+1} - v_t$. It comes from eq. (88),

$$\Delta_{t-1} = \left(\hat{X}_{t-1} \hat{X}_{t-1}^\top + b \cdot \Gamma - \mathbf{a}_t^+ \mathbf{b}_t^{+\top} - \mathbf{b}_t^+ \mathbf{a}_t^{+\top} \right)^{-1} - v_t. \quad (96)$$

To simplify this expression, we need two consecutive applications of Sherman-Morrison's inversion formula:

$$\begin{aligned} & \left(\hat{\mathbf{X}}_{t-1} \hat{\mathbf{X}}_{t-1}^\top + b \cdot \Gamma - \mathbf{a}_t^+ \mathbf{b}_t^{+\top} - \mathbf{b}_t^+ \mathbf{a}_t^{+\top} \right)^{-1} \\ = & \left(\hat{\mathbf{X}}_{t-1} \hat{\mathbf{X}}_{t-1}^\top + b \cdot \Gamma - \mathbf{a}_t^+ \mathbf{b}_t^{+\top} \right)^{-1} + \frac{1}{1 - \mathbf{a}_t^{+\top} \left(\hat{\mathbf{X}}_{t-1} \hat{\mathbf{X}}_{t-1}^\top + b \cdot \Gamma - \mathbf{a}_t^+ \mathbf{b}_t^{+\top} \right)^{-1} \mathbf{b}_t^+} \cdot Q_t(97) \end{aligned}$$

with

$$Q_t \doteq \left(\hat{\mathbf{X}}_{t-1} \hat{\mathbf{X}}_{t-1}^\top + b \cdot \Gamma - \mathbf{a}_t^+ \mathbf{b}_t^{+\top} \right)^{-1} \mathbf{b}_t^+ \mathbf{a}_t^{+\top} \left(\hat{\mathbf{X}}_{t-1} \hat{\mathbf{X}}_{t-1}^\top + b \cdot \Gamma - \mathbf{a}_t^+ \mathbf{b}_t^{+\top} \right)^{-1}, \quad (98)$$

and

$$\left(\hat{\mathbf{X}}_{t-1} \hat{\mathbf{X}}_{t-1}^\top + b \cdot \Gamma - \mathbf{a}_t^+ \mathbf{b}_t^{+\top} \right)^{-1} = v_{t-1} + \frac{1}{1 - \mathbf{b}_t^{+\top} v_{t-1} \mathbf{a}_t^+} \cdot v_{t-1} \mathbf{a}_t^+ \mathbf{b}_t^{+\top} v_{t-1}. \quad (99)$$

Let us define the following shorthand:

$$\Sigma_t \doteq v_{t-1} + \frac{1}{1 - \mathbf{b}_t^{+\top} v_{t-1} \mathbf{a}_t^+} \cdot v_{t-1} \mathbf{a}_t^+ \mathbf{b}_t^{+\top} v_{t-1}. \quad (100)$$

Then, plugging together eqs. (97) and (99), we get:

$$\begin{aligned}
& \left(\hat{\mathbf{X}}_{t-1} \hat{\mathbf{X}}_{t-1}^\top + b \cdot \Gamma - \mathbf{a}_t^+ \mathbf{b}_t^{+\top} - \mathbf{b}_t^+ \mathbf{a}_t^{+\top} \right)^{-1} \\
&= \mathbf{v}_{t-1} + \frac{1}{1 - \mathbf{b}_t^{+\top} \mathbf{v}_{t-1} \mathbf{a}_t^+} \cdot \mathbf{v}_{t-1} \mathbf{a}_t^+ \mathbf{b}_t^{+\top} \mathbf{v}_{t-1} \\
&\quad + \frac{1}{1 - \mathbf{a}_t^{+\top} \mathbf{v}_{t-1} \mathbf{b}_t^+ - \frac{\mathbf{a}_t^{+\top} \mathbf{v}_{t-1} \mathbf{a}_t^+ \cdot \mathbf{b}_t^{+\top} \mathbf{v}_{t-1} \mathbf{b}_t^+}{1 - \mathbf{b}_t^{+\top} \mathbf{v}_{t-1} \mathbf{a}_t^+}} \cdot \Sigma_t \mathbf{b}_t^+ \mathbf{a}_t^{+\top} \Sigma_t \\
&= \mathbf{v}_{t-1} + \frac{1}{1 - c_{1,t}} \cdot \mathbf{v}_{t-1} \mathbf{a}_t^+ \mathbf{b}_t^{+\top} \mathbf{v}_{t-1} \\
&\quad + \frac{1}{1 - c_{1,t} - \frac{c_{0,t}c_{2,t}}{1-c_{1,t}}} \cdot \left(\begin{array}{c} \mathbf{v}_{t-1} \\ + \\ \frac{1}{1-c_{1,t}} \cdot \mathbf{v}_{t-1} \mathbf{a}_t^+ \mathbf{b}_t^{+\top} \mathbf{v}_{t-1} \end{array} \right) \mathbf{b}_t^+ \mathbf{a}_t^{+\top} \left(\begin{array}{c} \mathbf{v}_{t-1} \\ + \\ \frac{1}{1-c_{1,t}} \cdot \mathbf{v}_{t-1} \mathbf{a}_t^+ \mathbf{b}_t^{+\top} \mathbf{v}_{t-1} \end{array} \right) \\
&= \mathbf{v}_{t-1} + \frac{1}{1 - c_{1,t}} \cdot \mathbf{v}_{t-1} \mathbf{a}_t^+ \mathbf{b}_t^{+\top} \mathbf{v}_{t-1} + \frac{1}{1 - c_{1,t} - \frac{c_{0,t}c_{2,t}}{1-c_{1,t}}} \cdot \mathbf{v}_{t-1} \mathbf{b}_t^+ \mathbf{a}_t^{+\top} \mathbf{v}_{t-1} \\
&\quad + \frac{c_{0,t}}{(1 - c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \mathbf{v}_{t-1} \mathbf{b}_t^+ \mathbf{b}_t^{+\top} \mathbf{v}_{t-1} + \frac{c_{2,t}}{(1 - c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \mathbf{v}_{t-1} \mathbf{a}_t^+ \mathbf{a}_t^{+\top} \mathbf{v}_{t-1} \\
&\quad + \frac{c_{0,t}c_{2,t}}{(1 - c_{1,t})((1 - c_{1,t})^2 - c_{0,t}c_{2,t})} \cdot \mathbf{v}_{t-1} \mathbf{a}_t^+ \mathbf{b}_t^{+\top} \mathbf{v}_{t-1} \\
&= \mathbf{v}_{t-1} + \frac{1 - c_{1,t}}{(1 - c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \left(\mathbf{v}_{t-1} \mathbf{a}_t^+ \mathbf{b}_t^{+\top} \mathbf{v}_{t-1} + \mathbf{v}_{t-1} \mathbf{b}_t^+ \mathbf{a}_t^{+\top} \mathbf{v}_{t-1} \right) \\
&\quad + \frac{c_{0,t}}{(1 - c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \mathbf{v}_{t-1} \mathbf{b}_t^+ \mathbf{b}_t^{+\top} \mathbf{v}_{t-1} + \frac{c_{2,t}}{(1 - c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \mathbf{v}_{t-1} \mathbf{a}_t^+ \mathbf{a}_t^{+\top} \mathbf{v}_{t-1} \\
&= \mathbf{v}_{t-1} + \frac{1}{(1 - c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \left\{ \begin{array}{l} (1 - c_{1,t}) \cdot (\mathbf{v}_{t-1} \mathbf{a}_t^+ \mathbf{b}_t^{+\top} \mathbf{v}_{t-1} + \mathbf{v}_{t-1} \mathbf{b}_t^+ \mathbf{a}_t^{+\top} \mathbf{v}_{t-1}) \\ + c_{0,t} \cdot \mathbf{v}_{t-1} \mathbf{b}_t^+ \mathbf{b}_t^{+\top} \mathbf{v}_{t-1} \\ + c_{2,t} \cdot \mathbf{v}_{t-1} \mathbf{a}_t^+ \mathbf{a}_t^{+\top} \mathbf{v}_{t-1} \end{array} \right\} \\
&= \mathbf{v}_{t-1} + \mathbf{v}_{t-1} \mathbf{U}_t \mathbf{v}_{t-1} .
\end{aligned} \tag{101}$$

So,

$$\begin{aligned}
\frac{1}{2} \cdot (\boldsymbol{\theta}_t^* - \boldsymbol{\theta}_{t-1}^*) &= \Delta_{t-1} \boldsymbol{\mu}_{t-1} + \mathbf{v}_t \boldsymbol{\varepsilon}_{t-1} \\
&= \mathbf{v}_{t-1} \mathbf{U}_t \mathbf{v}_{t-1} \boldsymbol{\mu}_{t-1} + \mathbf{v}_t \boldsymbol{\varepsilon}_{t-1} \\
&= \mathbf{v}_{t-1} \mathbf{U}_t \boldsymbol{\theta}_{t-1}^* + \mathbf{v}_t \boldsymbol{\varepsilon}_{t-1} ,
\end{aligned} \tag{102}$$

as claimed (end of the proof of Lemma P). ■

All that remains to do now is to unravel the relationship in Theorem P and quantify the exact variation $\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*$ as a function of $\boldsymbol{\theta}_0^*$ (which is the error-free optimal classifier), holding for any permutation P_* . We therefore suppose that the invertibility assumption holds.

Theorem Q Suppose the invertibility assumption holds. For any $T \geq 1$,

$$\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^* = (\mathbf{H}_{T,0} - \mathbf{I}_d) \boldsymbol{\theta}_0^* + \sum_{t=0}^{T-1} \mathbf{H}_{T,t+1} \boldsymbol{\lambda}_t . \tag{103}$$

Proof We recall first that we have from Theorem P, $\boldsymbol{\theta}_{t+1}^* - \boldsymbol{\theta}_t^* = \Lambda_t \boldsymbol{\theta}_t^* + \boldsymbol{\lambda}_t, \forall t \geq 0$. Equivalently,

$$\boldsymbol{\theta}_{t+1}^* = (\mathbf{I}_d + \Lambda_t) \boldsymbol{\theta}_t^* + \boldsymbol{\lambda}_t . \quad (104)$$

Unravelling, we easily get $\forall T \geq 1$,

$$\begin{aligned} \boldsymbol{\theta}_T^* &= \prod_{t=0}^{T-1} (\mathbf{I}_d + \Lambda_t) \boldsymbol{\theta}_0^* + \boldsymbol{\lambda}_{T-1} + \sum_{j=0}^{T-2} \prod_{t=j+1}^{T-1} (\mathbf{I}_d + \Lambda_t) \boldsymbol{\lambda}_j \\ &= \mathbf{H}_{T,0} \boldsymbol{\theta}_0^* + \sum_{t=0}^{T-1} \mathbf{H}_{T,t+1} \boldsymbol{\lambda}_t , \end{aligned} \quad (105)$$

which yields the statement of Theorem Q. ■

Since it applies to every permutation matrix, it applies to every entity resolution algorithm. Theorem Q gives us an interesting expression for the deviation $\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*$ which can be used to derive bounds on the distance between the two classifiers, even outside our privacy framework. We apply it below to derive one such bound.

V.3 Assumptions: details and discussion

Our result relies on several assumptions that concern the permutation P_* and its decomposition in eq. (55), as well as on the data size n and regularization parameters γ, Γ . We make in this Section more extensive comments on the assumptions we use.

((ε, τ)-accuracy) Equivalently, P_t is (ε, τ) -accurate iff both conditions below are satisfied:

1. the stretch in the *shuffle* space of errors on an observation due to permutations is bounded by the *total* stretch of the observation:

$$\varpi((\hat{\mathbf{x}}_{ti} - \mathbf{x}_i)_B, \mathbf{w}_B) \leq \varepsilon \cdot \varpi(\mathbf{x}_i, \mathbf{w}) + \tau, \forall i \in [n], \forall \mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 = 1 \quad (106)$$

2. recall that $\mathbf{x}_{u_A(t)}, \mathbf{x}_{v_A(t)}$ are the observations in X whose shuffle parts are *affected* by P_t , and $\mathbf{x}_{u_B(t)}, \mathbf{x}_{v_B(t)}$ are the observations in X whose shuffle parts are *permuted* by P_t . Then the stretch of the errors due to P_t , $(\mathbf{x}_{u_A(t)} - \mathbf{x}_{v_A(t)})_A$ and $(\mathbf{x}_{u_B(t)} - \mathbf{x}_{v_B(t)})_B$, is bounded by the *maximal* stretch of the related observations:

$$\begin{aligned} \varpi((\mathbf{x}_{u_F(t)} - \mathbf{x}_{v_F(t)})_F, \mathbf{w}_F) &\leq \varepsilon \cdot \max_{i \in \{u_F(t), v_F(t)\}} \varpi(\mathbf{x}_i, \mathbf{w}) + \tau, \\ \forall F \in \{A, B\}, \forall \mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 = 1 . \end{aligned} \quad (107)$$

→ Remarks on Definition 1:

Remark 1: eq. (18) in the (ε, τ) -accuracy assumption imposes that the stretch of all errors in \hat{X}_t , $\hat{\mathbf{x}}_{ti} - \mathbf{x}_i$, along direction \mathbf{w} , be bounded by the stretch of the corresponding observations in X along the same direction. Since the anchor parts of $\hat{\mathbf{x}}_{ti}$ and \mathbf{x}_i coincide by convention, the stretch of the error is in fact measured on the shuffle set of features. Figure 8 gives a visual for that.

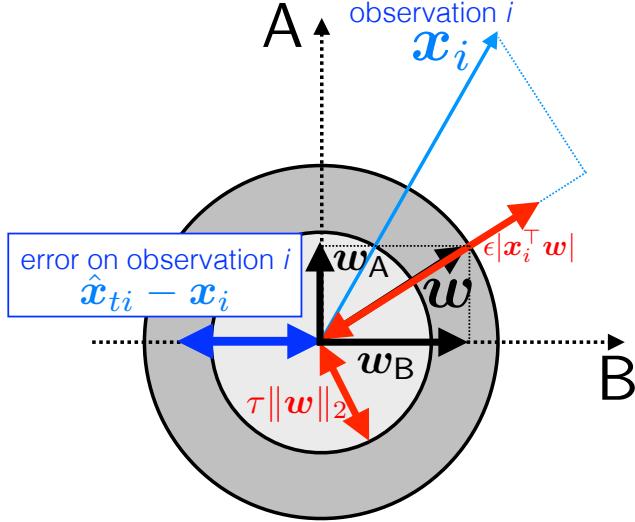


Figure 8: If P_t is (ε, τ) -accurate, eq. (18) in Definition 1 reads as follows: for any vector w , the norm of the error projected on w , *in the shuffle space*, is no more than a fraction of the norm of w (red) plus a fraction of the norm of the projection of the observation along w *in the complete feature space*.

Remark 2: Figure 9 illustrates the second part of the (ε, τ) -accuracy assumption. Informally, stretch of errors due to P_t is bounded by the maximal stretch of the related observations affected by P_t .

Remark 3: parameter τ is necessary in some way: if one picks w orthogonal to x_i , then $\varpi(x_i, w) = \|x_i\|_2 |\cos(x_i, w)| = 0$ while it may be the case that $\varpi((\hat{x}_{ti} - x_i)_B, w_B) = \|(\hat{x}_{ti} - x_i)_B\|_2 |\cos((\hat{x}_{ti} - x_i)_B, w_B)| \neq 0$.

→ Remarks on Definition 4:

Remark 1: both conditions are in fact weak. They barely impose that the strength of regularization ($\gamma \lambda_1^\uparrow(\Gamma)$) is proportional to a squared norm, while data size is no less than a potentially small constant.

Remark 2: X_*^2 can be of the same order as $\inf_w \sigma^2(\{\varpi(x_i, w)\}_{i=1}^n)$: indeed, if all observations have the same norm in \mathbb{R}^d , then $\sigma^2(\{\varpi(x_i, w)\}_{i=1}^n) = \beta(w) X_*^2$ for some $0 \leq \beta \leq 1$. So the problem can be reduced to roughly having

$$\gamma \lambda_1^\uparrow(\Gamma) \geq \left(1 - \frac{(1-\varepsilon)^2}{8} \cdot \beta(w)\right) \cdot X_*^2, \quad \forall w \in \mathbb{R}^d, \quad (108)$$

which also shows that $\gamma \lambda_1^\uparrow(\Gamma)$ has to be homogeneous to a square norm (*Cf* Remark 1).

V.4 Proof of Theorem 6

First, we shall see (Corollary W below) that the assumptions made guarantee the invertibility condition in Theorem Q. From Theorem Q, we now investigate a general bound of the type

$$\|\theta_T^* - \theta_0^*\|_2 \leq a \cdot \|\theta_0^*\|_2 + b, \quad \forall T \geq 1. \quad (109)$$

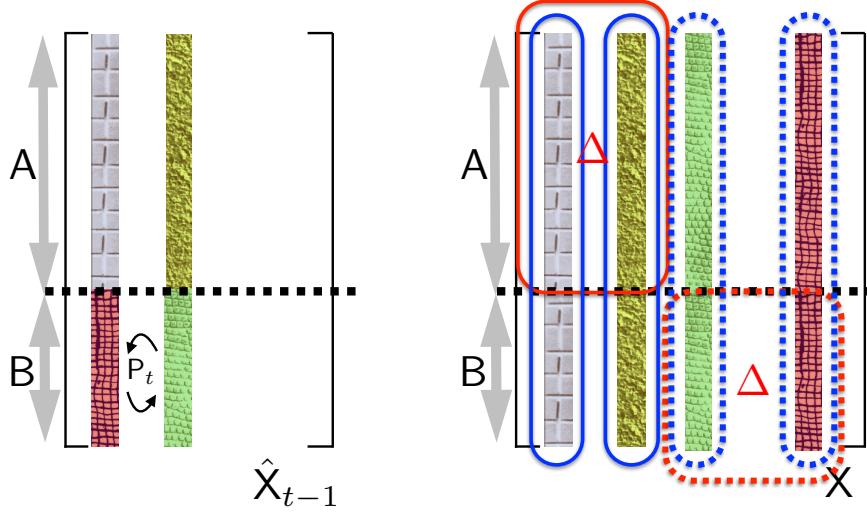


Figure 9: Overview of eq. (19) for the (ε, τ) -accuracy assumption for elementary permutation P_t . *Left:* effect of P_t on \hat{X}_{t-1} ; *right:* corresponding vectors whose stretch is used in the assumption. Informally, the assumption says that the stretch of the (dashed, resp. plain) red vectors is bounded by the maximal stretch among the (dashed, resp. plain) blue vectors (best viewed in color).

where $a \geq 0$ and $b \geq 0$ are two reals that we want as small as possible. We first need an intermediate technical Lemma. Let $\mu(\{a_i\}) \doteq (1/n) \cdot \sum_i a_i$ denote for short the average of $\{a_i\}$ with $a_i \geq 0, \forall i$. Let $\gamma' \geq 0$ be *any* real such that:

$$\frac{\mu^2(\{a_i\})}{\mu(\{a_i^2\})} \leq (1 - \gamma') . \quad (110)$$

Remark that the result is true for $\gamma' = 0$ since $\mu(\{a_i^2\}) - \mu^2(\{a_i\})$ is just the variance of $\{a_i\}$, which is non-negative. Remark also that we must have $\gamma' \leq 1$.

Lemma R $\sum_i ((1 - \varepsilon)a_i - q)^2 \geq \gamma'(1 - \varepsilon)^2 \sum_i a_i^2, \forall \varepsilon \leq 1, q \in \mathbb{R}$.

Proof Remark that

$$\sqrt{(1 - \gamma')\mu(\{a_i^2\})} = \inf_{k \geq 0} \frac{1}{2} \cdot \left(k + \frac{1}{k} \cdot (1 - \gamma')\mu(\{a_i^2\}) \right) , \quad (111)$$

so we have:

$$\mu(\{a_i\}) \leq \sqrt{(1 - \gamma')\mu(\{a_i^2\})} \quad (112)$$

$$\leq \frac{q}{2(1 - \varepsilon)} + \frac{(1 - \gamma')(1 - \varepsilon)}{2q} \cdot \mu(\{a_i^2\}) . \quad (113)$$

Ineq. (112) holds because of ineq. (110) and ineq. (112) holds because of eq. (111) and substituting $k \doteq q/(1 - \varepsilon) \geq 0$. After reorganising, we obtain:

$$nq^2 - 2(1 - \varepsilon)q \sum_i a_i + (1 - \gamma')(1 - \varepsilon)^2 \sum_i a_i^2 \geq 0 , \quad (114)$$

and so we obtain the inequality of:

$$\begin{aligned} \sum_i ((1 - \varepsilon)a_i - q)^2 &= nq^2 - 2(1 - \varepsilon)q \sum_i a_i + (1 - \varepsilon)^2 \sum_i a_i^2 \\ &\geq \gamma'(1 - \varepsilon)^2 \sum_i a_i^2 , \end{aligned} \quad (115)$$

which allows to conclude the proof of Lemma R. ■

Let

$$\gamma'(\mathbf{X}, \mathbf{w}) \doteq 1 - \frac{\mu^2(\{\varpi(\mathbf{x}_i, \mathbf{w})\}_{i=1}^n)}{\mu(\{\varpi^2(\mathbf{x}_i, \mathbf{w})\}_{i=1}^n)} , \quad (116)$$

where we recall that $\mu(\{a_i\})$ is the average in set $\{a_i\}$. It is easy to remark that $\gamma'(\mathbf{X}, \mathbf{w}) \in [0, 1]$ and it can be used in Lemma R for the choice

$$\{a_i\} \doteq \{\varpi(\mathbf{x}_i, \mathbf{w})\}_{i=1}^n . \quad (117)$$

It is also not hard to see that as long as there exists two \mathbf{x}_i in \mathbf{X} with a different *direction*, we shall have $\gamma'(\mathbf{X}, \mathbf{w}) > 0, \forall \mathbf{w}$.

Following Bhatia [1997], for any symmetric matrix \mathbf{M} , we let $\lambda_{\mathbf{M}}^{\downarrow}$ (resp. $\lambda_{\mathbf{M}}^{\uparrow}$) denote the vector of eigenvalues arranged in decreasing (resp. increasing) order. So, $\lambda_1^{\downarrow}(\mathbf{M})$ (resp. $\lambda_1^{\uparrow}(\mathbf{M})$) denotes the maximal (resp. minimal) eigenvalue of \mathbf{M} .

Lemma S *For any set $\mathcal{S} \doteq \{a_i\}_{i=1}^n$, let $\mu(\mathcal{S})$ and $\sigma(\mathcal{S})$ denote the mean and standard deviation of \mathcal{S} . If \mathbf{P}_t is (ε, τ) -accurate, then the eigenspectrum of \mathbf{v}_t is bounded as below:*

$$\lambda_1^{\downarrow}(\mathbf{v}_t) \leq \frac{1}{n} \cdot \frac{1}{(1 - \varepsilon)^2 \cdot \inf_{\mathbf{w}} \sigma^2(\{\varpi(\mathbf{x}_i, \mathbf{w})\}_{i=1}^n) + 8\gamma\lambda_1^{\uparrow}(\Gamma)} , \quad (118)$$

$$\lambda_1^{\uparrow}(\mathbf{v}_t) \geq \frac{1}{2n} \cdot \frac{1}{(1 + \varepsilon)^2 \cdot \sup_{\mathbf{w}} \mu(\{\varpi^2(\mathbf{x}_i, \mathbf{w})\}_{i=1}^n) + \tau^2 + 4\gamma\lambda_1^{\downarrow}(\Gamma)} . \quad (119)$$

Proof We first show the upperbound on $\lambda_1^{\downarrow}(\mathbf{v}_t)$, and we start by showing that

$$\lambda_1^{\downarrow}(\mathbf{v}_t) \leq \frac{1}{n} \cdot \frac{1}{(1 - \varepsilon)^2 \cdot \inf_{\mathbf{w}} \gamma'(\mathbf{X}, \mathbf{w})\varsigma(\mathbf{X}, \mathbf{w}) + 8\gamma\lambda_1^{\uparrow}(\Gamma)} , \quad (120)$$

with

$$\varsigma(\mathbf{X}, \mathbf{w}) \doteq \frac{1}{n} \cdot \sum_i \varpi^2(\mathbf{x}_i, \mathbf{w}) . \quad (121)$$

If \mathbf{P}_t is (ε, τ) -accurate, it comes from the triangle inequality

$$\begin{aligned} |\hat{\mathbf{x}}_{ti}^{\top} \mathbf{w}| &= |\mathbf{x}_i^{\top} \mathbf{w} + (\mathbf{x}_{ti_B} - \mathbf{x}_{i_B})^{\top} \mathbf{w}_B| \\ &\geq |\mathbf{x}_i^{\top} \mathbf{w}| - |(\mathbf{x}_{ti_B} - \mathbf{x}_{i_B})^{\top} \mathbf{w}_B| \\ &\geq (1 - \varepsilon)|\mathbf{x}_i^{\top} \mathbf{w}| - \tau \|\mathbf{w}\|_2 , \end{aligned} \quad (122)$$

and so, using Lemma R with $q \doteq \tau$ and $a_i \doteq \|\mathbf{x}_i\|_2 |\cos(\mathbf{x}_i, \mathbf{w})|$, we obtain the last inequality of:

$$\begin{aligned}
\|\hat{\mathbf{X}}_t^\top \mathbf{w}\|_2^2 &= \sum_i (\hat{\mathbf{x}}_i^\top \mathbf{w})^2 \\
&\geq \sum_i ((1 - \varepsilon)|\mathbf{x}_i^\top \mathbf{w}| - \tau \|\mathbf{w}\|_2)^2 \\
&= \|\mathbf{w}\|_2^2 \cdot \sum_i ((1 - \varepsilon)\varpi(\mathbf{x}_i, \mathbf{w}) - \tau)^2 \\
&\geq \|\mathbf{w}\|_2^2 (1 - \varepsilon)^2 \cdot \gamma'(\mathbf{X}, \mathbf{w}) \sum_i \varpi^2(\mathbf{x}_i, \mathbf{w}) .
\end{aligned} \tag{123}$$

Therefore, if P_t is (ε, τ) -accurate, we have

$$\begin{aligned}
\lambda_1^\downarrow(\mathbf{v}_t) &\doteq \left(\inf_{\mathbf{w}} \frac{\mathbf{w}^\top (\hat{\mathbf{X}}_t \hat{\mathbf{X}}_t^\top + b \cdot \Gamma) \mathbf{w}}{\|\mathbf{w}\|_2^2} \right)^{-1} \\
&\leq \frac{1}{((1 - \varepsilon)^2 \cdot \inf_{\mathbf{w}} \gamma'(\mathbf{X}, \mathbf{w}) \sum_i \varpi^2(\mathbf{x}_i, \mathbf{w}) + b \lambda_1^\uparrow(\Gamma))} \\
&= \frac{1}{n} \cdot \frac{1}{(1 - \varepsilon)^2 \inf_{\mathbf{w}} \sigma^2(\mathbf{X}, \mathbf{w}) + 8\gamma \lambda_1^\uparrow(\Gamma)} ,
\end{aligned} \tag{124}$$

where the last identity follows from

$$\begin{aligned}
\inf_{\mathbf{w}} \sigma^2(\mathbf{X}, \mathbf{w}) &\doteq \inf_{\mathbf{w}} \mu(\{\varpi^2(\mathbf{x}_i, \mathbf{w})\}) - \mu^2(\{\varpi(\mathbf{x}_i, \mathbf{w})\}) \\
&= \inf_{\mathbf{w}} \left(1 - \frac{\mu^2(\{\varpi(\mathbf{x}_i, \mathbf{w})\})}{\mu(\{\varpi^2(\mathbf{x}_i, \mathbf{w})\})} \right) \cdot \mu(\{\varpi^2(\mathbf{x}_i, \mathbf{w})\}) \\
&= \inf_{\mathbf{w}} \left(1 - \frac{\mu^2(\{\varpi(\mathbf{x}_i, \mathbf{w})\})}{\mu(\{\varpi^2(\mathbf{x}_i, \mathbf{w})\})} \right) \cdot \frac{1}{n} \cdot \sum_i \varpi^2(\mathbf{x}_i, \mathbf{w}) \\
&= \frac{1}{n} \cdot \inf_{\mathbf{w}} \gamma'(\mathbf{X}, \mathbf{w}) \sum_i \varpi^2(\mathbf{x}_i, \mathbf{w}) .
\end{aligned} \tag{125}$$

This finishes the proof for ineq. (118). To show ineq. (119), we remark that if P_t is (ε, τ) -accurate, it also comes from the triangle inequality

$$\begin{aligned}
|\hat{\mathbf{x}}_{ti}^\top \mathbf{w}| &= |\mathbf{x}_i^\top \mathbf{w} + (\mathbf{x}_{ti_B} - \mathbf{x}_{i_B})^\top \mathbf{w}_B| \\
&\leq |\mathbf{x}_i^\top \mathbf{w}| + |(\mathbf{x}_{ti_B} - \mathbf{x}_{i_B})^\top \mathbf{w}_B| \\
&\leq (1 + \varepsilon) |\mathbf{x}_i^\top \mathbf{w}| + \tau \|\mathbf{w}\|_2 ,
\end{aligned} \tag{126}$$

and so,

$$\begin{aligned}
\|\hat{\mathbf{X}}_t^\top \mathbf{w}\|_2^2 &= \sum_i (\hat{\mathbf{x}}_i^\top \mathbf{w})^2 \\
&\leq \|\mathbf{w}\|_2^2 \cdot \sum_i ((1 + \varepsilon)\varpi(\mathbf{x}_i, \mathbf{w}) + \tau)^2 \\
&\leq \|\mathbf{w}\|_2^2 \cdot \left(2(1 + \varepsilon)^2 \sum_i \varpi^2(\mathbf{x}_i, \mathbf{w}) + 2n\tau^2 \right) ,
\end{aligned} \tag{127}$$

because $(a + b)^2 \leq 2a^2 + 2b^2$. Therefore, if P_t is (ε, τ) -accurate, we have

$$\begin{aligned}\lambda_1^\uparrow(\mathbf{v}_t) &\doteq \left(\sup_{\mathbf{w}} \frac{\mathbf{w}^\top (\hat{\mathbf{X}}_t \hat{\mathbf{X}}_t^\top + b \cdot \Gamma) \mathbf{w}}{\|\mathbf{w}\|_2^2} \right)^{-1} \\ &\geq \frac{1}{2(1+\varepsilon)^2 \cdot \inf_{\mathbf{w}} \sum_i \varpi^2(\mathbf{x}_i, \mathbf{w}) + 2n\tau^2 + b\lambda_1^\uparrow(\Gamma)} \\ &= \frac{1}{2n} \cdot \frac{1}{(1+\varepsilon)^2 \sup_{\mathbf{w}} \mu(\{\varpi^2(\mathbf{x}_i, \mathbf{w})\}_{i=1}^n) + \tau^2 + 4\gamma\lambda_1^\uparrow(\Gamma)},\end{aligned}\quad (128)$$

This ends the proof of Lemma S. ■

Lemma T Suppose $(1 - c_{1,t})^2 - c_{0,t}c_{2,t} \neq 0$ ⁵ and $\mathbf{a}_t \neq \mathbf{0}$. Then \mathbf{U}_t is negative semi-definite iff $(1 - c_{1,t})^2 - c_{0,t}c_{2,t} < 0$. Otherwise, \mathbf{U}_t is indefinite. In all cases, for any $z \in \{\lambda_1^\downarrow(\mathbf{U}_t), |\lambda_1^\uparrow(\mathbf{U}_t)|\}$, we have

$$z \leq \frac{2 + 3(c_{0,t} + c_{2,t})}{2|(1 - c_{1,t})^2 - c_{0,t}c_{2,t}|} \cdot \max\{\|\mathbf{a}_t\|_2^2, \|\mathbf{b}_t\|_2^2\}. \quad (129)$$

Proof Consider a block-vector following the column-block partition of \mathbf{U}_t ,

$$\tilde{\mathbf{x}} \doteq \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}. \quad (130)$$

Denote for short $\rho \doteq (1 - c_{1,t})^2 - c_{0,t}c_{2,t}$. We have

$$\mathbf{U}_t \tilde{\mathbf{x}} = \frac{1}{\rho} \cdot \begin{bmatrix} (c_{2,t}(\mathbf{a}_t^\top \mathbf{x}) + (1 - c_{1,t})(\mathbf{b}_t^\top \mathbf{y})) \cdot \mathbf{a}_t \\ ((1 - c_{1,t})(\mathbf{a}_t^\top \mathbf{x}) + c_{0,t}(\mathbf{b}_t^\top \mathbf{y})) \cdot \mathbf{b}_t \end{bmatrix}. \quad (131)$$

We see that the only possibility for $\tilde{\mathbf{x}}$ to be an eigenvector is that $\mathbf{x} \propto \mathbf{a}_t$ and $\mathbf{y} \propto \mathbf{b}_t$ (including the null vector for at most one vector). We now distinguish two cases.

Case 1. $c_{1,t} = 1$. In this case, \mathbf{U}_t is block diagonal and so we get two eigenvectors:

$$\begin{aligned}\mathbf{U}_t \begin{bmatrix} \mathbf{a}_t \\ \mathbf{0} \end{bmatrix} &= -\frac{1}{c_{0,t}c_{2,t}} \cdot \begin{bmatrix} c_{2,t} \cdot \mathbf{a}_t \mathbf{a}_t^\top & 0 \\ 0 & c_{0,t} \cdot \mathbf{b}_t \mathbf{b}_t^\top \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{0} \end{bmatrix} \\ &= -\frac{1}{\lambda(\mathbf{a}_t^+)} \cdot \begin{bmatrix} \mathbf{a}_t \\ \mathbf{0} \end{bmatrix},\end{aligned}\quad (132)$$

with (since $\|\mathbf{a}_t^+\|_2^2 = \|\mathbf{a}_t\|_2^2$):

$$\lambda(\mathbf{a}_t^+) \doteq \frac{\mathbf{a}_t^{+\top} \mathbf{v}_{t-1} \mathbf{a}_t^+}{\|\mathbf{a}_t^+\|_2^2}, \quad (133)$$

⁵This is implied by the invertibility assumption.

and

$$U_t \begin{bmatrix} \mathbf{0} \\ \mathbf{b}_t \end{bmatrix} = -\frac{1}{\lambda(\mathbf{b}_t^+)} \cdot \begin{bmatrix} \mathbf{0} \\ \mathbf{b}_t \end{bmatrix}, \lambda(\mathbf{b}_t^+) \doteq \frac{\mathbf{b}_t^{+\top} \mathbf{v}_{t-1} \mathbf{b}_t^+}{\|\mathbf{b}_t^+\|_2^2}. \quad (134)$$

We also remark that U_t is negative semi-definite.

Case 2. $c_{1,t} \neq 1$. In this case, let us assume without loss of generality that for some $\alpha \in \mathbb{R}_*$,

$$\begin{aligned} \mathbf{x} &= \alpha \cdot \mathbf{a}_t, \\ \mathbf{y} &= \mathbf{b}_t. \end{aligned}$$

In this case, we obtain

$$\begin{aligned} U_t \tilde{\mathbf{x}} &= \frac{(1 - c_{1,t})(\mathbf{a}_t^\top \mathbf{x}) + c_{0,t}(\mathbf{b}_t^\top \mathbf{y})}{(1 - c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \begin{bmatrix} \frac{c_{2,t}(\mathbf{a}_t^\top \mathbf{x}) + (1 - c_{1,t})(\mathbf{b}_t^\top \mathbf{y})}{(1 - c_{1,t})(\mathbf{a}_t^\top \mathbf{x}) + c_{0,t}(\mathbf{b}_t^\top \mathbf{y})} \cdot \mathbf{a}_t \\ \mathbf{b}_t \end{bmatrix} \\ &= \frac{\alpha(1 - c_{1,t})\|\mathbf{a}_t\|_2^2 + c_{0,t}\|\mathbf{b}_t\|_2^2}{(1 - c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \begin{bmatrix} \frac{\alpha c_{2,t}\|\mathbf{a}_t\|_2^2 + (1 - c_{1,t})\|\mathbf{b}_t\|_2^2}{\alpha(1 - c_{1,t})\|\mathbf{a}_t\|_2^2 + c_{0,t}\|\mathbf{b}_t\|_2^2} \cdot \mathbf{a}_t \\ \mathbf{b}_t \end{bmatrix} \doteq \lambda \cdot \tilde{\mathbf{x}}, \end{aligned} \quad (135)$$

and so we obtain the eigenvalue

$$\lambda = \frac{\alpha(1 - c_{1,t})\|\mathbf{a}_t\|_2^2 + c_{0,t}\|\mathbf{b}_t\|_2^2}{(1 - c_{1,t})^2 - c_{0,t}c_{2,t}}, \quad (136)$$

and we get from the eigenvector that α satisfies

$$\alpha = \frac{\alpha c_{2,t}\|\mathbf{a}_t\|_2^2 + (1 - c_{1,t})\|\mathbf{b}_t\|_2^2}{\alpha(1 - c_{1,t})\|\mathbf{a}_t\|_2^2 + c_{0,t}\|\mathbf{b}_t\|_2^2}, \quad (137)$$

and so

$$(1 - c_{1,t})\|\mathbf{a}_t\|_2^2\alpha^2 + (c_{0,t}\|\mathbf{b}_t\|_2^2 - c_{2,t}\|\mathbf{a}_t\|_2^2)\alpha - (1 - c_{1,t})\|\mathbf{b}_t\|_2^2 = 0. \quad (138)$$

We note that the discriminant is

$$\tau = (c_{0,t}\|\mathbf{b}_t\|_2^2 - c_{2,t}\|\mathbf{a}_t\|_2^2)^2 + 4(1 - c_{1,t})^2\|\mathbf{a}_t\|_2^2\|\mathbf{b}_t\|_2^2, \quad (139)$$

which is always > 0 . Therefore we always have two roots,

$$\alpha_{\pm} = \frac{c_{2,t}\|\mathbf{a}_t\|_2^2 - c_{0,t}\|\mathbf{b}_t\|_2^2 \pm \sqrt{(c_{0,t}\|\mathbf{b}_t\|_2^2 - c_{2,t}\|\mathbf{a}_t\|_2^2)^2 + 4(1 - c_{1,t})^2\|\mathbf{a}_t\|_2^2\|\mathbf{b}_t\|_2^2}}{2(1 - c_{1,t})\|\mathbf{a}_t\|_2^2}. \quad (140)$$

yielding two non-zero eigenvalues,

$$\lambda_{\pm}(U_t) = \frac{1}{2\rho} \cdot \left(c_{2,t}\|\mathbf{a}_t\|_2^2 + c_{0,t}\|\mathbf{b}_t\|_2^2 \pm \sqrt{(c_{0,t}\|\mathbf{b}_t\|_2^2 - c_{2,t}\|\mathbf{a}_t\|_2^2)^2 + 4(1 - c_{1,t})^2\|\mathbf{a}_t\|_2^2\|\mathbf{b}_t\|_2^2} \right). \quad (141)$$

Let us analyze the sign of both eigenvalues. For the numerator of λ_- to be negative, we have equivalently after simplification

$$(c_{2,t}\|\mathbf{a}_t\|_2^2 + c_{0,t}\|\mathbf{b}_t\|_2^2)^2 < (c_{0,t}\|\mathbf{b}_t\|_2^2 - c_{2,t}\|\mathbf{a}_t\|_2^2)^2 + 4(1 - c_{1,t})^2\|\mathbf{a}_t\|_2^2\|\mathbf{b}_t\|_2^2, \quad (142)$$

which simplifies in $c_{0,t}c_{2,t} < (1 - c_{1,t})^2$, i.e. $\rho > 0$. Hence, $\lambda_- < 0$.

Now, for λ_+ , it is easy to check that its sign is that of ρ . When $\rho > 0$, we have $\lambda_+ \geq |\lambda_-|$, and because $a^2 + b^2 \leq (|a| + |b|)^2$, we get

$$\begin{aligned}\lambda_1^\downarrow(U_t) = \lambda_+ &\leq \frac{1}{2} \cdot (c_{2,t}\|\mathbf{a}_t\|_2^2 + c_{0,t}\|\mathbf{b}_t\|_2^2 + |c_{0,t}\|\mathbf{b}_t\|_2^2 - c_{2,t}\|\mathbf{a}_t\|_2^2) + 2(1 - c_{1,t})\|\mathbf{a}_t\|_2\|\mathbf{b}_t\|_2 \\ &\leq c_{2,t}\|\mathbf{a}_t\|_2^2 + c_{0,t}\|\mathbf{b}_t\|_2^2 + (1 - c_{1,t})\|\mathbf{a}_t\|_2\|\mathbf{b}_t\|_2.\end{aligned}\quad (143)$$

Now, remark that because v_t is positive definite,

$$\begin{aligned}c_{0,t} - 2c_{1,t} + c_{2,t} &\doteq \mathbf{a}_t^{+\top} v_t \mathbf{a}_t^+ - 2\mathbf{a}_t^{+\top} v_t \mathbf{b}_t^+ + \mathbf{b}_t^{+\top} v_t \mathbf{b}_t^+ \\ &= (\mathbf{a}_t^+ - \mathbf{b}_t^+)^{\top} v_t (\mathbf{a}_t^+ - \mathbf{b}_t^+) \\ &\geq 0,\end{aligned}\quad (144)$$

showing that $c_{1,t} \leq (c_{0,t} + c_{2,t})/2$. So we get from ineq. (143),

$$\begin{aligned}\lambda_1^\downarrow(U_t) &\leq \frac{1}{\rho} \cdot \left(c_{2,t}\|\mathbf{a}_t\|_2^2 + c_{0,t}\|\mathbf{b}_t\|_2^2 + \left(1 + \frac{c_{0,t} + c_{2,t}}{2}\right) \|\mathbf{a}_t\|_2\|\mathbf{b}_t\|_2 \right) \\ &\leq \frac{1}{\rho} \cdot \left(1 + \frac{3}{2} \cdot (c_{0,t} + c_{2,t}) \right) \cdot \max\{\|\mathbf{a}_t\|_2^2, \|\mathbf{b}_t\|_2^2\} \\ &\leq \frac{2 + 3(c_{0,t} + c_{2,t})}{2((1 - c_{1,t})^2 - c_{0,t}c_{2,t})} \cdot \max\{\|\mathbf{a}_t\|_2^2, \|\mathbf{b}_t\|_2^2\}.\end{aligned}\quad (145)$$

When $\rho < 0$, we remark that $\lambda_+ < \lambda_-$ and so U_t is negative semi-definite.

Whenever $c_{1,t} \neq 1$, it is then easy to check that for any $z \in \{|\lambda_+|, |\lambda_-|\}$, ineq. (145) brings

$$z \leq \frac{2 + 3(c_{0,t} + c_{2,t})}{2|(1 - c_{1,t})^2 - c_{0,t}c_{2,t}|} \cdot \max\{\|\mathbf{a}_t\|_2^2, \|\mathbf{b}_t\|_2^2\}. \quad (146)$$

Whenever $c_{1,t} = 1$ (Case 1.), it is also immediate to check that for any $z \in \{|-1/\lambda(\mathbf{a}_t^+)|, |-1/\lambda(\mathbf{b}_t^+)|\}$,

$$\begin{aligned}z &\leq \max\left\{\frac{1}{c_{0,t}}, \frac{1}{c_{2,t}}\right\} \cdot \max\{\|\mathbf{a}_t\|_2^2, \|\mathbf{b}_t\|_2^2\} \\ &< \left(1 + \frac{3}{c_{0,t}} + \frac{3}{c_{2,t}}\right) \cdot \max\{\|\mathbf{a}_t\|_2^2, \|\mathbf{b}_t\|_2^2\} \\ &= \frac{2 + 3(c_{0,t} + c_{2,t})}{2|(1 - c_{1,t})^2 - c_{0,t}c_{2,t}|} \cdot \max\{\|\mathbf{a}_t\|_2^2, \|\mathbf{b}_t\|_2^2\}.\end{aligned}\quad (147)$$

Once we remark that $c_{1,t} = 1$ implies $\rho < 0$, we obtain the statement of Lemma T. ■

Lemma U *If P_t is (ε, τ) -accurate, then the following holds true:*

$$\|\mathbf{b}_t^+\|_2^2 = \|\mathbf{b}_t\|_2^2 \leq 2\xi \cdot X_*^2, \quad (148)$$

$$\|\mathbf{a}_t^+\|_2^2 = \|\mathbf{a}_t\|_2^2 \leq 2\xi \cdot X_*^2, \quad (149)$$

where ξ is defined in eq. (20).

Proof To prove ineq. (148), we make two applications of point 2. in the (ε, τ) -accuracy assumption with $F \doteq B$:

$$\begin{aligned} \varpi((\mathbf{x}_{u_B(t)} - \mathbf{x}_{v_B(t)})_F, \mathbf{w}_F) &\leq \varepsilon \cdot \max_{i \in \{u_B(t), v_B(t)\}} \varpi(\mathbf{x}_i, \mathbf{w}) + \tau , \\ \forall \mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 = 1 . \end{aligned} \quad (150)$$

Fix $\mathbf{w} \doteq (1/\|\mathbf{x}_{v_B(t)}\|_2) \cdot \mathbf{x}_{v_B(t)}$. We get:

$$\begin{aligned} |(\mathbf{x}_{v_B(t)} - \mathbf{x}_{u_B(t)})_B^\top \mathbf{x}_{v_B(t)_B}| &\leq \varepsilon \cdot \max\{|\mathbf{x}_{u_B(t)}^\top \mathbf{x}_{v_B(t)}|, \|\mathbf{x}_{v_B(t)}\|_2^2\} + \tau \cdot \|\mathbf{x}_{v_B(t)}\|_2 \\ &\leq \varepsilon \cdot X_*^2 + \tau \cdot X_* = \xi \cdot X_*^2 . \end{aligned} \quad (151)$$

Fix $\mathbf{w} \doteq (1/\|\mathbf{x}_{u_B(t)}\|_2) \cdot \mathbf{x}_{u_B(t)}$. We get:

$$\begin{aligned} |(\mathbf{x}_{u_B(t)} - \mathbf{x}_{v_B(t)})_B^\top \mathbf{x}_{u_B(t)_B}| &\leq \varepsilon \cdot \max\{|\mathbf{x}_{u_B(t)}^\top \mathbf{x}_{v_B(t)}|, \|\mathbf{x}_{u_B(t)}\|_2^2\} + \tau \cdot \|\mathbf{x}_{u_B(t)}\|_2 \\ &\leq \varepsilon \cdot X_*^2 + \tau \cdot X_* = \xi \cdot X_*^2 . \end{aligned} \quad (152)$$

Folding together ineqs. (151) and (152) yields

$$\begin{aligned} \|(\mathbf{x}_{v_B(t)} - \mathbf{x}_{u_B(t)})_B\|_2^2 &= (\mathbf{x}_{v_B(t)} - \mathbf{x}_{u_B(t)})_B^\top (\mathbf{x}_{v_B(t)} - \mathbf{x}_{u_B(t)})_B \\ &\leq |(\mathbf{x}_{v_B(t)} - \mathbf{x}_{u_B(t)})_B^\top \mathbf{x}_{v_B(t)_B}| + |(\mathbf{x}_{v_B(t)} - \mathbf{x}_{u_B(t)})_B^\top \mathbf{x}_{u_B(t)_B}| \\ &\leq 2\xi \cdot X_*^2 . \end{aligned} \quad (153)$$

We get

$$\|\mathbf{b}_t^+\|_2^2 = \|\mathbf{b}_t\|_2^2 = \|(\mathbf{x}_{v_B(t)} - \mathbf{x}_{u_B(t)})_B\|_2^2 \leq 2\xi \cdot X_*^2 , \quad (154)$$

which yields ineq. (148). To get ineq. (149), we switch $F \doteq B$ by $F \doteq A$ in our application of point 2. in the (ε, τ) -accuracy assumption. ■

Lemma V If P_t is (ε, τ) -accurate and the data-model calibration assumption holds,

$$c_{i,t} \leq \frac{1}{12} , \forall i \in \{0, 1, 2\} . \quad (155)$$

Proof We remark that

$$\begin{aligned} c_{0,t} &\doteq \mathbf{a}_t^{+\top} \mathbf{v}_t \mathbf{a}_t^+ \\ &\leq \lambda_1^\downarrow(\mathbf{v}_t) \|\mathbf{a}_t^+\|_2^2 \\ &\leq 2\lambda_1^\downarrow(\mathbf{v}_t) \xi \cdot X_*^2 , \end{aligned}$$

and for the same reasons, $c_{2,t} \leq 2\lambda_1^\downarrow(\mathbf{v}_t) \xi \cdot X_*^2$. Hence, it comes from the proof of Lemma T that we also have $c_{2,t} \leq 2\lambda_1^\downarrow(\mathbf{v}_t) \xi \cdot X_*^2$. Using ineq. (118) in Lemma S, we thus obtain for any $i \in \{0, 1, 2\}$:

$$\begin{aligned} c_{i,t} &\leq \frac{1}{n} \cdot \frac{2\xi \cdot X_*^2}{(1-\varepsilon)^2 \cdot \inf_{\mathbf{w}} \sigma^2(\{\varpi(\mathbf{x}_i, \mathbf{w})\}_{i=1}^n) + 8\gamma\lambda_1^\uparrow(\Gamma)} \\ &= \frac{\xi}{n} \cdot \frac{1}{4} \cdot \frac{X_*^2}{\frac{(1-\varepsilon)^2}{8} \cdot \inf_{\mathbf{w}} \sigma^2(\{\varpi(\mathbf{x}_i, \mathbf{w})\}_{i=1}^n) + \gamma\lambda_1^\uparrow(\Gamma)} \\ &\leq \frac{1}{4} \cdot \frac{1}{4} \cdot 1 < \frac{1}{12} , \end{aligned} \quad (156)$$

as claimed. The last inequality uses the data-model calibration assumption. \blacksquare

Corollary W Suppose P_t is (ε, τ) -accurate for any $t \geq 1$ and the data-model calibration assumption holds. Then the invertibility assumption holds.

Proof From Lemma V, we conclude that $(1 - c_{1,t})^2 > 121/144 > 1/144 > c_{0,t}c_{2,t} > 0$, hence the invertibility assumption holds. \blacksquare

Lemma X If P_t is (ε, τ) -accurate and the data-model calibration assumption holds, the following holds true: $I_d + \Lambda_t \succ 0$ and

$$\lambda_1^\downarrow(\Lambda_t) \leq \frac{\xi}{n}. \quad (157)$$

Proof First note that $\lambda_1^\uparrow(v_t) \geq 1/(\gamma\lambda_1^\downarrow(\Gamma)) > 0$ and so $v_t \succ 0$, which implies that $\Lambda_t = 2v_tU_t = 2v_t^{1/2}(v_t^{1/2}U_tv_t^{1/2})v_t^{-1/2}$, i.e. Λ_t is similar to a symmetric matrix $(v_t^{1/2}U_tv_t^{1/2})$ and therefore has only real eigenvalues. We get

$$\begin{aligned} \lambda_1^\downarrow(\Lambda_t) &= \lambda_1^\downarrow(2v_tU_t) \\ &\leq 2 \cdot \lambda_1^\downarrow(v_t) \cdot \left(1 + \frac{3}{2} \cdot (c_{0,t} + c_{2,t})\right) \cdot \max\{\|\mathbf{a}_t\|_2^2, \|\mathbf{b}_t\|_2^2\} \end{aligned} \quad (158)$$

$$\leq \frac{2 + 3(c_{0,t} + c_{2,t})}{|(1 - c_{1,t})^2 - c_{0,t}c_{2,t}|} \cdot 2\lambda_1^\downarrow(v_t)\xi \cdot X_*^2. \quad (159)$$

Ineq. (158) is due to Lemma T and ineq. (159) is due to Lemma U. We now use Lemma V and its proof, which shows that

$$\begin{aligned} (1 - c_{1,t})^2 - c_{0,t}c_{2,t} &\geq \left(1 - \frac{1}{12}\right)^2 - \frac{1}{144} \\ &= \frac{5}{6}. \end{aligned} \quad (160)$$

Letting $U = 2\lambda_1^\downarrow(v_t)\xi \cdot X_*^2$ for short, we thus get from the proof of Lemma V:

$$\begin{aligned} \lambda_1^\downarrow(\Lambda_t) &\leq \frac{6}{5} \cdot (2 + 3(U + U))U \\ &= \frac{6}{5} \cdot (2U + 6U^2). \end{aligned} \quad (161)$$

Now we want $\lambda_1^\downarrow(\Lambda_t) \leq \xi/n$, which translates into a second-order inequality for U , whose solution imposes the following upperbound on U :

$$6U \leq -1 + \sqrt{1 + \frac{5\xi}{n}}. \quad (162)$$

We can indeed forget the lowerbound for U , whose sign is negative while $U \geq 0$.

Since $\sqrt{1+x} \geq 1 + (x/2) - (x^2/8)$ for $x \geq 0$ (and $\xi/n \geq 0$), we get the sufficient condition for ineq. (162) to be satisfied:

$$12\lambda_1^\downarrow(\mathbf{v}_t)\xi \cdot X_*^2 \leq \frac{5\xi}{2n} - \frac{25}{8} \cdot \left(\frac{\xi}{n}\right)^2. \quad (163)$$

Now, it comes from Lemma S that a sufficient condition for ineq. (163) is that

$$\frac{\xi}{n} \cdot \frac{12X_*^2}{(1-\varepsilon)^2 \cdot \inf_{\mathbf{w}} \sigma^2(\mathbf{X}, \mathbf{w}) + 8\gamma\lambda_1^\uparrow(\Gamma)} \leq \frac{5\xi}{2n} - \frac{25}{8} \cdot \left(\frac{\xi}{n}\right)^2, \quad (164)$$

which, after simplification, is equivalent to

$$\frac{3}{5} \cdot \frac{X_*^2}{\frac{(1-\varepsilon)^2}{8} \cdot \inf_{\mathbf{w}} \sigma^2(\mathbf{X}, \mathbf{w}) + \gamma\lambda_1^\uparrow(\Gamma)} + \frac{5\xi}{4n} \leq 1. \quad (165)$$

But, the data-model calibration assumption implies that the left-hand side is no more than $(3/5) + (5/16) = 73/80 < 1$, and ineq. (157) follows.

It also trivially follows that $\mathbf{I}_d + \Lambda_t$ has only real eigenvalues. To prove that they are all strictly positive, we know that the only potentially negative eigenvalue of \mathbf{U}_t , λ_- (Lemma T) is smaller in absolute value to $\lambda_1^\downarrow(\mathbf{U}_t)$. \mathbf{v}_t being positive definite, we thus have under the (ε, τ) -accuracy assumption and data-model calibration:

$$\begin{aligned} \lambda_1^\uparrow(\mathbf{I}_d + \Lambda_t) &\geq 1 - \frac{\xi}{n} \\ &\geq 1 - \frac{1}{4} = \frac{3}{4} > 0, \end{aligned} \quad (166)$$

showing $\mathbf{I}_d + \Lambda_t$ is positive definite. This ends the proof of Lemma X. ■

Let $0 \leq T_+ \leq T$ denote the number of elementary permutations that act between classes, and $\rho \doteq T_+/T$ denote the proportion of such elementary permutations among all.

Theorem Y Suppose \mathbf{P}_* is (ε, τ) -accurate and α -bounded, and the data-model calibration assumption holds. Then the following holds for all $T \geq 1$:

$$\begin{aligned} \|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*\|_2 &\leq \frac{\xi}{n} \cdot T^2 \cdot \left(\|\boldsymbol{\theta}_0^*\|_2 + \frac{\sqrt{\xi}}{4X_*} \cdot \rho \right) \\ &\leq \left(\frac{\xi}{n} \right)^\alpha \cdot \left(\|\boldsymbol{\theta}_0^*\|_2 + \frac{\sqrt{\xi}}{4X_*} \cdot \rho \right). \end{aligned} \quad (167)$$

Proof We use Theorem Q, which yields from the triangle inequality:

$$\|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*\|_2 = \|(\mathbf{H}_{T,0} - \mathbf{I}_d)\boldsymbol{\theta}_0^*\|_2 + \left\| \sum_{t=0}^{T-1} \mathbf{H}_{T,t+1} \boldsymbol{\lambda}_t \right\|_2. \quad (168)$$

Denote for short $q \doteq \xi/n$. It comes from the definition of $H_{i,j}$ and Lemma X the first inequality of:

$$\begin{aligned}\lambda_1^\downarrow(H_{T,0} - I_d) &\leq (1+q)^T - 1 \\ &\leq T^2 q ,\end{aligned}\quad (169)$$

where the second inequality holds because $\binom{T}{k} q^k \leq (Tq)^k \leq Tq$ for $k \geq 1$ whenever $Tq \leq 1$, which is equivalent to

$$T \leq \frac{n}{\xi} , \quad (170)$$

which is implied by the condition of α -bounded permutation size ($n/\xi \geq 4 \geq 1$ from the data-model calibration assumption). We thus get

$$\| (H_{T,0} - I_d) \theta_0^* \|_2 \leq T^2 q \cdot \|\theta_0^*\|_2 . \quad (171)$$

Using ineq. (168), this shows the statement of the Theorem with (187). The upperbound comes from the fact that the factor in the right hand side is no more than $(\xi/n)^\alpha$ for some $0 \leq \alpha \leq 1$ provided this time the stronger constraint holds:

$$T \leq \left(\frac{n}{\xi} \right)^{\frac{1-\alpha}{2}} , \quad (172)$$

which is the condition of α -boundedness.

Let us now have a look at the shift term in eq. (168), which depends only on the mistakes between classes done during the permutation (which changes the mean operator between permutations),

$$R \doteq \sum_{t=0}^{T-1} H_{T,t+1} \lambda_t . \quad (173)$$

Using eq. (77), we can simplify R since $\lambda_t = 2v_{t+1}\varepsilon_t$, so if we define $G_{i,j}$ from $H_{i,j}$ as follows, for $0 \leq j \leq i$:

$$G_{i,j} \doteq 2H_{i,j}v_j , \quad (174)$$

then we get

$$R \doteq \sum_{t=0}^{T-1} G_{T,t+1} \varepsilon_t , \quad (175)$$

where we recall that $\varepsilon_t \doteq \mu_{t+1} - \mu_t$ is the shift in the mean operator, *which is the null vector whenever P_t acts in a specific class ($y_{u_A(t)} = y_{v_A(t)}$)*. To see this, we remark

$$\begin{aligned}\varepsilon_t &\doteq \mu_{t+1} - \mu_t \\ &= \sum_i y_i \cdot \left[\frac{\mathbf{x}_{i_A}}{\mathbf{x}_{(t+1)i_B}} \right] - \sum_i y_i \cdot \left[\frac{\mathbf{x}_{i_A}}{\mathbf{x}_{ti_B}} \right] \\ &= \sum_i y_i \cdot \left[\frac{0}{\mathbf{x}_{(t+1)i_B}} \right] - \sum_i y_i \cdot \left[\frac{0}{\mathbf{x}_{ti_B}} \right] \\ &= \left[\frac{0}{\sum_i y_i \cdot (\mathbf{x}_{(t+1)i_B} - \mathbf{x}_{ti_B})} \right] \doteq \left[\frac{0}{\varepsilon_{tB}} \right] ,\end{aligned}\quad (176)$$

which can be simplified further since we work with the elementary permutation P_t ,

$$\begin{aligned}\boldsymbol{\varepsilon}_{t\mathbf{B}} &= y_{u_{\mathbf{A}}(t)} \cdot (\mathbf{x}_{v_{\mathbf{B}}(t)} - \mathbf{x}_{u_{\mathbf{B}}(t)})_{\mathbf{B}} + y_{v_{\mathbf{A}}(t)} \cdot (\mathbf{x}_{u_{\mathbf{B}}(t)} - \mathbf{x}_{v_{\mathbf{B}}(t)})_{\mathbf{B}} \\ &= (y_{u_{\mathbf{A}}(t)} - y_{v_{\mathbf{A}}(t)}) \cdot (\mathbf{x}_{v_{\mathbf{B}}(t)} - \mathbf{x}_{u_{\mathbf{B}}(t)})_{\mathbf{B}}.\end{aligned}\quad (177)$$

Hence,

$$\begin{aligned}\|\boldsymbol{\varepsilon}_t\|_2 &= \|\boldsymbol{\varepsilon}_{t\mathbf{B}}\|_2 = 1_{y_{u_{\mathbf{A}}(t)} \neq y_{v_{\mathbf{A}}(t)}} \cdot \|(\mathbf{x}_{v_{\mathbf{B}}(t)} - \mathbf{x}_{u_{\mathbf{B}}(t)})_{\mathbf{B}}\|_2 \\ &\leq 1_{y_{u_{\mathbf{A}}(t)} \neq y_{v_{\mathbf{A}}(t)}} \cdot \sqrt{2\xi} X_* ,\end{aligned}\quad (178)$$

from Lemma U, and we see that indeed $\|\boldsymbol{\varepsilon}_t\|_2 = 0$ when the elementary permutation occurs within observations of the same class.

It follows from the data-model calibration assumption and Lemma X that

$$\begin{aligned}\lambda_1^{\downarrow}(v_t) &\leq \frac{1}{n} \cdot \frac{1}{(1-\varepsilon)^2 \cdot \inf_{\mathbf{w}} \sigma^2(\{\varpi(\mathbf{x}_i, \mathbf{w})\}_{i=1}^n) + 8\gamma\lambda_1^{\uparrow}(\Gamma)} \\ &= \frac{1}{8nX_*^2} \cdot \frac{X_*^2}{\frac{(1-\varepsilon)^2}{8} \cdot \inf_{\mathbf{w}} \sigma^2(\{\varpi(\mathbf{x}_i, \mathbf{w})\}_{i=1}^n) + \gamma\lambda_1^{\uparrow}(\Gamma)} \\ &\leq \frac{1}{8nX_*^2} .\end{aligned}\quad (179)$$

Using [Bhatia, 1997, Problem III.6.14], Lemma X and ineq. (179), we also obtain

$$\lambda_1^{\downarrow}(G_{T,t+1}) \leq 2 \cdot \left(1 + \frac{\xi}{n}\right)^{T-t-1} \cdot \frac{1}{8nX_*^2} .\quad (180)$$

So,

$$\begin{aligned}\|\mathbf{R}\|_2 &\leq \sum_{t=0}^{T-1} \lambda_{\max}(G_{T,t+1}) \|\boldsymbol{\varepsilon}_t\|_2 \\ &\leq \frac{1}{2\sqrt{2}} \cdot \sum_{t=0}^{T-1} 1_{y_{u_{\mathbf{A}}(t)} \neq y_{v_{\mathbf{A}}(t)}} \cdot \left(1 + \frac{\xi}{n}\right)^{T-t-1} \cdot \frac{\sqrt{\xi}}{nX_*} \\ &= \frac{1}{2X_*} \cdot \sqrt{\frac{\xi}{2}} \cdot \sum_{t=0}^{T-1} 1_{y_{u_{\mathbf{A}}(t)} \neq y_{v_{\mathbf{A}}(t)}} \cdot \left(1 + \frac{\xi}{n}\right)^{T-t-1} \cdot \frac{\xi}{n} ,\end{aligned}\quad (181)$$

from ineq. (178). Assuming $T_+ \leq T$ errors are made by permutations between classes and recalling $q \doteq \xi/n$, we see that the largest upperbound for $\|\mathbf{R}\|_2$ in ineq. (181) is obtained when all T_+ errors happen at the last elementary permutations in the sequence in P_* , so we get that

$$\begin{aligned}\|\mathbf{R}\|_2 &\leq \frac{1}{2X_*} \cdot \sqrt{\frac{\xi}{2}} \cdot \sum_{t=0}^{T_+-1} q(1+q)^{T-t-1} \\ &= \frac{1}{2X_*} \cdot \sqrt{\frac{\xi}{2}} \cdot q(1+q)^{T-T_+} \sum_{t=0}^{T_+-1} (1+q)^{T_+-t-1} \\ &= \frac{1}{2X_*} \cdot \sqrt{\frac{\xi}{2}} \cdot (1+q)^{T-T_+} ((1+q)^{T_+} - 1) .\end{aligned}\quad (182)$$

It comes from ineq. (169) $(1 + q)^{T_+} - 1 \leq T_+^2 q$ and

$$\begin{aligned}
(1 + q)^{T-T_+} &\leq (T - T_+)^2 q + 1 \\
&\leq \left(\frac{n}{\xi}\right)^{1-\alpha} \cdot \frac{\xi}{n} + 1 \\
&= \left(\frac{\xi}{n}\right)^\alpha + 1 \\
&\leq \frac{1}{4} + 1 < \sqrt{2} .
\end{aligned} \tag{183}$$

The last line is due to the data-model calibration assumption. We finally get from ineq. (182)

$$\begin{aligned}
\|\mathbf{R}\|_2 &\leq \frac{\sqrt{\xi}}{2X_*} \cdot \frac{\xi}{n} \cdot T_+^2 \\
&= \frac{\xi^{3/2}}{4X_* n} \cdot T_+^2 .
\end{aligned} \tag{184}$$

We also remark that if P_* is α -bounded, since $T_+ \leq T$, we also have:

$$\begin{aligned}
\frac{\xi^{3/2}}{4X_* n} \cdot T_+^2 &\leq \frac{\xi^{3/2}}{4X_* n} \cdot \left(\frac{n}{\xi}\right)^{1-\alpha} \\
&= \frac{\sqrt{\xi}}{4X_*} \cdot \left(\frac{\xi}{n}\right)^\alpha .
\end{aligned} \tag{185}$$

Summarizing, we get

$$\|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*\|_2 \leq a(T) \cdot \|\boldsymbol{\theta}_0^*\|_2 + b(T_+) , \tag{186}$$

where

$$a(T) \doteq \frac{\xi}{n} \cdot T^2 \leq \left(\frac{\xi}{n}\right)^\alpha , \tag{187}$$

$$b(T_+) \doteq \frac{\xi^{3/2}}{4X_* n} \cdot T_+^2 \leq \frac{\sqrt{\xi}}{4X_*} \cdot \left(\frac{\xi}{n}\right)^\alpha , \tag{188}$$

which yields the proof of Theorem Y. ■

Theorem Y easily yields the proof of Theorem 6.

V.5 Proof of Theorem 8

Remark that for any example (\mathbf{x}, y) , we have from Cauchy-Schwartz inequality:

$$\begin{aligned}
|y(\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*)^\top \mathbf{x}| &= |(\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*)^\top \mathbf{x}| \leq \|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*\|_2 \|\mathbf{x}\|_2 \\
&\leq \left(\frac{\xi}{n}\right)^\alpha \cdot \left(\|\boldsymbol{\theta}_0^*\|_2 + \frac{\sqrt{\xi}}{4X_*} \cdot \rho\right) \cdot X_* \\
&= \left(\frac{\xi}{n}\right)^\alpha \cdot \left(\|\boldsymbol{\theta}_0^*\|_2 X_* + \frac{\sqrt{\xi}}{4} \cdot \rho\right) .
\end{aligned} \tag{189}$$

So, to have $|y(\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*)^\top \mathbf{x}| < \kappa$ for some $\kappa > 0$, it is sufficient that

$$n > \xi \cdot \left(\frac{\|\boldsymbol{\theta}_0^*\|_2 X_*}{\kappa} + \frac{\sqrt{\xi}}{4\kappa} \cdot \rho \right)^{\frac{1}{\alpha}} . \quad (190)$$

In this case, for any example (\mathbf{x}, y) such that $y(\boldsymbol{\theta}_0^*)^\top \mathbf{x} > \kappa$, then

$$\begin{aligned} y(\boldsymbol{\theta}_T^*)^\top \mathbf{x} &= y(\boldsymbol{\theta}_0^*)^\top \mathbf{x} + y(\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*)^\top \mathbf{x} \\ &\geq y(\boldsymbol{\theta}_0^*)^\top \mathbf{x} - |y(\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*)^\top \mathbf{x}| \\ &> \kappa - \kappa = 0 , \end{aligned} \quad (191)$$

and we get the statement of the Theorem.

V.6 Proof of Theorem 9

We want to bound the difference between the loss *over the true data* for the optimal (unknown) classifier $\boldsymbol{\theta}_0^*$ and the classifier we learn from entity resolved data, $\boldsymbol{\theta}_T^*$,

$$\Delta_S(\boldsymbol{\theta}_0^*, \boldsymbol{\theta}_T^*) \doteq \ell_{S,\gamma}(\boldsymbol{\theta}_T^*) - \ell_{S,\gamma}(\boldsymbol{\theta}_0^*) . \quad (192)$$

Simple arithmetics and Cauchy-Schwartz inequality allow to derive:

$$\begin{aligned} \Delta_S(\boldsymbol{\theta}_0^*, \boldsymbol{\theta}_T^*) &= \frac{1}{2n} \cdot \left((\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*)^\top \left(\sum_i y_i \mathbf{x}_i \right) + \frac{1}{4} \cdot \sum_i ((\boldsymbol{\theta}_0^*)^\top \mathbf{x}_i)^2 - ((\boldsymbol{\theta}_T^*)^\top \mathbf{x}_i)^2 \right) \\ &= \frac{1}{2n} \cdot \left((\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*)^\top \boldsymbol{\mu}_0 + \frac{1}{4} \cdot \sum_i ((\boldsymbol{\theta}_0^* - \boldsymbol{\theta}_T^*)^\top \mathbf{x}_i) ((\boldsymbol{\theta}_0^* + \boldsymbol{\theta}_T^*)^\top \mathbf{x}_i) \right) \\ &\leq \frac{1}{2} \cdot \left(\frac{1}{n} \cdot \|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*\|_2 \|\boldsymbol{\mu}_0\|_2 + \|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*\|_2 \|\boldsymbol{\theta}_T^* + \boldsymbol{\theta}_0^*\|_2 X_*^2 \right) \\ &= \|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*\|_2 \cdot \left(\frac{\|\boldsymbol{\mu}_0\|_2}{2n} + \frac{1}{2} \cdot \|\boldsymbol{\theta}_T^* + \boldsymbol{\theta}_0^*\|_2 X_*^2 \right) . \end{aligned} \quad (193)$$

We now need to bound $\|\boldsymbol{\theta}_T^* + \boldsymbol{\theta}_0^*\|_2$, which is easy since the triangle inequality yields

$$\begin{aligned} \|\boldsymbol{\theta}_T^* + \boldsymbol{\theta}_0^*\|_2 &= \|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^* + 2\boldsymbol{\theta}_0^*\|_2 \\ &\leq \|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*\|_2 + 2\|\boldsymbol{\theta}_0^*\|_2 , \end{aligned} \quad (194)$$

and so Theorem Y yields:

$$\begin{aligned} \|\boldsymbol{\theta}_T^* + \boldsymbol{\theta}_0^*\|_2 &\leq 2\|\boldsymbol{\theta}_0^*\|_2 + \frac{\xi}{n} \cdot T^2 \cdot \left(\|\boldsymbol{\theta}_0^*\|_2 + \frac{\sqrt{\xi}}{4X_*} \cdot \rho \right) \\ &= 2\|\boldsymbol{\theta}_0^*\|_2 + \frac{\xi(\delta_m + \delta_\rho)}{nX_*} \cdot T^2 , \end{aligned} \quad (195)$$

where $\delta_m \doteq \|\boldsymbol{\theta}_0^*\|_2 X_*$ is the maximum margin for the optimal (unknown) classifier and $\delta_\rho \doteq \sqrt{\xi}\rho/4$ is a penalty due to class-mismatch permutations. Denote for short

$$\eta \doteq \|\boldsymbol{\theta}_0^*\|_2 + \frac{\sqrt{\xi}}{4X_*} \cdot \rho = \frac{\delta_m + \delta_\rho}{X_*} . \quad (196)$$

We finally obtain, letting $\delta_{\mu_0} \doteq \|\boldsymbol{\mu}_0\|_2/(nX_*)$ ($\in [0, 1]$) denote the normalized mean-operator for the true dataset,

$$\begin{aligned}
\Delta_S(\boldsymbol{\theta}_0^*, \boldsymbol{\theta}_T^*) &\leq \frac{\xi\eta}{n} \cdot T^2 \cdot \left(\frac{\|\boldsymbol{\mu}_0\|_2}{2n} + \frac{\xi\eta X_*^2}{2n} \cdot T^2 + \|\boldsymbol{\theta}_0^*\|_2 X_*^2 \right) \\
&= \frac{\xi\eta}{n} \cdot T^2 \cdot \left(\frac{\|\boldsymbol{\mu}_0\|_2}{2n} + \|\boldsymbol{\theta}_0^*\|_2 X_*^2 \cdot \left(1 + \frac{\xi}{2n} \cdot T^2 \right) + \frac{\xi^{3/2}\rho X_*}{8n} \cdot T^2 \right) \\
&= \frac{\xi\eta X_*}{n} \cdot T^2 \cdot \left(\frac{\|\boldsymbol{\mu}_0\|_2}{2nX_*} + \|\boldsymbol{\theta}_0^*\|_2 X_* \cdot \left(1 + \frac{\xi}{2n} \cdot T^2 \right) + \frac{\xi^{3/2}\rho}{8n} \cdot T^2 \right) \\
&= \frac{\xi(\delta_m + \delta_\rho)}{n} \cdot T^2 \cdot \left(\frac{\delta_{\mu_0}}{2} + \delta_m \cdot \left(1 + \frac{\xi}{2n} \cdot T^2 \right) + \delta_\rho \cdot \frac{\xi}{2n} \cdot T^2 \right). \quad (197)
\end{aligned}$$

Let us denote for short

$$C(n) \doteq \frac{\xi}{n} \cdot T^2. \quad (198)$$

We know that under the α -boundedness condition,

$$\begin{aligned}
C(n) &\leq \left(\frac{\xi}{n} \right)^\alpha \\
&= o(1),
\end{aligned} \quad (199)$$

and we finally get from eq. (197),

$$\begin{aligned}
\Delta_S(\boldsymbol{\theta}_0^*, \boldsymbol{\theta}_T^*) &\leq \frac{\delta_m + \delta_\rho}{2} \cdot C(n) \cdot (\delta_{\mu_0} + \delta_m \cdot (2 + C(n)) + \delta_\rho \cdot C(n)) \\
&\leq \bar{\delta}_{m,\rho} (\delta_{\mu_0} + 6\bar{\delta}_{m,\rho}) \cdot C(n),
\end{aligned} \quad (200)$$

with $\bar{\delta}_{m,\rho} \doteq (\delta_m + \delta_\rho)/2$ the average of the margin and class-mismatch penalties. We have used in the last inequality the fact that under the data-model calibration assumption, $C(n) \leq (1/4)^\alpha \leq 1$. This ends the proof of Theorem 9.

V.7 Proof of Theorem 10

The Rademacher complexity is a fundamental notion in learning [Bartlett and Mendelson, 2002]. Letting $\Sigma_n \doteq \{-1, 1\}^n$, the empirical Rademacher complexity of hypothesis class \mathcal{H} is:

$$R_n \doteq \mathbb{E}_{\sigma \sim \Sigma_n} \sup_{h \in \mathcal{H}} \{ \mathbb{E}_S[\sigma(\mathbf{x})h(\mathbf{x})] \}. \quad (201)$$

When it comes to linear classifiers, two parameters are fundamental to bound the empirical Rademacher complexity: the maximum norm of the classifier and the maximum norm of an observation [Kakade et al., 2008, Patrini et al., 2016a]. There are therefore two Rademacher complexities that are relevant to our setting:

- the one related to the unknown optimal classifier $\boldsymbol{\theta}_0$, $R_n(\boldsymbol{\theta}_0)$, to which we attach maximum classifier norm θ_* and example norm X_* . It is well-known that we have [Kakade et al., 2008, Theorem 3], [Patrini et al., 2016a, Lemma 1]:

$$R_n(\boldsymbol{\theta}_0) \leq \frac{X_* \theta_*}{\sqrt{n}} . \quad (202)$$

- the one related to the optimal classifier that we build on our observed data, $R_n(\boldsymbol{\theta}_T)$.

Notice that this latter one should *also* be computed over the optimal dataset S . One may wonder how this would degrade any empirically computable bound, that would depend on \hat{S}_T instead of S in the supremum in eq. (201). It can be shown that entity resolution has this very desirable property in the vertical partition setting that we can in fact use any empirical upperbound on X_* , the way the bound in eq. (201) is computed is not going to be affected. This is what we now show.

Lemma Z *Let $\hat{X}_{T*} = \max_i \|\hat{x}_{Ti}\|_2$ and $\hat{\theta}_*$ be any upperbound for $\|\boldsymbol{\theta}_T\|_2$. We have*

$$R_n(\boldsymbol{\theta}_T) \leq \frac{\hat{X}_{T*} \hat{\theta}_*}{\sqrt{n}} \quad (203)$$

Proof The proof follows the basic steps of [Patrini et al., 2016a, Lemma 1], with a twist to handle the replacement of X_* by \hat{X}_{T*} . We give the proof for completeness. We have the key observation that $\forall \boldsymbol{\sigma} \in \Sigma_n$,

$$\begin{aligned} \arg \sup_{\boldsymbol{\theta} \in \mathcal{H}_*} \{ \mathbb{E}_S[\sigma(\mathbf{x}) \boldsymbol{\theta}^\top \mathbf{x}] \} &= \frac{1}{n} \arg \sup_{\boldsymbol{\theta} \in \mathcal{H}_*} \left\{ \sum_i \sigma_i \boldsymbol{\theta}^\top \mathbf{x}_{Ti} \right\} \\ &= \frac{\sup_{\mathcal{H}_*} \|\boldsymbol{\theta}\|_2}{\|\sum_i \sigma_i \mathbf{x}_{Ti}\|_2} \sum_i \sigma_i \mathbf{x}_{Ti} . \end{aligned} \quad (204)$$

So,

$$\begin{aligned} R_n &= \mathbb{E}_{\Sigma_n} \sup_{h \in \mathcal{H}} \{ \mathbb{E}_S[\sigma(\mathbf{x}) h(\mathbf{x})] \} \\ &= \frac{\sup_{\mathcal{H}_*} \|\boldsymbol{\theta}\|_2}{n} \mathbb{E}_{\Sigma_n} \left[\frac{(\sum_{\mathbf{x}} \sigma(\mathbf{x}) \mathbf{x})^\top (\sum_{\mathbf{x}} \sigma(\mathbf{x}) \mathbf{x})}{\|\sum_{\mathbf{x}} \sigma(\mathbf{x}) \mathbf{x}\|_2} \right] \\ &= \frac{\sup_{\mathcal{H}_*} \|\boldsymbol{\theta}\|_2}{n} \mathbb{E}_{\Sigma_n} \left[\left\| \sum_{\mathbf{x}} \sigma(\mathbf{x}) \mathbf{x} \right\|_2 \right] \\ &= \frac{\sup_{\mathcal{H}_*} \|\boldsymbol{\theta}\|_2}{n} \times \frac{1}{|\Sigma_n|} \sum_{\Sigma_n} \sqrt{\sum_i \|\mathbf{x}_i\|_2^2 + \sum_{i_1 \neq i_2} \sigma_{i_1} \sigma_{i_2} \mathbf{x}_{i_1}^\top \mathbf{x}_{i_2}} \\ &= \frac{\sup_{\mathcal{H}_*} \|\boldsymbol{\theta}\|_2 \sqrt{\sum_i \|\mathbf{x}_i\|_2^2}}{n} \times \frac{1}{|\Sigma_n|} \sum_{\Sigma_n} \sqrt{1 + \frac{\sum_{i_1 \neq i_2} \sigma_{i_1} \sigma_{i_2} \mathbf{x}_{i_1}^\top \mathbf{x}_{i_2}}{\sum_i \|\mathbf{x}_i\|_2^2}} . \end{aligned}$$

Using the fact that $\sqrt{1+x} \leq 1 + x/2$, we have

$$\begin{aligned}
R_n &\leq \frac{\sup_{\mathcal{H}_*} \|\theta\|_2 \sqrt{\sum_i \|\mathbf{x}_i\|_2^2}}{n} \times \frac{1}{|\Sigma_n|} \sum_{\Sigma_n} \left(1 + \frac{\sum_{i_1 \neq i_2} \sigma_{i_1} \sigma_{i_2} \mathbf{x}_{i_1}^\top \mathbf{x}_{i_2}}{2 \sum_i \|\mathbf{x}_i\|_2^2} \right) \\
&= \frac{\sup_{\mathcal{H}_*} \|\theta\|_2 \sqrt{\sum_i \|\mathbf{x}_i\|_2^2}}{n} \left(1 + \frac{1}{2|\Sigma_n|} \sum_{\Sigma_n} \frac{\sum_{i_1 \neq i_2} \sigma_{i_1} \sigma_{i_2} \mathbf{x}_{i_1}^\top \mathbf{x}_{i_2}}{\sum_i \|\mathbf{x}_i\|_2^2} \right) \\
&= \frac{\sup_{\mathcal{H}_*} \|\theta\|_2 \sqrt{\sum_i \|\hat{\mathbf{x}}_{Ti}\|_2^2}}{n} \left(1 + \frac{1}{2|\Sigma_n|} \sum_{\Sigma_n} \frac{\sum_{i_1 \neq i_2} \sigma_{i_1} \sigma_{i_2} \mathbf{x}_{i_1}^\top \mathbf{x}_{i_2}}{\sum_i \|\mathbf{x}_i\|_2^2} \right). \quad (205)
\end{aligned}$$

We prove eq. (205). Because P_* is a permutation, we have the penultimate identity of:

$$\sum_i \|\hat{\mathbf{x}}_{Ti}\|_2^2 = \sum_i \|(\hat{\mathbf{x}}_{Ti})_{\mathbb{A}}\|_2^2 + \sum_i \|(\hat{\mathbf{x}}_{Ti})_{\mathbb{B}}\|_2^2 \quad (206)$$

$$= \sum_i \|(\mathbf{x}_i)_{\mathbb{A}}\|_2^2 + \sum_i \|(\hat{\mathbf{x}}_{Ti})_{\mathbb{B}}\|_2^2 \quad (207)$$

$$= \sum_i \|(\mathbf{x}_i)_{\mathbb{A}}\|_2^2 + \sum_i \|(\mathbf{x}_i)_{\mathbb{B}}\|_2^2 \quad (208)$$

$$= \sum_i \|\mathbf{x}_i\|_2^2. \quad (209)$$

The second identity follows from the fact that P_* acts on indexes from \mathbb{B} . We obtain

$$R_n \leq \frac{\hat{X}_{T*} \hat{\theta}_*}{\sqrt{n}} \quad (210)$$

after remarking, following [Patrini et al., 2016a, Lemma 1], that because Rademacher variables are i.i.d., $\sum_{\Sigma_n} \sigma_{i_1} \sigma_{i_2} = 0$, which yields:

$$\sum_{\Sigma_n} \frac{\sum_{i_1 \neq i_2} \sigma_{i_1} \sigma_{i_2} \mathbf{x}_{i_1}^\top \mathbf{x}_{i_2}}{\sum_i \|\mathbf{x}_i\|_2^2} = \frac{1}{\sum_i \|\mathbf{x}_i\|_2^2} \sum_{i_1 \neq i_2} \mathbf{x}_{i_1}^\top \mathbf{x}_{i_2} \sum_{\Sigma_n} \sigma_{i_1} \sigma_{i_2} = 0, \quad (211)$$

and ends the proof of Lemma Z. ■

We now remark that we have from Theorem 6 and the triangle inequality:

$$\|\boldsymbol{\theta}_T^*\|_2 \leq \left(1 + C(n) \cdot \left(1 + \frac{\delta_\rho}{\delta_m} \right) \right) \cdot \|\boldsymbol{\theta}_0^*\|_2, \quad (212)$$

which means that we can let

$$\hat{\theta}_* \doteq \left(1 + C(n) \cdot \left(1 + \frac{\delta_\rho}{\delta_m} \right) \right) \cdot \theta_*. \quad (213)$$

Letting L denote the Lipschitz constant for the Taylor loss, we get from [Bartlett and Mendelson, 2002, Theorem 7] that with probability $\geq 1 - \delta$ over the drawing of $S \sim \mathcal{D}^n$,

$$\Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [y(\boldsymbol{\theta}_T^*)^\top \mathbf{x} \leq 0] \leq \ell_{S, \gamma}(\boldsymbol{\theta}_T^*) + 2L R_n(\boldsymbol{\theta}_T^*) + \sqrt{\frac{\ln(2/\delta)}{2n}}. \quad (214)$$

So, using Theorem 9, ineq. (214) implies

$$\begin{aligned}
\Pr_{(\mathbf{x}, y) \sim \mathcal{D}} [y(\boldsymbol{\theta}_T^*)^\top \mathbf{x} \leq 0] &\leq \ell_{S, \gamma}(\boldsymbol{\theta}_0^*) + \bar{\delta}_{m, \rho} (\delta_{\mu_0} + 6\bar{\delta}_{m, \rho}) \cdot C(n) + \frac{2LX_*}{\sqrt{n}} \cdot \hat{\theta}_* + \sqrt{\frac{\ln(2/\delta)}{2n}} \\
&\leq \ell_{S, \gamma}(\boldsymbol{\theta}_0^*) + \bar{\delta}_{m, \rho} (\delta_{\mu_0} + 6\bar{\delta}_{m, \rho}) \cdot C(n) \\
&\quad + \frac{2LX_*}{\sqrt{n}} \cdot \left(1 + C(n) \cdot \left(1 + \frac{\delta_\rho}{\delta_m}\right)\right) \cdot \theta_* + \sqrt{\frac{\ln(2/\delta)}{2n}} \\
&\leq \ell_{S, \gamma}(\boldsymbol{\theta}_0^*) + \bar{\delta}_{m, \rho} (\delta_{\mu_0} + 6\bar{\delta}_{m, \rho}) \cdot C(n) \\
&\quad + \frac{2LX_*}{\sqrt{n}} \cdot \left(1 + C(n) \cdot \left(1 + \frac{\delta_\rho}{\delta_m}\right)\right) \cdot \theta_* + \sqrt{\frac{\ln(2/\delta)}{2n}} \\
&= \ell_{S, \gamma}(\boldsymbol{\theta}_0^*) + \frac{2LX_* \theta_*}{\sqrt{n}} + \sqrt{\frac{\ln(2/\delta)}{2n}} + U(n) ,
\end{aligned} \tag{215}$$

with

$$\begin{aligned}
U(n) &= \left(\bar{\delta}_{m, \rho} (\delta_{\mu_0} + 6\bar{\delta}_{m, \rho}) + \frac{2LX_* \theta_*}{\sqrt{n}} \cdot \left(1 + \frac{\delta_\rho}{\delta_m}\right) \right) \cdot C(n) \\
&= \left(\bar{\delta}_{m, \rho} (\delta_{\mu_0} + 6\bar{\delta}_{m, \rho}) + \frac{2L}{\sqrt{n}} \cdot (\delta_m + \delta_\rho) \right) \cdot C(n) \\
&= \bar{\delta}_{m, \rho} \cdot \left(\delta_{\mu_0} + 6\bar{\delta}_{m, \rho} + \frac{4L}{\sqrt{n}} \right) \cdot C(n) ,
\end{aligned} \tag{216}$$

achieving the proof of Theorem 10.

VI Cryptographic longterm keys, entity resolution and learning

We stressed the importance of Theorem 6, which is applicable regardless of the permutation matrix and in fact also relevant to the non private setting. While we think this Theorem may be useful to optimize entity resolution algorithms with the objective to learn more accurately afterwards, a question is how does our theory fits to the actual system we are using, *cryptographic longterm keys*, (CLKs, Schnell et al. [2011]).

To summarize, a CLK proceed in two phases: (i) it hashes the input in a set of objects (*e.g.* bigrams, digits, etc.), (ii) for each object, it hashes *several times* its binary code into a Bloom filter of fixed size, initialized to 0, flipping the zeroes to 1 for each hashing value. Hence, computing a CLK amounts to putting at most hs bits to 1 in an array of l bits, where h is the number of hash functions and s is the number of objects. Comparing two binary CLKs z and z' is done with the dice coefficient:

$$D(z, z') \doteq 2 \times \frac{1(z \wedge z')}{1(z) + 1(z')} , \tag{217}$$

where $1(z)$ is the set of indexes having 1 in the Bloom filter (binary) encoding of z . Since $D(z, z') \in [0, 1]$, a threshold $\tau \in [0, 1]$ is in general learned, above which z and z' are considered representing the same inputs.

CLKs have the key property that if the two inputs are the same, then $D(z, z') = 1 \geq \tau$, for any applicable τ . Therefore, whenever two entities from A and B have the same input values, *if* these inputs do not have errors, then their Dice coefficient is maximal and they are recognized as representing the same entity.

Suppose for simplicity that we study the (ε, δ) -accuracy assumptions on binary vectors, the common part of A and B being the representation of the CLK. In this case, ignoring the features not in the Bloom filter, it is not hard to see that the unit vector \mathbf{w} which maximizes the stretch $\varpi((\hat{\mathbf{x}}_{ti} - \mathbf{x}_i)_{\text{CLK}}, \mathbf{w}_{\text{CLK}})$ (the features of CLK are included in those of B) is the one whose coordinates over the l bits of the CLK, in absolute value, are proportional to the coordinates of $\hat{\mathbf{x}}_{ti} \oplus \mathbf{x}_i$, where \oplus is the exclusive or. We get for this \mathbf{w}

$$\varpi((\hat{\mathbf{x}}_{ti} - \mathbf{x}_i)_{\text{CLK}}, \mathbf{w}_{\text{CLK}}) = \sqrt{1(\hat{\mathbf{x}}_{ti} \oplus \mathbf{x}_i)} , \quad (218)$$

while for this \mathbf{w} , we also get

$$\varpi(\mathbf{x}_i, \mathbf{w}) = \sqrt{1(\mathbf{x}_i \wedge \neg \hat{\mathbf{x}}_{ti})} . \quad (219)$$

Assuming the CLKs have exactly hs bits to 1, we have $1(\hat{\mathbf{x}}_{ti} \oplus \mathbf{x}_i) = 2 \cdot 1(\mathbf{x}_i \wedge \neg \hat{\mathbf{x}}_{ti})$. We have $\sqrt{2} - 1 \approx 0.414$, so depending on the actual norm of the observations, we might be able to find $\delta > 0$ such that (ε, δ) -accuracy holds for $\varepsilon < 1$. Of course, (i) we forgot all the other features of B and (ii) we have only analyzed (ε, δ) -accuracy for the choices of \mathbf{w} that maximize the left-hand side of eq. (218). However, this simplistic analysis hints on the fact that entity resolution based on CLKs is not just relevant to private entity resolution, it may also be a good choice for learning in our setting.