The Impact of Record Linkage on Learning from Feature Partitioned Data
— Supplementary Material —

Richard Nock
Google Research, Brain team
richardnock@google.com

Stephen Hardy
Ambiata
stephen.hardy@ambiata.com

Wilko Henecka
Ambiata
wilko.henecka@ambiata.com

Hamish Ivey-Law
The Australian National University
Hamish.Ivey-Law@anu.edu.au

Jakub Nabaglo
The Australian National University
jakub@nab.gl

Giorgio Patrini
Sensity
giorgio@sensity.ai

Guillaume Smith
Ambiata
guillaume.smith@ambiata.com

Brian Thorne
HardByte
brian@hardbyte.nz

Abstract

This is the Supplementary Material to Paper ”The Impact of Record Linkage on Learning from Feature Partitioned Data”, appearing in the proceedings of ICML 2021. To differentiate with the numberings in the main file, the numbering of Theorems is letter-based (A, B, ...).

Table of contents

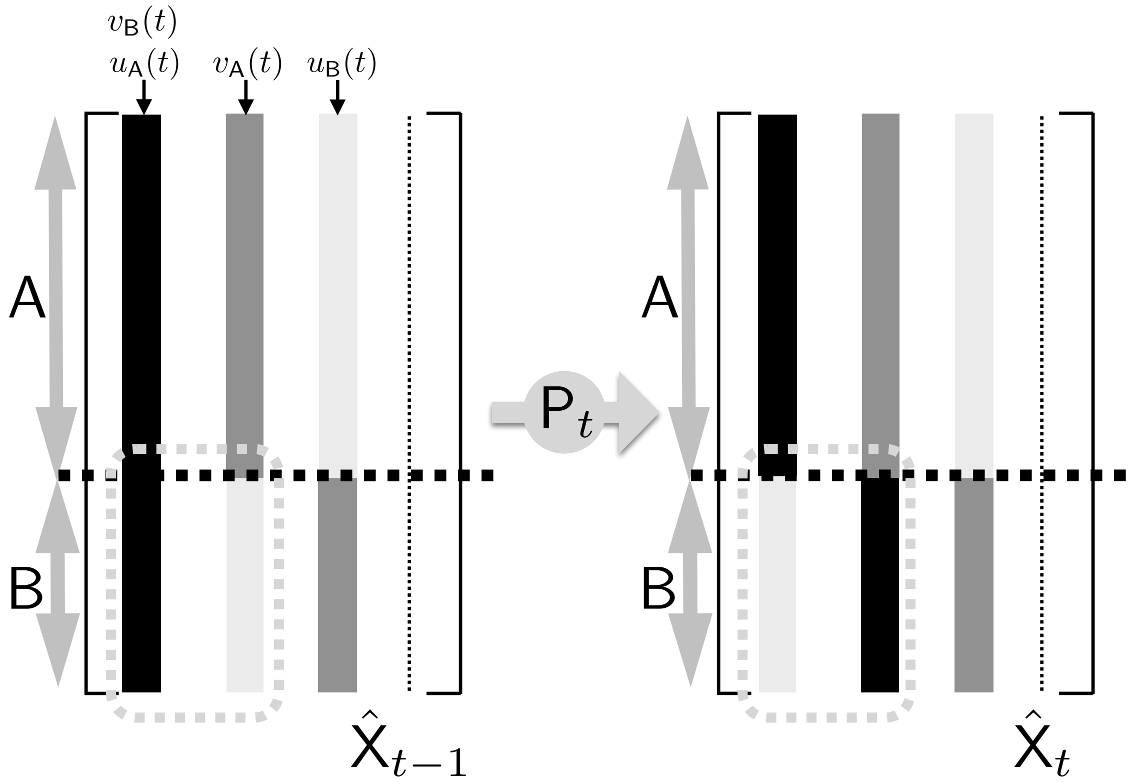


Figure 1: Permutation P_t applied to observation matrix $\hat{\mathsf{X}}_{t-1}$ and subsequent matrix $\hat{\mathsf{X}}_t$, using notations $u_{\mathsf{A}}(t)$, $u_{\mathsf{B}}(t)$, $v_{\mathsf{A}}(t)$ and $v_{\mathsf{B}}(t)$. Different shades represent different observations of X.

I Supplementary material on proofs

I.1 Notations

We need to refine the definitions in Section 2 (main file) to now incorporate iteration number t in the observations affected by P_t. First, because the features of A are not affected by record linkage, we call them *anchor* features. Because the features of B are affected by linkage, we call them *shuffle* features. From now on, we let $u_{\mathsf{B}}(t)$ (resp. $v_{\mathsf{B}}(t)$) denote the indices in $[m]$ of the shuffle features in X that are in observation $u_{\mathsf{A}}(t)$ (resp. $v_{\mathsf{A}}(t)$) and that will be permuted by P_t, creating $\hat{\mathsf{X}}_t$ from $\hat{\mathsf{X}}_{t-1}$. For example, if $u_{\mathsf{B}}(t) = v_{\mathsf{A}}(t), v_{\mathsf{B}}(t) = u_{\mathsf{A}}(t)$, then P_t correctly reconstructs observations in indexes $u_{\mathsf{A}}(t)$ and $v_{\mathsf{A}}(t)$ in X. Figure 1 illustrates the use of these notations. We shall also need to analyse not just classifiers $\tilde{\boldsymbol{\theta}}^*$ and $\boldsymbol{\theta}^*$ (Figure 1 in the main file) but also the sequence of optimal classifiers built over the sequence of datasets affected by the sequence of permutations in $\hat{\mathsf{P}}$. Each $\mathsf{P}_.$ denotes an elementary permutation matrix, and T, the *size* of $\hat{\mathsf{P}}$, is unknown. We thus have the sequence $\hat{\mathsf{X}}_0, \hat{\mathsf{X}}_1, ..., \hat{\mathsf{X}}_T$ constructing $\hat{\mathsf{X}}_T \doteq \hat{\mathsf{X}}$ from $\hat{\mathsf{X}}_0 \doteq \mathsf{X}$. We sometimes write $\hat{\mathsf{X}}_t \doteq [\hat{\boldsymbol{x}}_{t1} \;\; \hat{\boldsymbol{x}}_{t2} \;\; \cdots \;\; \hat{\boldsymbol{x}}_{tm}]$ to denote the column vector decomposition of $\hat{\mathsf{X}}_t$ (with $\hat{\boldsymbol{x}}_{0i} \doteq \boldsymbol{x}_i$) and let $\hat{S}_t$ be the training sample obtained from the t first permutations in the sequence. Hence, $\hat{S}_0 \doteq S$, $\hat{S}_T \doteq \hat{S}$ and $\hat{S}_t \doteq \{(\hat{\boldsymbol{x}}_{ti}, y_i), i \in [m]\}$. We focus on the sequence of minimizers of a given Taylor loss:

$$\boldsymbol{\theta}^*_t \doteq \arg\min_{\boldsymbol{\theta}} \ell_F(\hat{S}_t, \boldsymbol{\theta}; \Gamma) \ . \quad (1)$$

We let $\mu_t \doteq (1/m) \cdot \|\hat{\mathsf{X}}_t\|_F^2 = (1/m) \cdot \|\hat{\mathsf{X}}\|_F^2 = \mu$ (in the main file), where $\|.\|_F$ denotes Frobenius norm. μ_t is just the average square observation norm in $\hat{S}_t$, which remains invariant through permutations. Notice that we have the relationship

$$\begin{aligned}\boldsymbol{\theta}^*_0 &= \boldsymbol{\theta}^*,\\ \boldsymbol{\theta}^*_T &= \tilde{\boldsymbol{\theta}}^*\end{aligned}$$

with respect to notations in the main file (Figure 1).

I.2 Proof of Theorem 3

The proof is obtained in three steps: we first define additional assumptions useful for the proof, then prove a helper Theorem of independent interest, and finally prove Theorem 3.

I.I.2.1 Related notations and additional properties

Definition 1 from the main file now reads more concisely as follows. Suppose that P_t is (ε, τ)-inexact for some $\varepsilon, \tau \geq 0, \varepsilon \leq 1$. Then we have:

$$|(\boldsymbol{x}_{u_{\mathsf{F}}(t)} - \boldsymbol{x}_{v_{\mathsf{F}}(t)})_{\mathsf{F}}^\top \boldsymbol{w}_{\mathsf{F}}| \leq \varepsilon \cdot \max_{i \in \{u_{\mathsf{F}}(t), v_{\mathsf{F}}(t)\}} |\boldsymbol{x}_i^\top \boldsymbol{w}| + \tau \|\boldsymbol{w}\|_2 \ , \forall \mathsf{F} \in \{\mathsf{A}, \mathsf{B}\} \ . \quad (2)$$

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

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.* $\boldsymbol{\mu}_T = \boldsymbol{\mu}_0$ if $\hat{\mathsf{P}}$ factorizes as permutations mixing observations within one or the other class — but not between both. Since the optimal classifier for any Taylor loss is a linear mapping of the mean operator (Lemma C below), our bounds will appear significantly better when $\hat{\mathsf{P}}$ factorizes in such a convenient way.

We now show an additional property of our notations in (main file, Section 2).

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

$$\begin{aligned}(\hat{\boldsymbol{x}}_{tu_{\mathsf{A}}(t)})_{\mathsf{B}} &= (\boldsymbol{x}_{u_{\mathsf{B}}(t)})_{\mathsf{B}} \ , \quad (3)\\ (\hat{\boldsymbol{x}}_{tv_{\mathsf{A}}(t)})_{\mathsf{B}} &= (\boldsymbol{x}_{v_{\mathsf{B}}(t)})_{\mathsf{B}} \ . \quad (4)\end{aligned}$$

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

$$\begin{aligned}u_{\mathsf{B}}(1) &= v_{\mathsf{A}}(1) \ , \quad (5)\\ v_{\mathsf{B}}(1) &= u_{\mathsf{A}}(1) \ . \quad (6)\end{aligned}$$

Thus, it follows:

$$\begin{aligned}
&\mathsf{X}_{\mathsf{A}}\Theta_{u_{\mathsf{A}}(1),v_{\mathsf{A}}(1)}\hat{\mathsf{X}}_{1\mathsf{B}}^{\top} \\
&= (\boldsymbol{x}_{u_{\mathsf{A}}(1)})_{\mathsf{A}}(\boldsymbol{x}_{1v_{\mathsf{A}}(1)})_{\mathsf{B}}^{\top} + (\boldsymbol{x}_{v_{\mathsf{A}}(1)})_{\mathsf{A}}(\boldsymbol{x}_{1u_{\mathsf{A}}(1)})_{\mathsf{B}}^{\top} - (\boldsymbol{x}_{v_{\mathsf{A}}(1)})_{\mathsf{A}}(\boldsymbol{x}_{1v_{\mathsf{A}}(1)})_{\mathsf{B}}^{\top} - (\boldsymbol{x}_{u_{\mathsf{A}}(1)})_{\mathsf{A}}(\boldsymbol{x}_{1u_{\mathsf{A}}(1)})_{\mathsf{B}}^{\top} \\
&= (\boldsymbol{x}_{u_{\mathsf{A}}(1)})_{\mathsf{A}}(\boldsymbol{x}_{v_{\mathsf{B}}(1)})_{\mathsf{B}}^{\top} + (\boldsymbol{x}_{v_{\mathsf{A}}(1)})_{\mathsf{A}}(\boldsymbol{x}_{u_{\mathsf{B}}(1)})_{\mathsf{B}}^{\top} - (\boldsymbol{x}_{v_{\mathsf{A}}(1)})_{\mathsf{A}}(\boldsymbol{x}_{v_{\mathsf{B}}(1)})_{\mathsf{B}}^{\top} - (\boldsymbol{x}_{u_{\mathsf{A}}(1)})_{\mathsf{A}}(\boldsymbol{x}_{u_{\mathsf{B}}(1)})_{\mathsf{B}}^{\top} \qquad (7) \\
&= (\boldsymbol{x}_{u_{\mathsf{A}}(1)})_{\mathsf{A}}(\boldsymbol{x}_{u_{\mathsf{A}}(1)})_{\mathsf{B}}^{\top} + (\boldsymbol{x}_{v_{\mathsf{A}}(1)})_{\mathsf{A}}(\boldsymbol{x}_{v_{\mathsf{A}}(1)})_{\mathsf{B}}^{\top} - (\boldsymbol{x}_{v_{\mathsf{A}}(1)})_{\mathsf{A}}(\boldsymbol{x}_{u_{\mathsf{A}}(1)})_{\mathsf{B}}^{\top} - (\boldsymbol{x}_{u_{\mathsf{A}}(1)})_{\mathsf{A}}(\boldsymbol{x}_{v_{\mathsf{A}}(1)})_{\mathsf{B}}^{\top} \qquad (8) \\
&= (\boldsymbol{x}_{u_{\mathsf{A}}(1)} - \boldsymbol{x}_{v_{\mathsf{A}}(1)})_{\mathsf{A}}(\boldsymbol{x}_{u_{\mathsf{A}}(1)} - \boldsymbol{x}_{v_{\mathsf{A}}(1)})_{\mathsf{B}}^{\top} \ . \qquad (9)
\end{aligned}$$

In eq. (7), we have used eqs (3, 4) and in eq. (8), we have used eqs (5, 6).

Key matrices — The proof of our 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. We first define the matrix we will use most often:

$$\mathrm{v}_t \doteq \left(\mathrm{sign}(c)\cdot\hat{\mathsf{X}}_t\hat{\mathsf{X}}_t^{\top} + \frac{m}{|c|}\cdot\Gamma\right)^{-1} \ , t = 0, 1, ..., T \ , \qquad (10)$$

where c, Γ are the loss parameters in eq. 1. Another matrix u_t, quantifies precisely the local mistake made by each elementary permutation. To define it, we first let (for $t = 1, 2, ..., T$):

$$\begin{aligned}
\boldsymbol{a}_t &\doteq (\boldsymbol{x}_{u_{\mathsf{A}}(t)} - \boldsymbol{x}_{v_{\mathsf{A}}(t)})_{\mathsf{A}} \ , \qquad (11) \\
\boldsymbol{b}_t &\doteq (\boldsymbol{x}_{u_{\mathsf{B}}(t)} - \boldsymbol{x}_{v_{\mathsf{B}}(t)})_{\mathsf{B}} \ . \qquad (12)
\end{aligned}$$

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

$$\begin{aligned}
\boldsymbol{a}_t^{+} &\doteq \left[\begin{array}{c} (\boldsymbol{x}_{u_{\mathsf{A}}(t)} - \boldsymbol{x}_{v_{\mathsf{A}}(t)})_{\mathsf{A}} \\ \hline \boldsymbol{0} \end{array}\right] \in \mathbb{R}^d \ , \qquad (13) \\
\boldsymbol{b}_t^{+} &\doteq \left[\begin{array}{c} \boldsymbol{0} \\ \hline (\boldsymbol{x}_{u_{\mathsf{B}}(t)} - \boldsymbol{x}_{v_{\mathsf{B}}(t)})_{\mathsf{B}} \end{array}\right] \in \mathbb{R}^d \ , \qquad (14)
\end{aligned}$$

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

$$\begin{aligned}
c_{0,t} &\doteq \boldsymbol{a}_t^{+\top}\mathrm{v}_{t-1}\boldsymbol{a}_t^{+} \ , \qquad (15) \\
c_{1,t} &\doteq \boldsymbol{a}_t^{+\top}\mathrm{v}_{t-1}\boldsymbol{b}_t^{+} \ , \qquad (16) \\
c_{2,t} &\doteq \boldsymbol{b}_t^{+\top}\mathrm{v}_{t-1}\boldsymbol{b}_t^{+} \ . \qquad (17)
\end{aligned}$$

We now define u_t as the following block matrix for $t = 1, 2, ..., T$:

$$\mathrm{u}_t \doteq \frac{1}{(1-\mathrm{sign}(c)\cdot c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \left[\begin{array}{c|c} c_{2,t}\cdot\boldsymbol{a}_t\boldsymbol{a}_t^{\top} & (1-\mathrm{sign}(c)\cdot c_{1,t})\cdot\boldsymbol{a}_t\boldsymbol{b}_t^{\top} \\ \hline (1-\mathrm{sign}(c)\cdot c_{1,t})\cdot\boldsymbol{b}_t\boldsymbol{a}_t^{\top} & c_{0,t}\cdot\boldsymbol{b}_t\boldsymbol{b}_t^{\top} \end{array}\right] \qquad (18)$$

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

$$\begin{tikzcd}
\mathrm{U}_t \arrow[dr, dotted] & & \\
& \Lambda_t \arrow[r, dotted] & \mathrm{H}_{i,j} \\
\mathrm{V}_t \arrow[ur] \arrow[dr, dotted] & & \\
& \boldsymbol{\lambda}_t & \\
\boldsymbol{\varepsilon}_t \arrow[ur] & &
\end{tikzcd}$$

Figure 2: Summary of our key notations on matrices and vectors, and dependencies. The dashed arrow means indexes do not match (eq. (22)).

There is one important thing to remark on U_t: it is defined from the indices $u_{\mathsf{A}}(t)$ and $v_{\mathsf{A}}(t)$ in A that are affected by P_t. Hence, U_1 collects the two first such indices (see Figure 1). We also define matrix Λ_t as follows:

$$\Lambda_t \doteq -\frac{b}{|c|} \cdot \mathrm{V}_t \mathrm{U}_{t+1} \ , t = 0, 1, ..., T-1 \ , \tag{19}$$

where parameters b, c are those defined in the Taylor loss after eq. (1). To finish up with matrices, we define doubly indexed matrices that shall be crucial to our proofs, $\mathrm{H}_{i,j}$ for $0 \leq j \leq i \leq T$:

$$\mathrm{H}_{i,j} \doteq \left\{ \begin{array}{ccl} \prod_{k=j}^{i-1}(\mathrm{I}_d + \Lambda_k) & \text{if} & 0 \leq j < i \\ \mathrm{I}_d & \text{if} & j = i \end{array} \right. . \tag{20}$$

Key vectors — we let

$$\boldsymbol{\varepsilon}_t \doteq \boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t \ , t = 0, 1, ..., T-1 \ , \tag{21}$$

which is the difference between two successive mean operators, and

$$\boldsymbol{\lambda}_t \doteq -\frac{b}{|c|} \cdot \mathrm{V}_{t+1} \boldsymbol{\varepsilon}_t \ , t = 0, 1, ..., T-1 \ . \tag{22}$$

Figure 2 summarizes our key notations in this Section. We are now ready to proceed through the proof of our key helper Theorem.

I.I.2.2 Helper Theorem

In this Section, we first show (Theorem E 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:

$$\begin{aligned} \boldsymbol{\theta}^*_{t+1} - \boldsymbol{\theta}^*_t &= \nu \cdot \mathrm{V}_t \mathrm{U}_{t+1} \boldsymbol{\theta}^*_t + \nu \cdot \mathrm{V}_{t+1} \boldsymbol{\varepsilon}_t \\ &= \Lambda_t \boldsymbol{\theta}^*_t + \boldsymbol{\lambda}_t \ , \forall t \geq 0 \ , \end{aligned} \tag{23}$$

where $\Lambda_t, \boldsymbol{\varepsilon}_t, \boldsymbol{\lambda}_t$ are defined in eqs (21, 19, 22) and ν is defined in Lemma C below. This holds regardless of the permutation matrices in the sequence.

We start by the trivial solutions to the minimization of a convex Taylor loss.

Lemma C *The minimum of any convex Taylor loss $\ell_F(\hat{S}_t, \boldsymbol{\theta}; \Gamma)$, is:*

$$\begin{aligned}\boldsymbol{\theta}^*_t &= \nu \cdot \left(\mathrm{sign}(c) \cdot \hat{\mathsf{X}}_t\hat{\mathsf{X}}_t^\top + \nu' \cdot \Gamma\right)^{-1} \boldsymbol{\mu}(\hat{S})_t \\ &= \nu \cdot \mathrm{v}_t \boldsymbol{\mu}_t \ ,\end{aligned} \tag{24}$$

with

$$\begin{aligned}\nu &\doteq -\frac{b}{2|c|} \ , \\ \nu' &\doteq \frac{m}{|c|} \ ,\end{aligned} \tag{25}$$

meeting $\nu, \nu' \neq 0$.

Proof The minimizer $\boldsymbol{\theta}^*_t$ trivially satisfies:

$$2c\hat{\mathsf{X}}_t\hat{\mathsf{X}}_t^\top \boldsymbol{\theta}^*_t + 2m \cdot \Gamma \boldsymbol{\theta}^*_t = -b \cdot \boldsymbol{\mu}_t \ , \tag{26}$$

giving $\boldsymbol{\theta}^*_t = -b \cdot (2c\hat{\mathsf{X}}_t\hat{\mathsf{X}}_t^\top + 2m \cdot \Gamma)^{-1}\boldsymbol{\mu}_t = \nu \cdot \left(\mathrm{sign}(c) \cdot \hat{\mathsf{X}}_t\hat{\mathsf{X}}_t^\top + \nu' \cdot \Gamma\right)^{-1} \boldsymbol{\mu}(\hat{S})_t$, as claimed. ■

Lemma D *Suppose v_{t-1} exists. Then v_t exists if the following holds:*

$$\left\{\begin{array}{rcl}\mathrm{sign}(c) \cdot c_{1,t} & \neq & 1 \ , \\ (1 - \mathrm{sign}(c) \cdot c_{1,t})^2 & \neq & c_{0,t}c_{2,t} \ .\end{array}\right. \tag{27}$$

Proof Throughout the proof, we let

$$\varsigma \doteq \mathrm{sign}(c) \tag{28}$$

for short. We know that $\hat{\mathsf{X}}_t$ is obtained from $\hat{\mathsf{X}}_{t-1}$ after permuting the shuffle part of observations at indexes $u_{\mathsf{A}}(t)$ and $v_{\mathsf{A}}(t)$ in $\hat{\mathsf{X}}_{(t-1)\mathsf{B}}$ by P_t (see Figure 1). So,

$$\begin{aligned}\hat{\mathsf{X}}_{t\mathsf{B}} &= \hat{\mathsf{X}}_{(t-1)\mathsf{B}} + \hat{\mathsf{X}}_{(t-1)\mathsf{B}}(\mathsf{P}_t - \mathrm{I}_n) \\ &= \hat{\mathsf{X}}_{(t-1)\mathsf{B}} + \hat{\mathsf{X}}_{(t-1)\mathsf{B}}(\mathbf{1}_{u_{\mathsf{A}}(t)}\mathbf{1}_{v_{\mathsf{A}}(t)}^\top + \mathbf{1}_{v_{\mathsf{A}}(t)}\mathbf{1}_{u_{\mathsf{A}}(t)}^\top - \mathbf{1}_{v_{\mathsf{A}}(t)}\mathbf{1}_{v_{\mathsf{A}}(t)}^\top - \mathbf{1}_{u_{\mathsf{A}}(t)}\mathbf{1}_{u_{\mathsf{A}}(t)}^\top) \ ,\end{aligned} \tag{29}$$

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

$$\begin{aligned}\hat{\mathsf{X}}_t\hat{\mathsf{X}}_t^\top &= \left[\begin{array}{c|c}\mathsf{X}_\mathsf{A}\mathsf{X}_\mathsf{A}^\top & \mathsf{X}_\mathsf{A}\hat{\mathsf{X}}_{t\mathsf{B}}^\top \\ \hline \hat{\mathsf{X}}_{t\mathsf{B}}\mathsf{X}_\mathsf{A}^\top & \hat{\mathsf{X}}_{t\mathsf{B}}\hat{\mathsf{X}}_{t\mathsf{B}}^\top\end{array}\right] \\ &= \left[\begin{array}{c|c}\mathsf{X}_\mathsf{A}\mathsf{X}_\mathsf{A}^\top & \mathsf{X}_\mathsf{A}\hat{\mathsf{X}}_{t\mathsf{B}}^\top \\ \hline \hat{\mathsf{X}}_{t\mathsf{B}}\mathsf{X}_\mathsf{A}^\top & \hat{\mathsf{X}}_{(t-1)\mathsf{B}}\mathsf{P}_t\mathsf{P}_t^\top\hat{\mathsf{X}}_{(t-1)\mathsf{B}}^\top\end{array}\right] \\ &= \left[\begin{array}{c|c}\mathsf{X}_\mathsf{A}\mathsf{X}_\mathsf{A}^\top & \mathsf{X}_\mathsf{A}\hat{\mathsf{X}}_{t\mathsf{B}}^\top \\ \hline \hat{\mathsf{X}}_{t\mathsf{B}}\mathsf{X}_\mathsf{A}^\top & \hat{\mathsf{X}}_{(t-1)\mathsf{B}}\hat{\mathsf{X}}_{(t-1)\mathsf{B}}^\top\end{array}\right] \ ,\end{aligned} \tag{30}$$

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,

$$\begin{aligned}
\hat{\mathsf{X}}_t\hat{\mathsf{X}}_t^\top &= \hat{\mathsf{X}}_{t-1}\hat{\mathsf{X}}_{t-1}^\top + \left[\begin{array}{c|c} 0 & \mathsf{X}_\mathsf{A}(\hat{\mathsf{X}}_{t\mathsf{B}} - \mathsf{X}_{(t-1)\mathsf{B}})^\top \\ \hline (\hat{\mathsf{X}}_{t\mathsf{B}} - \mathsf{X}_{(t-1)\mathsf{B}})\mathsf{X}_\mathsf{A}^\top & 0 \end{array}\right] \\
&= \hat{\mathsf{X}}_{t-1}\hat{\mathsf{X}}_{t-1}^\top + \left[\begin{array}{c|c} 0 & \mathsf{X}_\mathsf{A}\Theta_{u_\mathsf{A}(t),v_\mathsf{A}(t)}\hat{\mathsf{X}}_{(t-1)\mathsf{B}}^\top \\ \hline \hat{\mathsf{X}}_{(t-1)\mathsf{B}}\Theta_{u_\mathsf{A}(t),v_\mathsf{A}(t)}\mathsf{X}_\mathsf{A}^\top & 0 \end{array}\right] \ , \qquad (31)
\end{aligned}$$

with $\Theta_{u_\mathsf{A}(t),v_\mathsf{A}(t)} \doteq \mathbf{1}_{u_\mathsf{A}(t)}\mathbf{1}_{v_\mathsf{A}(t)}^\top + \mathbf{1}_{v_\mathsf{A}(t)}\mathbf{1}_{u_\mathsf{A}(t)}^\top - \mathbf{1}_{v_\mathsf{A}(t)}\mathbf{1}_{v_\mathsf{A}(t)}^\top - \mathbf{1}_{u_\mathsf{A}(t)}\mathbf{1}_{u_\mathsf{A}(t)}^\top$ (symmetric, see eq. (29) and example 1). Now, remark that

$$\begin{aligned}
&\mathsf{X}_\mathsf{A}\Theta_{u_\mathsf{A}(t),v_\mathsf{A}(t)}\hat{\mathsf{X}}_{(t-1)\mathsf{B}}^\top \\
&= \mathsf{X}_\mathsf{A}(\mathbf{1}_{u_\mathsf{A}(t)}\mathbf{1}_{v_\mathsf{A}(t)}^\top + \mathbf{1}_{v_\mathsf{A}(t)}\mathbf{1}_{u_\mathsf{A}(t)}^\top - \mathbf{1}_{v_\mathsf{A}(t)}\mathbf{1}_{v_\mathsf{A}(t)}^\top - \mathbf{1}_{u_\mathsf{A}(t)}\mathbf{1}_{u_\mathsf{A}(t)}^\top)\hat{\mathsf{X}}_{t\mathsf{B}}^\top \\
&= (\boldsymbol{x}_{u_\mathsf{A}(t)})_\mathsf{A}(\boldsymbol{x}_{tv_\mathsf{A}(t)})_\mathsf{B}^\top + (\boldsymbol{x}_{v_\mathsf{A}(t)})_\mathsf{A}(\boldsymbol{x}_{tu_\mathsf{A}(t)})_\mathsf{B}^\top - (\boldsymbol{x}_{v_\mathsf{A}(t)})_\mathsf{A}(\boldsymbol{x}_{tv_\mathsf{A}(t)})_\mathsf{B}^\top - (\boldsymbol{x}_{u_\mathsf{A}(t)})_\mathsf{A}(\boldsymbol{x}_{tu_\mathsf{A}(t)})_\mathsf{B}^\top \\
&= (\boldsymbol{x}_{u_\mathsf{A}(t)})_\mathsf{A}(\boldsymbol{x}_{v_\mathsf{B}(t)})_\mathsf{B}^\top + (\boldsymbol{x}_{v_\mathsf{A}(t)})_\mathsf{A}(\boldsymbol{x}_{u_\mathsf{B}(t)})_\mathsf{B}^\top - (\boldsymbol{x}_{v_\mathsf{A}(t)})_\mathsf{A}(\boldsymbol{x}_{v_\mathsf{B}(t)})_\mathsf{B}^\top - (\boldsymbol{x}_{u_\mathsf{A}(t)})_\mathsf{A}(\boldsymbol{x}_{u_\mathsf{B}(t)})_\mathsf{B}^\top \qquad (32) \\
&= -((\boldsymbol{x}_{u_\mathsf{A}(t)})_\mathsf{A} - (\boldsymbol{x}_{v_\mathsf{A}(t)})_\mathsf{A})((\boldsymbol{x}_{u_\mathsf{B}(t)})_\mathsf{B} - (\boldsymbol{x}_{v_\mathsf{B}(t)})_\mathsf{B})^\top \\
&= -(\boldsymbol{x}_{u_\mathsf{A}(t)} - \boldsymbol{x}_{v_\mathsf{A}(t)})_\mathsf{A}(\boldsymbol{x}_{u_\mathsf{B}(t)} - \boldsymbol{x}_{v_\mathsf{B}(t)})_\mathsf{B}^\top = -\boldsymbol{a}_t\boldsymbol{b}_t^\top \ . \qquad (33)
\end{aligned}$$

Eq. (32) holds because of Lemma B. We finally get

$$\hat{\mathsf{X}}_t\hat{\mathsf{X}}_t^\top = \hat{\mathsf{X}}_{t-1}\hat{\mathsf{X}}_{t-1}^\top - \varsigma \cdot \boldsymbol{a}_t^+\boldsymbol{b}_t^{+\top} - \varsigma \cdot \boldsymbol{b}_t^+\boldsymbol{a}_t^{+\top} \ , \qquad (34)$$

and so we have

$$\mathrm{v}_t = \left(\mathrm{v}_{t-1}^{-1} - \varsigma \cdot \boldsymbol{a}_t^+\boldsymbol{b}_t^{+\top} - \varsigma \cdot \boldsymbol{b}_t^+\boldsymbol{a}_t^{+\top}\right)^{-1} \ . \qquad (35)$$

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

$$\begin{aligned}
\det(\mathrm{v}_{t-1}^{-1} - \varsigma \cdot \boldsymbol{a}_t^+\boldsymbol{b}_t^{+\top}) &= \det(\mathrm{v}_{t-1}^{-1})\det(\mathrm{I}_d - \varsigma \cdot \mathrm{v}_{t-1}\boldsymbol{a}_t^+\boldsymbol{b}_t^{+\top}) \\
&= \det(\mathrm{v}_{t-1}^{-1})(1 - \varsigma \cdot \boldsymbol{b}_t^{+\top}\mathrm{v}_{t-1}\boldsymbol{a}_t^+) \\
&= \det(\mathrm{v}_{t-1}^{-1})(1 - \varsigma \cdot c_{1,t}) \ , \qquad (36)
\end{aligned}$$

where the penultimate identity comes from Sylvester's determinant formula. So, if in addition

$1 - \varsigma \cdot c_{1,t} \neq 0$, then

$$\begin{aligned}
&\det\left(\mathrm{v}_{t-1}^{-1} - \varsigma \cdot \boldsymbol{a}_t^+ \boldsymbol{b}_t^{+\top} - \varsigma \cdot \boldsymbol{b}_t^+ \boldsymbol{a}_t^{+\top}\right) \\
&\quad = \det(\mathrm{v}_{t-1}^{-1} - \varsigma \cdot \boldsymbol{a}_t^+ \boldsymbol{b}_t^{+\top})\det\left(\mathrm{I}_d - \varsigma \cdot \left(\mathrm{v}_{t-1}^{-1} - \varsigma \cdot \boldsymbol{a}_t^+ \boldsymbol{b}_t^{+\top}\right) \boldsymbol{b}_t^+ \boldsymbol{a}_t^{+\top}\right) \\
&\quad = \det(\mathrm{v}_{t-1}^{-1})(1 - \varsigma \cdot c_{1,t})\det\left(\mathrm{I}_d - \varsigma \cdot \left(\mathrm{v}_{t-1}^{-1} - \varsigma \cdot \boldsymbol{a}_t^+ \boldsymbol{b}_t^{+\top}\right)^{-1} \boldsymbol{b}_t^+ \boldsymbol{a}_t^{+\top}\right) && (37) \\
&\quad = \det(\mathrm{v}_{t-1}^{-1})(1 - \varsigma \cdot c_{1,t})\left(1 - \varsigma \cdot \boldsymbol{a}_t^{+\top} \left(\mathrm{v}_{t-1}^{-1} - \varsigma \cdot \boldsymbol{a}_t^+ \boldsymbol{b}_t^{+\top}\right)^{-1} \boldsymbol{b}_t^+\right) && (38) \\
&\quad = \det(\mathrm{v}_{t-1}^{-1})(1 - \varsigma \cdot c_{1,t})\left(1 - \varsigma \cdot \boldsymbol{a}_t^{+\top} \left(\mathrm{v}_{t-1} + \frac{\varsigma}{1 - \varsigma \cdot \boldsymbol{b}_t^{+\top} \mathrm{v}_{t-1} \boldsymbol{a}_t^+} \cdot \mathrm{v}_{t-1} \boldsymbol{a}_t^+ \boldsymbol{b}_t^{+\top} \mathrm{v}_{t-1}\right) \boldsymbol{b}_t^+\right) && (39) \\
&\quad = \det(\mathrm{v}_{t-1}^{-1})(1 - \varsigma \cdot c_{1,t})\left(1 - \varsigma \cdot c_{1,t} - \varsigma^2 \cdot \frac{c_{0,t}c_{2,t}}{1 - \varsigma \cdot c_{1,t}}\right) \\
&\quad = \frac{(1 - \varsigma \cdot c_{1,t})^2 - c_{0,t}c_{2,t}}{\det(\mathrm{v}_{t-1})} \ . && (40)
\end{aligned}$$

Here, eq. (37) comes from eq. (36). Eq. (38) is another application of Sylvester's determinant formula. Eq. (38) is Sherman-Morrison formula and the last equation uses the fact that $\varsigma^2 = 1$. We immediately conclude on Lemma D. ■

If we now assume without loss of generality that v_0 exists — which boils down to taking $\Gamma \succ 0$ —, then we get the existence of the complete sequence of matrices v_t (and thus the existence of the sequence of optimal classifiers $\boldsymbol{\theta}_0^*, \boldsymbol{\theta}_1^*, ...$) provided the following **invertibility** condition is satisfied.

(**invertibility**) For any $t \geq 1$, $(1 - \mathrm{sign}(c) \cdot c_{1,t})^2 \not\in \{0, c_{0,t}c_{2,t}\}$.

We shall check later (Corollary K) that the invertibility condition indeed holds in our setting.

Theorem E *Suppose the invertibility assumption holds. Then we have:*

$$\frac{1}{\nu} \cdot (\boldsymbol{\theta}_{t+1}^* - \boldsymbol{\theta}_t^*) \ = \ \mathrm{v}_t \mathrm{u}_{t+1} \boldsymbol{\theta}_t^* + \mathrm{v}_{t+1} \boldsymbol{\varepsilon}_t \ , \forall t \geq 0 \ ,$$

where $\boldsymbol{\varepsilon}_t$ is defined in eq. (21).

Proof Throughout the proof, we let

$$\varsigma \ \doteq \ \mathrm{sign}(c) \tag{41}$$

for short. We have from Lemma C, for any $t \geq 1$,

$$\begin{aligned}
\frac{1}{\nu} \cdot (\boldsymbol{\theta}_t^* - \boldsymbol{\theta}_{t-1}^*) &= \mathrm{v}_t \boldsymbol{\mu}_t - \mathrm{v}_{t-1} \boldsymbol{\mu}_{t-1} \\
&= \Delta_{t-1} \boldsymbol{\mu}_{t-1} + \mathrm{v}_t \boldsymbol{\varepsilon}_{t-1} \ , && (42)
\end{aligned}$$

with $\Delta_t \doteq \mathrm{v}_{t+1} - \mathrm{v}_t$. It comes from eq. (34),

$$\Delta_{t-1} = \left(\hat{\mathsf{X}}_{t-1}\hat{\mathsf{X}}_{t-1}^\top + \nu' \cdot \Gamma - \varsigma \cdot \boldsymbol{a}_t^+ \boldsymbol{b}_t^{+\top} - \varsigma \cdot \boldsymbol{b}_t^+ \boldsymbol{a}_t^{+\top}\right)^{-1} - \mathrm{v}_t \ . \quad (43)$$

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

$$\begin{aligned} & \left(\hat{\mathsf{X}}_{t-1}\hat{\mathsf{X}}_{t-1}^\top + \nu' \cdot \Gamma - \varsigma \cdot \boldsymbol{a}_t^+ \boldsymbol{b}_t^{+\top} - \varsigma \cdot \boldsymbol{b}_t^+ \boldsymbol{a}_t^{+\top}\right)^{-1} \\ & \quad = \left(\hat{\mathsf{X}}_{t-1}\hat{\mathsf{X}}_{t-1}^\top + \nu' \cdot \Gamma - \varsigma \cdot \boldsymbol{a}_t^+ \boldsymbol{b}_t^{+\top}\right)^{-1} \\ & \quad\quad + \frac{\varsigma}{1 - \varsigma \cdot \boldsymbol{a}_t^{+\top} \left(\hat{\mathsf{X}}_{t-1}\hat{\mathsf{X}}_{t-1}^\top + \nu' \cdot \Gamma - \varsigma \cdot \boldsymbol{a}_t^+ \boldsymbol{b}_t^{+\top}\right)^{-1} \boldsymbol{b}_t^+} \cdot \mathsf{Q}_t \ , \end{aligned} \quad (44)$$

with

$$\mathsf{Q}_t \doteq \left(\hat{\mathsf{X}}_{t-1}\hat{\mathsf{X}}_{t-1}^\top + \nu' \cdot \Gamma - \varsigma \cdot \boldsymbol{a}_t^+ \boldsymbol{b}_t^{+\top}\right)^{-1} \boldsymbol{b}_t^+ \boldsymbol{a}_t^{+\top} \left(\hat{\mathsf{X}}_{t-1}\hat{\mathsf{X}}_{t-1}^\top + \nu' \cdot \Gamma - \varsigma \cdot \boldsymbol{a}_t^+ \boldsymbol{b}_t^{+\top}\right)^{-1} \ , \quad (45)$$

and

$$\left(\hat{\mathsf{X}}_{t-1}\hat{\mathsf{X}}_{t-1}^\top + \nu' \cdot \Gamma - \varsigma \cdot \boldsymbol{a}_t^+ \boldsymbol{b}_t^{+\top}\right)^{-1} = \mathrm{v}_{t-1} + \frac{\varsigma}{1 - \varsigma \cdot \boldsymbol{b}_t^{+\top} \mathrm{v}_{t-1} \boldsymbol{a}_t^+} \cdot \mathrm{v}_{t-1} \boldsymbol{a}_t^+ \boldsymbol{b}_t^{+\top} \mathrm{v}_{t-1} \ . (46)$$

Let us define the following shorthand:

$$\Sigma_t \doteq \mathrm{v}_{t-1} + \frac{\varsigma}{1 - \varsigma \cdot \boldsymbol{b}_t^{+\top} \mathrm{v}_{t-1} \boldsymbol{a}_t^+} \cdot \mathrm{v}_{t-1} \boldsymbol{a}_t^+ \boldsymbol{b}_t^{+\top} \mathrm{v}_{t-1} \ . \quad (47)$$

Then, plugging together eqs. (44) and (46), we get:

$$\begin{aligned}
&\left(\hat{\mathsf{X}}_{t-1}\hat{\mathsf{X}}_{t-1}^{\top} + \nu' \cdot \Gamma - \varsigma \cdot \boldsymbol{a}_t^{+}\boldsymbol{b}_t^{+\top} - \varsigma \cdot \boldsymbol{b}_t^{+}\boldsymbol{a}_t^{+\top}\right)^{-1}\\
= \quad & \mathrm{V}_{t-1} + \frac{\varsigma}{1-\varsigma\cdot \boldsymbol{b}_t^{+\top}\mathrm{V}_{t-1}\boldsymbol{a}_t^{+}} \cdot \mathrm{V}_{t-1}\boldsymbol{a}_t^{+}\boldsymbol{b}_t^{+\top}\mathrm{V}_{t-1}\\
& + \frac{\varsigma}{1-\varsigma\cdot \boldsymbol{a}_t^{+\top}\mathrm{V}_{t-1}\boldsymbol{b}_t^{+} - \frac{\boldsymbol{a}_t^{+\top}\mathrm{V}_{t-1}\boldsymbol{a}_t^{+}\cdot\boldsymbol{b}_t^{+\top}\mathrm{V}_{t-1}\boldsymbol{b}_t^{+}}{1-\varsigma\cdot\boldsymbol{b}_t^{+\top}\mathrm{V}_{t-1}\boldsymbol{a}_t^{+}}} \cdot \Sigma_t\boldsymbol{b}_t^{+}\boldsymbol{a}_t^{+\top}\Sigma_t\\
= \quad & \mathrm{V}_{t-1} + \frac{\varsigma}{1-\varsigma\cdot c_{1,t}} \cdot \mathrm{V}_{t-1}\boldsymbol{a}_t^{+}\boldsymbol{b}_t^{+\top}\mathrm{V}_{t-1}\\
& + \frac{\varsigma}{1-\varsigma\cdot c_{1,t} - \frac{c_{0,t}c_{2,t}}{1-\varsigma\cdot c_{1,t}}} \cdot \left(\begin{array}{c}\mathrm{V}_{t-1}\\ +\\ \frac{\varsigma}{1-\varsigma\cdot c_{1,t}}\cdot \mathrm{V}_{t-1}\boldsymbol{a}_t^{+}\boldsymbol{b}_t^{+\top}\mathrm{V}_{t-1}\end{array}\right)\boldsymbol{b}_t^{+}\boldsymbol{a}_t^{+\top}\left(\begin{array}{c}\mathrm{V}_{t-1}\\ +\\ \frac{\varsigma}{1-\varsigma\cdot c_{1,t}}\cdot \mathrm{V}_{t-1}\boldsymbol{a}_t^{+}\boldsymbol{b}_t^{+\top}\mathrm{V}_{t-1}\end{array}\right)\\
= \quad & \mathrm{V}_{t-1} + \frac{\varsigma}{1-\varsigma\cdot c_{1,t}} \cdot \mathrm{V}_{t-1}\boldsymbol{a}_t^{+}\boldsymbol{b}_t^{+\top}\mathrm{V}_{t-1} + \frac{\varsigma\cdot}{1-\varsigma\cdot c_{1,t} - \frac{c_{0,t}c_{2,t}}{1-\varsigma\cdot c_{1,t}}} \cdot \mathrm{V}_{t-1}\boldsymbol{b}_t^{+}\boldsymbol{a}_t^{+\top}\mathrm{V}_{t-1}\\
& + \frac{c_{0,t}}{(1-\varsigma\cdot c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \mathrm{V}_{t-1}\boldsymbol{b}_t^{+}\boldsymbol{b}_t^{+\top}\mathrm{V}_{t-1} + \frac{c_{2,t}}{(1-\varsigma\cdot c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \mathrm{V}_{t-1}\boldsymbol{a}_t^{+}\boldsymbol{a}_t^{+\top}\mathrm{V}_{t-1}\\
& + \frac{\varsigma c_{0,t}c_{2,t}}{(1-\varsigma\cdot c_{1,t})((1-\varsigma\cdot c_{1,t})^2 - c_{0,t}c_{2,t})} \cdot \mathrm{V}_{t-1}\boldsymbol{a}_t^{+}\boldsymbol{b}_t^{+\top}\mathrm{V}_{t-1}\\
= \quad & \mathrm{V}_{t-1} + \frac{1-\varsigma\cdot c_{1,t}}{(1-\varsigma\cdot c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \left(\mathrm{V}_{t-1}\boldsymbol{a}_t^{+}\boldsymbol{b}_t^{+\top}\mathrm{V}_{t-1} + \mathrm{V}_{t-1}\boldsymbol{b}_t^{+}\boldsymbol{a}_t^{+\top}\mathrm{V}_{t-1}\right)\\
& + \frac{c_{0,t}}{(1-\varsigma\cdot c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \mathrm{V}_{t-1}\boldsymbol{b}_t^{+}\boldsymbol{b}_t^{+\top}\mathrm{V}_{t-1} + \frac{c_{2,t}}{(1-\varsigma\cdot c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \mathrm{V}_{t-1}\boldsymbol{a}_t^{+}\boldsymbol{a}_t^{+\top}\mathrm{V}_{t-1}\\
= \quad & \mathrm{V}_{t-1} + \frac{1}{(1-\varsigma\cdot c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \left\{\begin{array}{c}(1-\varsigma\cdot c_{1,t})\cdot(\mathrm{V}_{t-1}\boldsymbol{a}_t^{+}\boldsymbol{b}_t^{+\top}\mathrm{V}_{t-1} + \mathrm{V}_{t-1}\boldsymbol{b}_t^{+}\boldsymbol{a}_t^{+\top}\mathrm{V}_{t-1})\\ +c_{0,t}\cdot\mathrm{V}_{t-1}\boldsymbol{b}_t^{+}\boldsymbol{b}_t^{+\top}\mathrm{V}_{t-1}\\ +c_{2,t}\cdot\mathrm{V}_{t-1}\boldsymbol{a}_t^{+}\boldsymbol{a}_t^{+\top}\mathrm{V}_{t-1}\end{array}\right\}\\
= \quad & \mathrm{V}_{t-1} + \mathrm{V}_{t-1}\mathrm{U}_t\mathrm{V}_{t-1} \quad .
\end{aligned}\tag{48}$$

So,

$$\begin{aligned}
\frac{1}{\nu}\cdot(\boldsymbol{\theta}_t^* - \boldsymbol{\theta}_{t-1}^*) &= \Delta_{t-1}\boldsymbol{\mu}_{t-1} + \mathrm{V}_t\boldsymbol{\varepsilon}_{t-1}\\
&= \mathrm{V}_{t-1}\mathrm{U}_t\mathrm{V}_{t-1}\boldsymbol{\mu}_{t-1} + \mathrm{V}_t\boldsymbol{\varepsilon}_{t-1}\\
&= \mathrm{V}_{t-1}\mathrm{U}_t\boldsymbol{\theta}_{t-1}^* + \mathrm{V}_t\boldsymbol{\varepsilon}_{t-1} \quad ,
\end{aligned}\tag{49}$$

as claimed (end of the proof of Theorem E). ■

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

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

$$\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^* = (\mathrm{H}_{T,0} - \mathrm{I}_d)\boldsymbol{\theta}_0^* + \sum_{t=0}^{T-1}\mathrm{H}_{T,t+1}\boldsymbol{\lambda}_t \quad . \tag{50}$$

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

$$\boldsymbol{\theta}^*_{t+1} \quad = \quad (\mathrm{I}_d + \Lambda_t)\boldsymbol{\theta}^*_t + \boldsymbol{\lambda}_t \ . \tag{51}$$

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

$$\begin{aligned}
\boldsymbol{\theta}^*_T &= \prod_{t=0}^{T-1} (\mathrm{I}_d + \Lambda_t)\boldsymbol{\theta}^*_0 + \boldsymbol{\lambda}_{T-1} + \sum_{j=0}^{T-2} \prod_{t=j+1}^{T-1} (\mathrm{I}_d + \Lambda_t)\boldsymbol{\lambda}_j \\
&= \mathrm{H}_{T,0}\boldsymbol{\theta}^*_0 + \sum_{t=0}^{T-1} \mathrm{H}_{T,t+1}\boldsymbol{\lambda}_t \ , \qquad (52)
\end{aligned}$$

which yields the statement of Theorem F. ■

Since it applies to every permutation matrix, Theorem F applies to *every* record linkage algorithm. Theorem F gives us a 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. We apply it now to derive one such bound.

I.I.2.3 Finalizing the proof of Theorem 3

Assuming it is diagonalizable, we let $\lambda^*(\mathrm{M})$ (resp. $\lambda^\circ(\mathrm{M})$) denotes the maximal (resp. minimal) eigenvalue of M.

Lemma G *Suppose $\lambda^\circ(\Gamma) > \max\{0, -c\mu_t\}$. Then V_t is positive definite and its eigenspectrum satisfies:*

$$\begin{aligned}
\lambda^*(\mathrm{V}_t) &\leq \frac{1}{m} \cdot \frac{1}{\mathrm{sign}(c)\mu_t + \frac{1}{|c|}\lambda^\circ(\Gamma)} \ , \qquad (53) \\
\lambda^\circ(\mathrm{V}_t) &\geq \frac{1}{m} \cdot \frac{1}{\mathrm{sign}(c)\mu_t + \frac{1}{|c|}\lambda^*(\Gamma)} \ , \qquad (54)
\end{aligned}$$

where c is defined in Lemma C and $\mu_t \doteq (1/m) \cdot \|\hat{\mathsf{X}}_t\|_F^2$, where $\|.\|_F$ denotes Frobenius norm.

Proof We have $\|\hat{\mathsf{X}}_t^\top \boldsymbol{w}\|_2^2 \leq \|\boldsymbol{w}\|_2^2 \|\hat{\mathsf{X}}_t\|_F^2$. We get:

$$\begin{aligned}
\lambda^*(\mathrm{V}_t) &\doteq \left(\inf_{\boldsymbol{w}} \frac{\boldsymbol{w}^\top \left(\mathrm{sign}(c)\hat{\mathsf{X}}_t\hat{\mathsf{X}}_t^\top + \nu' \cdot \Gamma \right) \boldsymbol{w}}{\|\boldsymbol{w}\|_2^2} \right)^{-1} \\
&\leq \frac{1}{\mathrm{sign}(c) m \mu_t + \nu' \lambda^\circ(\Gamma)} \\
&= \frac{1}{m} \cdot \frac{1}{\mathrm{sign}(c)\mu_t + \frac{1}{|c|}\lambda^\circ(\Gamma)} \ , \qquad (55)
\end{aligned}$$

assuming $\mathrm{sign}(c)\mu_t + \frac{1}{|c|}\lambda^\circ(\Gamma) > 0$, which trivially holds if $c \geq 0$, and implies otherwise $\lambda^\circ(\Gamma) >$

$-c\mu_t$. Similarly,

$$\begin{aligned}
\lambda^\circ(\mathrm{v}_t) &\doteq \left(\sup_{\boldsymbol{w}} \frac{\boldsymbol{w}^\top \left(\mathrm{sign}(c)\hat{\mathsf{X}}_t\hat{\mathsf{X}}_t^\top + \nu' \cdot \Gamma\right)\boldsymbol{w}}{\|\boldsymbol{w}\|_2^2}\right)^{-1} \\
&\geq \frac{1}{\mathrm{sign}(c) m\mu_t + \nu'\lambda^*(\Gamma)} \\
&= \frac{1}{m} \cdot \frac{1}{\mathrm{sign}(c)\mu_t + \frac{1}{|c|}\lambda^*(\Gamma)} \ , \qquad (56)
\end{aligned}$$

as claimed. This ends the proof of Lemma G. ■

Lemma H *Using notations of Subsection I.1.2.1, suppose* $(1-c_{1,t})^2 - c_{0,t}c_{2,t} \neq 0$[1] *and* $\boldsymbol{a}_t \neq \boldsymbol{0}$. *Then* u_t *is negative semi-definite iff* $(1-c_{1,t})^2 - c_{0,t}c_{2,t} < 0$. *Otherwise,* u_t *is indefinite. In all cases,* u_t *is diagonalizable and for any* $z \in \{\lambda^*(\mathrm{u}_t), |\lambda^\circ(\mathrm{u}_t)|\}$, *we have*

$$z \leq \frac{2+3(c_{0,t}+c_{2,t})}{2|(1-c_{1,t})^2 - c_{0,t}c_{2,t}|} \cdot \max\{\|\boldsymbol{a}_t\|_2^2, \|\boldsymbol{b}_t\|_2^2\} \ . \qquad (57)$$

Proof Consider a block-vector following the column-block partition of u_t,

$$\tilde{\boldsymbol{x}} \doteq \left[\begin{array}{c} \boldsymbol{x} \\ \hline \boldsymbol{y} \end{array}\right] \ . \qquad (58)$$

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

$$\mathrm{u}_t\tilde{\boldsymbol{x}} = \frac{1}{\zeta} \cdot \left[\begin{array}{c} (c_{2,t}(\boldsymbol{a}_t^\top\boldsymbol{x}) + (1-c_{1,t})(\boldsymbol{b}_t^\top\boldsymbol{y})) \cdot \boldsymbol{a}_t \\ \hline ((1-c_{1,t})(\boldsymbol{a}_t^\top\boldsymbol{x}) + c_{0,t}(\boldsymbol{b}_t^\top\boldsymbol{y})) \cdot \boldsymbol{b}_t \end{array}\right] \ . \qquad (59)$$

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

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

$$\begin{aligned}
\mathrm{u}_t \left[\begin{array}{c} \boldsymbol{a}_t \\ \hline \boldsymbol{0} \end{array}\right] &= -\frac{1}{c_{0,t}c_{2,t}} \cdot \left[\begin{array}{c|c} c_{2,t} \cdot \boldsymbol{a}_t\boldsymbol{a}_t^\top & 0 \\ \hline 0 & c_{0,t} \cdot \boldsymbol{b}_t\boldsymbol{b}_t^\top \end{array}\right] \left[\begin{array}{c} \boldsymbol{a} \\ \hline \boldsymbol{0} \end{array}\right] \\
&= -\frac{1}{\lambda(\boldsymbol{a}_t^+)} \cdot \left[\begin{array}{c} \boldsymbol{a}_t \\ \hline \boldsymbol{0} \end{array}\right] \ , \qquad (60)
\end{aligned}$$

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

$$\lambda(\boldsymbol{a}_t^+) \doteq \frac{{\boldsymbol{a}_t^+}^\top \mathrm{v}_{t-1}\boldsymbol{a}_t^+}{\|\boldsymbol{a}_t^+\|_2^2} \ , \qquad (61)$$

[1] This is implied by the invertibility assumption.

and

$$\mathrm{U}_t \left[\begin{array}{c}\mathbf{0}\\ \hline \boldsymbol{b}_t\end{array}\right] = -\frac{1}{\lambda(\boldsymbol{b}_t^+)} \cdot \left[\begin{array}{c}\mathbf{0}\\ \hline \boldsymbol{b}_t\end{array}\right] \ , \lambda(\boldsymbol{b}_t^+) \doteq \frac{\boldsymbol{b}_t^{+\top}\mathrm{V}_{t-1}\boldsymbol{b}_t^+}{\|\boldsymbol{b}_t^+\|_2^2} \ . \tag{62}$$

We also remark that U_t is negative semi-definite. We finally remark that U_t is diagonalizable: both non-zero eigenvalues have multiplicity 1 and eigenvalue 0 is associated to eigen subspace complementary of the span of the two eigenvectors.

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

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

In this case, we obtain

$$\begin{aligned}\mathrm{U}_t\tilde{\boldsymbol{x}} &= \frac{(1-c_{1,t})(\boldsymbol{a}_t^\top\boldsymbol{x}) + c_{0,t}(\boldsymbol{b}_t^\top\boldsymbol{y})}{(1-c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \left[\begin{array}{c}\frac{c_{2,t}(\boldsymbol{a}_t^\top\boldsymbol{x})+(1-c_{1,t})(\boldsymbol{b}_t^\top\boldsymbol{y})}{(1-c_{1,t})(\boldsymbol{a}_t^\top\boldsymbol{x})+c_{0,t}(\boldsymbol{b}_t^\top\boldsymbol{y})} \cdot \boldsymbol{a}_t\\ \hline \boldsymbol{b}_t\end{array}\right]\\ &= \frac{\alpha(1-c_{1,t})\|\boldsymbol{a}_t\|_2^2 + c_{0,t}\|\boldsymbol{b}_t\|_2^2}{(1-c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot \left[\begin{array}{c}\frac{\alpha c_{2,t}\|\boldsymbol{a}_t\|_2^2+(1-c_{1,t})\|\boldsymbol{b}_t\|_2^2}{\alpha(1-c_{1,t})\|\boldsymbol{a}_t\|_2^2+c_{0,t}\|\boldsymbol{b}_t\|_2^2} \cdot \boldsymbol{a}_t\\ \hline \boldsymbol{b}_t\end{array}\right] \doteq \lambda \cdot \tilde{\boldsymbol{x}} \ ,\end{aligned} \tag{63}$$

and so we obtain the eigenvalue

$$\lambda = \frac{\alpha(1-c_{1,t})\|\boldsymbol{a}_t\|_2^2 + c_{0,t}\|\boldsymbol{b}_t\|_2^2}{(1-c_{1,t})^2 - c_{0,t}c_{2,t}} \ , \tag{64}$$

and we get from the eigenvector that α satisfies

$$\alpha = \frac{\alpha c_{2,t}\|\boldsymbol{a}_t\|_2^2 + (1-c_{1,t})\|\boldsymbol{b}_t\|_2^2}{\alpha(1-c_{1,t})\|\boldsymbol{a}_t\|_2^2 + c_{0,t}\|\boldsymbol{b}_t\|_2^2} \ , \tag{65}$$

and so

$$(1-c_{1,t})\|\boldsymbol{a}_t\|_2^2\alpha^2 + (c_{0,t}\|\boldsymbol{b}_t\|_2^2 - c_{2,t}\|\boldsymbol{a}_t\|_2^2)\alpha - (1-c_{1,t})\|\boldsymbol{b}_t\|_2^2 = 0 \ . \tag{66}$$

We note that the discriminant is

$$\tau = (c_{0,t}\|\boldsymbol{b}_t\|_2^2 - c_{2,t}\|\boldsymbol{a}_t\|_2^2)^2 + 4(1-c_{1,t})^2\|\boldsymbol{a}_t\|_2^2\|\boldsymbol{b}_t\|_2^2 \ , \tag{67}$$

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

$$\alpha_\pm = \frac{c_{2,t}\|\boldsymbol{a}_t\|_2^2 - c_{0,t}\|\boldsymbol{b}_t\|_2^2 \pm \sqrt{(c_{0,t}\|\boldsymbol{b}_t\|_2^2 - c_{2,t}\|\boldsymbol{a}_t\|_2^2)^2 + 4(1-c_{1,t})^2\|\boldsymbol{a}_t\|_2^2\|\boldsymbol{b}_t\|_2^2}}{2(1-c_{1,t})\|\boldsymbol{a}_t\|_2^2} \ . \tag{68}$$

yielding two non-zero eigenvalues,

$$\lambda_\pm(\mathrm{U}_t) = \frac{1}{2\zeta} \cdot \left(c_{2,t}\|\boldsymbol{a}_t\|_2^2 + c_{0,t}\|\boldsymbol{b}_t\|_2^2 \pm \sqrt{(c_{0,t}\|\boldsymbol{b}_t\|_2^2 - c_{2,t}\|\boldsymbol{a}_t\|_2^2)^2 + 4(1-c_{1,t})^2\|\boldsymbol{a}_t\|_2^2\|\boldsymbol{b}_t\|_2^2}\right) \tag{69}$$

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

$$(c_{2,t}\|\boldsymbol{a}_t\|_2^2 + c_{0,t}\|\boldsymbol{b}_t\|_2^2)^2 \quad < \quad (c_{0,t}\|\boldsymbol{b}_t\|_2^2 - c_{2,t}\|\boldsymbol{a}_t\|_2^2)^2 + 4(1-c_{1,t})^2\|\boldsymbol{a}_t\|_2^2\|\boldsymbol{b}_t\|_2^2 \ , \tag{70}$$

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

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

$$\begin{aligned} \lambda^*(\mathrm{U}_t) = \lambda_+ \quad &\leq \quad \frac{1}{2}\cdot\left(c_{2,t}\|\boldsymbol{a}_t\|_2^2 + c_{0,t}\|\boldsymbol{b}_t\|_2^2 + |c_{0,t}\|\boldsymbol{b}_t\|_2^2 - c_{2,t}\|\boldsymbol{a}_t\|_2^2| + 2(1-c_{1,t})\|\boldsymbol{a}_t\|_2\|\boldsymbol{b}_t\|_2\right) \\ &\leq \quad c_{2,t}\|\boldsymbol{a}_t\|_2^2 + c_{0,t}\|\boldsymbol{b}_t\|_2^2 + (1-c_{1,t})\|\boldsymbol{a}_t\|_2\|\boldsymbol{b}_t\|_2 \ . \end{aligned} \tag{71}$$

Now, remark that because V_t is positive definite,

$$\begin{aligned} c_{0,t} - 2c_{1,t} + c_{2,t} \quad &\doteq \quad \boldsymbol{a}_t^{+\top}\mathrm{V}_t\boldsymbol{a}_t^+ - 2\boldsymbol{a}_t^{+\top}\mathrm{V}_t\boldsymbol{b}_t^+ + \boldsymbol{b}_t^{+\top}\mathrm{V}_t\boldsymbol{b}_t^+ \\ &= \quad (\boldsymbol{a}_t^+ - \boldsymbol{b}_t^+)^\top\mathrm{V}_t(\boldsymbol{a}_t^+ - \boldsymbol{b}_t^+) \\ &\geq \quad 0 \ , \end{aligned} \tag{72}$$

showing that $c_{1,t} \leq (c_{0,t} + c_{2,t})/2$. So we get from ineq. (71),

$$\begin{aligned} \lambda^*(\mathrm{U}_t) \quad &\leq \quad \frac{1}{\zeta}\cdot\left(c_{2,t}\|\boldsymbol{a}_t\|_2^2 + c_{0,t}\|\boldsymbol{b}_t\|_2^2 + \left(1 + \frac{c_{0,t}+c_{2,t}}{2}\right)\|\boldsymbol{a}_t\|_2\|\boldsymbol{b}_t\|_2\right) \\ &\leq \quad \frac{1}{\zeta}\cdot\left(1 + \frac{3}{2}\cdot(c_{0,t}+c_{2,t})\right)\cdot\max\{\|\boldsymbol{a}_t\|_2^2, \|\boldsymbol{b}_t\|_2^2\} \\ &\leq \quad \frac{2 + 3(c_{0,t}+c_{2,t})}{2((1-c_{1,t})^2 - c_{0,t}c_{2,t})}\cdot\max\{\|\boldsymbol{a}_t\|_2^2, \|\boldsymbol{b}_t\|_2^2\} \ . \end{aligned} \tag{73}$$

When $\zeta < 0$, we remark that $\lambda_+ < \lambda_-$ and so U_t is negative semi-definite. We also remark that U_t is diagonalizable: non-zero eigenvalues have distinct eigenvectors eigenvalue and 0 is associated to the eigen subspace complementary of the span of those two eigenvectors.

Whenever $c_{1,t} \neq 1$, it is then easy to check that for any $z \in \{|\lambda_+|, |\lambda_-|\}$, ineq. (73) brings

$$z \quad \leq \quad \frac{2 + 3(c_{0,t}+c_{2,t})}{2|(1-c_{1,t})^2 - c_{0,t}c_{2,t}|}\cdot\max\{\|\boldsymbol{a}_t\|_2^2, \|\boldsymbol{b}_t\|_2^2\} \ . \tag{74}$$

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

$$\begin{aligned} z \quad &\leq \quad \max\left\{\frac{1}{c_{0,t}}, \frac{1}{c_{2,t}}\right\}\cdot\max\{\|\boldsymbol{a}_t\|_2^2, \|\boldsymbol{b}_t\|_2^2\} \\ &< \quad \left(1 + \frac{3}{c_{0,t}} + \frac{3}{c_{2,t}}\right)\cdot\max\{\|\boldsymbol{a}_t\|_2^2, \|\boldsymbol{b}_t\|_2^2\} \\ &= \frac{2 + 3(c_{0,t}+c_{2,t})}{2|(1-c_{1,t})^2 - c_{0,t}c_{2,t}|}\cdot\max\{\|\boldsymbol{a}_t\|_2^2, \|\boldsymbol{b}_t\|_2^2\} \ . \end{aligned} \tag{75}$$

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

Lemma I *The following holds true:*

$$\begin{aligned}\|\boldsymbol{b}_t^+\|_2^2 = \|\boldsymbol{b}_t\|_2^2 & \leq & 2\xi \cdot X_*^2 \ , \qquad (76)\\ \|\boldsymbol{a}_t^+\|_2^2 = \|\boldsymbol{a}_t\|_2^2 & \leq & 2\xi \cdot X_*^2 \ , \qquad (77)\end{aligned}$$

where ξ is defined in eq. (6).

Proof Suppose P_t (ε, τ)-inexact with the ε, τ parameters used to compute ξ. To prove ineq. (76), we make two applications of (2) with $\mathsf{F} \doteq \mathsf{B}$. Fix $\boldsymbol{w} \doteq (1/\|\boldsymbol{x}_{v_{\mathsf{B}(t)}}\|_2) \cdot \boldsymbol{x}_{v_{\mathsf{B}(t)}}$. We get:

$$\begin{aligned}|(\boldsymbol{x}_{v_{\mathsf{B}(t)}} - \boldsymbol{x}_{u_{\mathsf{B}(t)}})_{\mathsf{B}}^\top \boldsymbol{x}_{v_{\mathsf{B}(t)\mathsf{B}}}| & \leq & \varepsilon \cdot \max\{|\boldsymbol{x}_{u_{\mathsf{B}(t)}}^\top \boldsymbol{x}_{v_{\mathsf{B}(t)}}|, \|\boldsymbol{x}_{v_{\mathsf{B}(t)}}\|_2^2\} + \tau \cdot \|\boldsymbol{x}_{v_{\mathsf{B}(t)}}\|_2\\ & \leq & \varepsilon \cdot X_*^2 + \tau \cdot X_* = \xi \cdot X_*^2 \ . \qquad (78)\end{aligned}$$

Fix $\boldsymbol{w} \doteq (1/\|\boldsymbol{x}_{u_{\mathsf{B}(t)}}\|_2) \cdot \boldsymbol{x}_{u_{\mathsf{B}(t)}}$. We get:

$$\begin{aligned}|(\boldsymbol{x}_{u_{\mathsf{B}(t)}} - \boldsymbol{x}_{v_{\mathsf{B}(t)}})_{\mathsf{B}}^\top \boldsymbol{x}_{u_{\mathsf{B}(t)\mathsf{B}}}| & \leq & \varepsilon \cdot \max\{|\boldsymbol{x}_{u_{\mathsf{B}(t)}}^\top \boldsymbol{x}_{v_{\mathsf{B}(t)}}|, \|\boldsymbol{x}_{u_{\mathsf{B}(t)}}\|_2^2\} + \tau \cdot \|\boldsymbol{x}_{u_{\mathsf{B}(t)}}\|_2\\ & \leq & \varepsilon \cdot X_*^2 + \tau \cdot X_* = \xi \cdot X_*^2 \ . \qquad (79)\end{aligned}$$

Folding together ineqs. (78) and (79) yields

$$\begin{aligned}&\|(\boldsymbol{x}_{v_{\mathsf{B}(t)}} - \boldsymbol{x}_{u_{\mathsf{B}(t)}})_{\mathsf{B}}\|_2^2 = (\boldsymbol{x}_{v_{\mathsf{B}(t)}} - \boldsymbol{x}_{u_{\mathsf{B}(t)}})_{\mathsf{B}}^\top (\boldsymbol{x}_{v_{\mathsf{B}(t)}} - \boldsymbol{x}_{u_{\mathsf{B}(t)}})_{\mathsf{B}}\\ &\leq \ |(\boldsymbol{x}_{v_{\mathsf{B}(t)}} - \boldsymbol{x}_{u_{\mathsf{B}(t)}})_{\mathsf{B}}^\top \boldsymbol{x}_{v_{\mathsf{B}(t)\mathsf{B}}}| + |(\boldsymbol{x}_{v_{\mathsf{B}(t)}} - \boldsymbol{x}_{u_{\mathsf{B}(t)}})_{\mathsf{B}}^\top \boldsymbol{x}_{u_{\mathsf{B}(t)\mathsf{B}}}|\\ &\leq \ 2\xi \cdot X_*^2 \ . \qquad (80)\end{aligned}$$

We get

$$\|\boldsymbol{b}_t^+\|_2^2 = \|\boldsymbol{b}_t\|_2^2 = \|(\boldsymbol{x}_{v_{\mathsf{B}(t)}} - \boldsymbol{x}_{u_{\mathsf{B}(t)}})_{\mathsf{B}}\|_2^2 \ \leq \ 2\xi \cdot X_*^2 \ , \qquad (81)$$

which yields ineq. (76). To get ineq. (77), we switch $\mathsf{F} \doteq \mathsf{B}$ by $\mathsf{F} \doteq \mathsf{A}$ in our application of point (2). ■

Lemma J *If the data-loss calibration assumption holds,*

$$c_{i,t} \ \leq \ \frac{1}{12} \ , \forall i \in \{0, 1, 2\} \ . \qquad (82)$$

Proof We remark that

$$\begin{aligned}c_{0,t} & \doteq & \boldsymbol{a}_t^{+\top} \mathrm{v}_t \boldsymbol{a}_t^+\\ & \leq & \lambda^*(\mathrm{v}_t)\|\boldsymbol{a}_t^+\|_2^2\\ & \leq & 2\lambda^*(\mathrm{v}_t)\xi \cdot X_*^2 \ , \qquad (83)\end{aligned}$$

and for the same reasons,

$$c_{2,t} \ \leq \ 2\lambda^*(\mathrm{v}_t)\xi \cdot X_*^2. \qquad (84)$$

Hence, it comes from the proof of Lemma H that we also have $c_{2,t} \leq 2\lambda^*(\mathrm{v}_t)\xi \cdot X_*^2$. Using ineq. (53) in Lemma G, we thus obtain for any $i \in \{0, 1, 2\}$:

$$\begin{aligned} c_{i,t} &\leq \frac{1}{m} \cdot \frac{2\xi \cdot X_*^2}{\mathrm{sign}(c)\mu_t + \frac{1}{|c|}\lambda^\circ(\Gamma)} \\ &= \frac{\xi}{m} \cdot \frac{2|c|X_*^2}{c\mu_t + \lambda^\circ(\Gamma)} \\ &\leq \frac{1}{4} \cdot \frac{1}{4} < \frac{1}{12} \ , \end{aligned} \tag{85}$$

as claimed. The last inequality uses the data-loss calibration assumption. ■

Corollary K *If the data-loss calibration assumption holds, then the invertibility condition holds.*

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

Lemma L *Using notations of Subsection I.I.2.1, if the data-loss calibration assumption holds, the following holds true:* $\mathrm{I}_d + \Lambda_t \succ 0$ *and*

$$\lambda^*(\Lambda_t) \ \leq \ \frac{\xi}{m} \ . \tag{86}$$

Proof To simplify notations, let us shift $t \rightarrow t-1$. First note that $\lambda^\circ(\mathrm{v}_{t-1}) \geq 1/(\gamma\lambda^*(\Gamma)) > 0$ and so $\mathrm{v}_{t-1} \succ 0$, which implies that $\Lambda_{t-1} \doteq \nu \cdot \mathrm{v}_{t-1}\mathrm{u}_t = \nu \mathrm{v}_{t-1}^{1/2}(\mathrm{v}_{t-1}^{1/2}\mathrm{u}_t\mathrm{v}_{t-1}^{1/2})\mathrm{v}_{t-1}^{-1/2}$, *i.e.* Λ_{t-1} is similar to a symmetric matrix $(\mathrm{v}_{t-1}^{1/2}\mathrm{u}_t\mathrm{v}_{t-1}^{1/2})$ and therefore has only real eigenvalues. We have seen from the proof of Corollary K that $(1 - c_{1,t})^2 - c_{0,t}c_{2,t} > 0, \forall t \geq 1$, which implies from Lemma H that u_t is indefinite with two distinct eigenvalues, say $-\lambda_- \leq 0$ and $\lambda_+ > 0$, with $\lambda_+ \geq \lambda_-$. We know from Lemma H that we can write $\mathrm{u}_t = -\lambda_-\boldsymbol{u}_t\boldsymbol{u}_t^\top + \lambda_+\boldsymbol{v}_t\boldsymbol{v}_t^\top$ with $\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1$, so

$$\boldsymbol{x}^\top \Lambda_{t-1}\boldsymbol{x} \ = \ -\lambda_-\nu \cdot (\boldsymbol{u}_t^\top\boldsymbol{x}) \cdot \boldsymbol{x}^\top \mathrm{v}_{t-1}\boldsymbol{u}_t + \lambda_+\nu \cdot (\boldsymbol{v}_t^\top\boldsymbol{x}) \cdot \boldsymbol{x}^\top \mathrm{v}_{t-1}\boldsymbol{v}_t. \tag{87}$$

We now remark that since v_{t-1} is positive definite (Lemma G), $(\boldsymbol{u}_t^\top\boldsymbol{x}) \cdot \boldsymbol{x}^\top\mathrm{v}_{t-1}\boldsymbol{u}_t \geq 0, (\boldsymbol{v}_t^\top\boldsymbol{x}) \cdot \boldsymbol{x}^\top\mathrm{v}_{t-1}\boldsymbol{v}_t \geq 0, \forall \boldsymbol{x}$, which, together with (87) shows that we have the first inequality of the following:

$$\begin{aligned} \lambda^*(\Lambda_{t-1}) &\leq |\nu|\lambda_+\lambda^*(\mathrm{v}_{t-1}) && (88) \\ &\leq |\nu| \cdot \frac{2 + 3(c_{0,t} + c_{2,t})}{2((1 - c_{1,t})^2 - c_{0,t}c_{2,t})} \cdot \max\{\|\boldsymbol{a}\|_2^2, \|\boldsymbol{b}_t\|_2^2\} \cdot \lambda^*(\mathrm{v}_{t-1}) && (89) \\ &\leq \frac{2 + 3(c_{0,t} + c_{2,t})}{(1 - c_{1,t})^2 - c_{0,t}c_{2,t}} \cdot |\nu|\lambda^*(\mathrm{v}_{t-1})\xi \cdot X_*^2. && (90) \end{aligned}$$

Ineq. (89) is due to Lemma H and ineq. (90) is due to Lemma I. We do not put the absolute value in the denominator of (90) because of Corollary K. We now use Lemma J and get:

$$\begin{aligned} (1 - c_{1,t})^2 - c_{0,t}c_{2,t} &\geq \left(1 - \frac{1}{12}\right)^2 - \frac{1}{144} \\ &= \frac{5}{6} \ . \end{aligned} \tag{91}$$

Letting $U \doteq 2\lambda^*(\mathrm{v}_t)\xi \cdot X_*^2$ for short, we thus get from (83), (84) in the proof of Lemma J:

$$\begin{aligned} \lambda^*\left(\Lambda_t\right) &\leq |\nu| \cdot \frac{6}{10} \cdot (2 + 3(U+U))U \\ &= |\nu| \cdot \frac{6}{10} \cdot (2U + 6U^2) \ . \end{aligned} \tag{92}$$

Now we want $\lambda^*\left(\Lambda_t\right) \leq \xi/m$, which translates into a second-order inequality for U, whose solution imposes the following upperbound on U:

$$6U \leq -1 + \sqrt{1 + \frac{10\xi}{|\nu|m}} \ . \tag{93}$$

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 $\xi/m \geq 0$), we get the sufficient condition for ineq. (93) to be satisfied:

$$12\lambda^*(\mathrm{v}_t)\xi \cdot X_*^2 \leq \frac{5\xi}{|\nu|m} - \frac{100}{8} \cdot \left(\frac{\xi}{|\nu|m}\right)^2 \ . \tag{94}$$

Now, it comes from Lemma G that a sufficient condition for ineq. (94) is that

$$\frac{\xi}{m} \cdot \frac{12X_*^2}{\mathrm{sign}(c)\mu_t + \frac{1}{|c|}\lambda^\circ(\Gamma)} \leq \frac{5\xi}{|\nu|m} - \frac{100}{8} \cdot \left(\frac{\xi}{|\nu|m}\right)^2 \ , \tag{95}$$

which, after simplification, is equivalent to

$$\frac{12}{5} \cdot \frac{|\nu|X_*^2}{\mathrm{sign}(c)\mu_t + \frac{1}{|c|}\lambda^\circ(\Gamma)} + \frac{5\xi}{2|\nu|m} \leq 1 \ , \tag{96}$$

or,

$$\frac{6|b|}{5} \cdot \frac{X_*^2}{c\mu_t + \lambda^\circ(\Gamma)} + \frac{5\xi}{2|\nu|m} \leq 1 \ , \tag{97}$$

But, the data-loss calibration assumption implies that the left-hand side is no more than $(3/5) + (5/16) = 73/80 < 1$, and ineq. (86) follows.

It also trivially follows that $\mathrm{I}_d + \Lambda_t$ has only real eigenvalues. To prove that they are all strictly positive, we know that the only potentially negative eigenvalue of U_t, λ_- (Lemma H) is smaller in absolute value to $\lambda^*(\mathrm{U}_t)$. V_t being positive definite, we thus have under the data-loss calibration assumption:

$$\begin{aligned} \lambda^\circ(\mathrm{I}_d + \Lambda_t) &\geq 1 - \frac{\xi}{m} \\ &\geq 1 - \frac{1}{4} = \frac{3}{4} > 0 \ , \end{aligned} \tag{98}$$

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

We recall that $0 \leq T_+ \leq T$ denote the number of elementary permutations that act between classes, and $\rho \doteq T_+/T$ denote the proportion of such elementary permutations among all. If $\hat{\mathsf{P}}$ is (ε, τ)-inexact then let

$$K_\xi \doteq \frac{\exp(\xi)+1}{8\sqrt{2\xi}}. \quad (99)$$

Theorem M *If the data-loss calibration assumption holds, then the following holds for all* $T \geq 1$*:*

$$\|\boldsymbol{\theta}^*_T - \boldsymbol{\theta}^*_0\|_2 \leq \frac{\xi^{\frac{1}{4}}}{m}\cdot T^2 \cdot \left(\xi^{\frac{3}{4}}\|\boldsymbol{\theta}^*_0\|_2 + \frac{\rho|\nu|\xi^{\frac{3}{4}}K_\xi}{X_*}\right), \quad (100)$$

and in addition, if $\hat{\mathsf{P}}$ *is* α*-bounded then*

$$\|\boldsymbol{\theta}^*_T - \boldsymbol{\theta}^*_0\|_2 \leq \left(\frac{\xi^{1/4}}{m}\right)^\alpha \cdot \left(\xi^{\frac{3}{4}}\|\boldsymbol{\theta}^*_0\|_2 + \frac{\rho|\nu|\xi^{\frac{3}{4}}K_\xi}{X_*}\right), \quad (101)$$

where we recall $\nu \doteq -b/(2|c|)$.

Remark: we have $\lim_{0,+\infty} K_\xi = +\infty$ *but* $\lim_0 \xi^{\frac{3}{4}}K_\xi = 0$, so the RHS of (100) and (101) converge to zero with $\xi \rightarrow 0$ (so, as $\hat{\mathsf{P}}$ gets 'better').

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

$$\|\boldsymbol{\theta}^*_T - \boldsymbol{\theta}^*_0\|_2 \leq \|(\mathrm{H}_{T,0} - \mathrm{I}_d)\boldsymbol{\theta}^*_0\|_2 + \left\|\sum_{t=0}^{T-1} \mathrm{H}_{T,t+1}\boldsymbol{\lambda}_t\right\|_2 . \quad (102)$$

To prove Theorem M, we prove upperbounds on the two terms of (102).

Lemma N *If the data-loss calibration holds then we have for all* $T \geq 1$*:*

$$\|(\mathrm{H}_{T,0} - \mathrm{I}_d)\boldsymbol{\theta}^*_0\|_2 \leq \frac{\xi}{m}\cdot T^2 \cdot \|\boldsymbol{\theta}^*_0\|_2. \quad (103)$$

If furthermore $\hat{\mathsf{P}}$ *is* α*-bounded, then the following holds for all* $T \geq 1$*:*

$$\|(\mathrm{H}_{T,0} - \mathrm{I}_d)\boldsymbol{\theta}^*_0\|_2 \leq \left(\frac{\xi^{1/4}}{m}\right)^\alpha \cdot \xi^{\frac{3}{4}}\|\boldsymbol{\theta}^*_0\|_2 . \quad (104)$$

Proof Denote for short $q \doteq \xi/m$. Remember that we can suppose without loss of generality that $\xi \leq 2$ (Lemma 2). We would like

$$(1+q)^T \leq 1 + T^2 q, \forall q \in \left[0, \frac{2}{m}\right], \forall T \in \{0, 1, ..., m-1\}. \quad (105)$$

This is trivially true for $T = 0, 1$. Let $f_q(T) \doteq (1+q)^T, g_q(T) \doteq 1 + T^2 q$. We have

$$\frac{f_q'(T)}{g_q'(T)} = \frac{(1+q)^{T-1}}{2q} \geq 1, \forall T \geq 1, \tag{106}$$

because of the data-loss calibration assumption (implies $q \leq 1/4$). Hence, to show (105), it is sufficient to show that it holds for an upperbound on T which for simplicity we take as $T = m$, *i.e.*, replacing the expression of q, we want

$$\left(1 + \frac{\xi}{m}\right)^m \leq 1 + \xi m \tag{107}$$

Since $1 + z \leq \exp(z)$, we want $\exp(m(\xi/m)) = \exp \xi \leq 1 + \xi m$. Since $\exp(z) \leq 1 + z((\exp(a) - 1)/a), \forall z \in [0, a]$ (the RHS is the chord of $\exp$ on points $0, a$), we get, by fixing $a = 2$ that it is sufficient that

$$m \geq \frac{\exp(2) - 1}{2}, \tag{108}$$

a sufficient condition for which is $m \geq 4$. Hence, if the data-loss calibration assumption holds, then (105) is true and so we get (letting again $q \doteq \xi/m$)

$$\begin{aligned} \| \left(\mathrm{H}_{T,0} - \mathrm{I}_d\right) \boldsymbol{\theta}_0^* \|_2 &\leq \lambda^*(\mathrm{H}_{T,0} - \mathrm{I}_d) \cdot \|\boldsymbol{\theta}_0^*\|_2 \\ &\leq ((1+q)^T - 1) \cdot \|\boldsymbol{\theta}_0^*\|_2 \qquad (109) \\ &\leq T^2 q \cdot \|\boldsymbol{\theta}_0^*\|_2 \ . \qquad (110) \end{aligned}$$

(109) holds given definition (20), Lemma L; (110) holds because of (105). If furthermore $\hat{\mathsf{P}}$ is α -bounded, then we write

$$T^2 q \doteq \frac{T^2 \xi}{m} = \frac{T^2 \xi^{1/4}}{m} \cdot \xi^{\frac{3}{4}},$$

and conclude

$$\| \left(\mathrm{H}_{T,0} - \mathrm{I}_d\right) \boldsymbol{\theta}_0^* \|_2 \leq \left(\frac{\xi^{1/4}}{m}\right)^\alpha \cdot \xi^{\frac{3}{4}} \|\boldsymbol{\theta}_0^*\|_2 \ . \tag{111}$$

This ends the proof of Lemma N. ■

We now have a look at the rightmost term in (102). The key part of the lemma to follow relies on remarking that whenever an elementary permutation does not permute observations between classes, its contribution to the sums is in fact zero.

Lemma O *If the data-loss calibration assumption holds then we have for all $T \geq 1$:*

$$\left\|\sum_{t=0}^{T-1} \mathrm{H}_{T,t+1} \boldsymbol{\lambda}_t\right\|_2 \leq \frac{\xi}{m} \cdot T^2 \cdot \frac{\rho |\nu| K_\xi}{X_*} \ . \tag{112}$$

If furthermore $\hat{\mathsf{P}}$ is α-bounded, then the following holds for all $T \geq 1$:

$$\left\|\sum_{t=0}^{T-1} \mathrm{H}_{T,t+1} \boldsymbol{\lambda}_t\right\|_2 \leq \left(\frac{\xi^{1/4}}{m}\right)^\alpha \cdot \frac{\rho |\nu| \xi^{\frac{3}{4}} K_\xi}{X_*} \ . \tag{113}$$

Proof Denote for short

$$\mathrm{R} \doteq \sum_{t=0}^{T-1} \mathrm{H}_{T,t+1}\boldsymbol{\lambda}_t \quad . \tag{114}$$

Using eq. (22), we can simplify R since $\boldsymbol{\lambda}_t = \nu \mathrm{V}_{t+1}\boldsymbol{\varepsilon}_t$, so if we define $\mathrm{G}_{.,.}$ from $\mathrm{H}_{.,.}$ as follows, for $0 \leq j \leq i$:

$$\mathrm{G}_{i,j} \doteq \nu \mathrm{H}_{i,j}\mathrm{V}_j \quad , \tag{115}$$

then we get

$$\mathrm{R} \doteq \sum_{t=0}^{T-1} \mathrm{G}_{T,t+1}\boldsymbol{\varepsilon}_t \quad , \tag{116}$$

where we recall that $\boldsymbol{\varepsilon}_t \doteq \boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t$ is the shift in the mean operator, *which is the null vector whenever* P_t *acts in a specific class* $(y_{u_{\mathsf{A}}(t)} = y_{v_{\mathsf{A}}(t)})$. To see this, we remark

$$\begin{aligned} \boldsymbol{\varepsilon}_t &\doteq \boldsymbol{\mu}_{t+1} - \boldsymbol{\mu}_t \\ &= \sum_i y_i \cdot \left[\begin{array}{c} \boldsymbol{x}_{i_{\mathsf{A}}} \\ \hline \boldsymbol{x}_{(t+1)i_{\mathsf{B}}} \end{array}\right] - \sum_i y_i \cdot \left[\begin{array}{c} \boldsymbol{x}_{i_{\mathsf{A}}} \\ \hline \boldsymbol{x}_{ti_{\mathsf{B}}} \end{array}\right] \\ &= \sum_i y_i \cdot \left[\begin{array}{c} 0 \\ \hline \boldsymbol{x}_{(t+1)i_{\mathsf{B}}} \end{array}\right] - \sum_i y_i \cdot \left[\begin{array}{c} 0 \\ \hline \boldsymbol{x}_{ti_{\mathsf{B}}} \end{array}\right] \\ &= \left[\begin{array}{c} 0 \\ \hline \sum_i y_i \cdot (\boldsymbol{x}_{(t+1)i_{\mathsf{B}}} - \boldsymbol{x}_{ti_{\mathsf{B}}}) \end{array}\right] \doteq \left[\begin{array}{c} 0 \\ \hline \boldsymbol{\varepsilon}_{t\mathsf{B}} \end{array}\right] \quad , \end{aligned} \tag{117}$$

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

$$\begin{aligned} \boldsymbol{\varepsilon}_{t\mathsf{B}} &= y_{u_{\mathsf{A}}(t)} \cdot (\boldsymbol{x}_{v_{\mathsf{B}}(t)} - \boldsymbol{x}_{u_{\mathsf{B}}(t)})_{\mathsf{B}} + y_{v_{\mathsf{A}}(t)} \cdot (\boldsymbol{x}_{u_{\mathsf{B}}(t)} - \boldsymbol{x}_{v_{\mathsf{B}}(t)})_{\mathsf{B}} \\ &= (y_{u_{\mathsf{A}}(t)} - y_{v_{\mathsf{A}}(t)}) \cdot (\boldsymbol{x}_{v_{\mathsf{B}}(t)} - \boldsymbol{x}_{u_{\mathsf{B}}(t)})_{\mathsf{B}} \quad . \end{aligned} \tag{118}$$

Hence,

$$\begin{aligned} \|\boldsymbol{\varepsilon}_t\|_2 = \|\boldsymbol{\varepsilon}_{t\mathsf{B}}\|_2 &= 1_{y_{u_{\mathsf{A}}(t)} \neq y_{v_{\mathsf{A}}(t)}} \cdot \|(\boldsymbol{x}_{v_{\mathsf{B}}(t)} - \boldsymbol{x}_{u_{\mathsf{B}}(t)})_{\mathsf{B}}\|_2 \\ &\leq 1_{y_{u_{\mathsf{A}}(t)} \neq y_{v_{\mathsf{A}}(t)}} \cdot \sqrt{2\xi} X_* \quad , \end{aligned} \tag{119}$$

from Lemma I, 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-loss calibration assumption and Lemma G that

$$\begin{aligned} \lambda^*(\mathrm{V}_t) &\leq \frac{1}{m} \cdot \frac{1}{\mathrm{sign}(c)\mu_t + \frac{1}{|c|}\lambda^\circ(\Gamma)} \\ &= \frac{|c|}{mX_*^2} \cdot \frac{X_*^2}{c\mu_t + \lambda^\circ(\Gamma)} \\ &\leq \frac{|c|}{mX_*^2} \cdot \frac{1}{2} \cdot \min\left\{\frac{1}{|b|}, \frac{1}{4|c|}\right\} \\ &\leq \frac{1}{8mX_*^2} \quad . \end{aligned} \tag{120}$$

We know that $\mathrm{H}_{.,.}$ is a product of positive definite matrices ((20) and Lemma L), and $\mathrm{V}_{.}$ is also positive definite (Lemma G), so using (Bhatia, 1997, Problem III.6.14), Lemma L and ineq. (120), we obtain

$$\begin{aligned}
\lambda^{\downarrow}_1(\mathrm{G}_{T,t+1}) & \doteq & \lambda^{\downarrow}_1(\nu \mathrm{H}_{T,t+1}\mathrm{V}_{t+1}) \\
& \leq & |\nu| \cdot \left(1+\frac{\xi}{m}\right)^{T-t-1} \cdot \frac{1}{8mX_*^2}\ . \qquad (121)
\end{aligned}$$

So,

$$\begin{aligned}
\|\mathrm{R}\|_2 & \leq & \sum_{t=0}^{T-1} \lambda^{\downarrow}_1(\mathrm{G}_{T,t+1})\|\boldsymbol{\varepsilon}_t\|_2 \\
& \leq & \frac{|\nu|}{4\sqrt{2}} \cdot \sum_{t=0}^{T-1} 1_{y_{u_{\mathsf{A}(t)}} \neq y_{v_{\mathsf{A}(t)}}} \cdot \left(1+\frac{\xi}{m}\right)^{T-t-1} \cdot \frac{\sqrt{\xi}}{mX_*} \\
& & = \frac{|\nu|}{4X_*} \cdot \frac{1}{\sqrt{2\xi}} \cdot \sum_{t=0}^{T-1} 1_{y_{u_{\mathsf{A}(t)}} \neq y_{v_{\mathsf{A}(t)}}} \cdot \left(1+\frac{\xi}{m}\right)^{T-t-1} \cdot \frac{\xi}{m}\ , \qquad (122)
\end{aligned}$$

from ineq. (119). Assuming $T_+ \leq T$ errors are made by permutations between classes and recalling $q \doteq \xi/m$, we see that the largest upperbound for $\|\mathrm{R}\|_2$ in ineq. (122) is obtained when all T_+ errors happen during the first elementary permutations indexes in the sequence in $\hat{\mathsf{P}}$, so we get

$$\begin{aligned}
\|\mathrm{R}\|_2 & \leq & \frac{|\nu|}{4X_*} \cdot \frac{1}{\sqrt{2\xi}} \cdot \sum_{t=0}^{T_+-1} q(1+q)^{T-t-1} \\
& & = \frac{|\nu|}{4X_*} \cdot \frac{1}{\sqrt{2\xi}} \cdot q(1+q)^{T-T_+} \sum_{t=0}^{T_+-1} (1+q)^{T_+-t-1} \\
& = & \frac{|\nu|}{4X_*} \cdot \frac{1}{\sqrt{2\xi}} \cdot q(1+q)^{T-T_+} \sum_{t=0}^{T_+-1} (1+q)^{t} \\
& = & \frac{|\nu|}{4X_*} \cdot \frac{1}{\sqrt{2\xi}} \cdot \frac{q(1+q)^{T-T_+}(1-(1+q)^{T_+})}{1-(1+q)} \\
& = & \frac{|\nu|}{4X_*} \cdot \frac{1}{\sqrt{2\xi}} \cdot (1+q)^{T-T_+}((1+q)^{T_+}-1) \\
& = & \frac{|\nu|}{4X_*} \cdot \frac{1}{\sqrt{2\xi}} \cdot ((1+q)^{T}-1) \cdot \frac{\left(1-\frac{1}{(1+q)^{\rho T}}\right)}{\left(1-\frac{1}{(1+q)^{T}}\right)} \\
& \leq & T^2 q \cdot \frac{|\nu|}{4X_*} \cdot \frac{1}{\sqrt{2\xi}} \cdot \frac{\left(1-\frac{1}{(1+q)^{\rho T}}\right)}{\left(1-\frac{1}{(1+q)^{T}}\right)}\ . \qquad (123)
\end{aligned}$$

Ineq. (123) holds because of ineq. (105). For $z > 1$, we let

$$f(z) \ \doteq\ \frac{1-\frac{1}{z^\rho}}{1-\frac{1}{z}} = \frac{z-z^{1-\rho}}{z-1}. \qquad (124)$$

We have $\lim_1 f(z) = \rho$,

$$f'(z) \quad = \quad \frac{1-\rho+z\rho-z^\rho}{z^\rho(z-1)^2}, \tag{125}$$

and we check that its numerator is ≥ 0 for $\rho \in [0,1]$ (it zeroes in $\rho = 0, 1$ and is convex in the interval). We also have $\lim_1 f'(z) = \rho(1-\rho)/2$,

$$f''(z) \quad = \quad \frac{z^{1-\rho}\cdot g_\rho(z)}{(z-1)^3}, \tag{126}$$

with $g_\rho(z) \doteq \rho(1-\rho) - 2(1-\rho)(1+\rho)z - \rho(1+\rho)z^2 + 2z^{1+\rho}$, which satisfies $\lim_1 f''(z) = -(1/3)\cdot\rho(1-\rho)(1+\rho) \leq 0$. We see that $g_0(z) = g_1(z) = g_\rho(1) = 0$ and $g'_\rho(z) = -2(1-\rho)(1+\rho) - 2\rho(1+\rho)z + 2\rho z^\rho \leq -2(1-\rho)(1+\rho) - 2\rho(1+\rho)z + 2\rho z = -2(1-\rho)(1+\rho) - 2\rho^2 z \leq 0$, which shows f is concave increasing for $z \geq 1$ and so

$$f(z) \quad \leq \quad \frac{\rho(1-\rho)}{2}z + \frac{\rho(1+\rho)}{2} = \frac{\rho}{2}\cdot((1-\rho)z+1+\rho), \forall z \geq 1, \tag{127}$$

and so $f((1+q)^T) \leq f((1+q)^m) \leq f(\exp(qm)) = f(\exp(\xi))$ which brings

$$f((1+q)^T) \quad \leq \quad \frac{\rho}{2}\cdot((1-\rho)\exp(\xi)+1+\rho), \tag{128}$$

and we get after (123)

$$\begin{aligned} \|\mathsf{R}\|_2 \quad &\leq \quad T^2 q\cdot\frac{|\nu|}{4X_*}\cdot\frac{1}{\sqrt{2\xi}}\cdot\frac{\rho}{2}\cdot((1-\rho)\exp(\xi)+1+\rho) \\ &= T^2 q\cdot\frac{|\nu|\rho}{X_*}\cdot\frac{(1-\rho)\exp(\xi)+1+\rho}{8\sqrt{2\xi}}. \end{aligned} \tag{129}$$

Finally, we remark that $(1-\rho)\exp(\xi)+1+\rho = \exp(\xi)+1-\rho(\exp(\xi)-1) \leq \exp(\xi)+1$ and get the statement of (112). We then get (113) by following the same proof as for (104) in Lemma N. This ends the proof of Lemma O. ■

We summarize the statements of Lemmata N and O. If $\hat{\mathsf{P}}$ is (ε,τ)-inexact then

$$\begin{aligned} \|\boldsymbol{\theta}^*_T - \boldsymbol{\theta}^*_0\|_2 \quad &\leq \quad \frac{\xi}{m}\cdot T^2\cdot\|\boldsymbol{\theta}^*_0\|_2 + \frac{\xi}{m}\cdot T^2\cdot\frac{\rho|\nu|K_\xi}{X_*} \\ &= \frac{\xi}{m}\cdot T^2\cdot\left(\|\boldsymbol{\theta}^*_0\|_2 + \frac{\rho|\nu|K_\xi}{X_*}\right) \qquad (130) \\ &= \frac{\xi^{\frac{1}{4}}}{m}\cdot T^2\cdot\left(\xi^{\frac{3}{4}}\|\boldsymbol{\theta}^*_0\|_2 + \frac{\rho|\nu|\xi^{\frac{3}{4}}K_\xi}{X_*}\right), \qquad (131) \end{aligned}$$

and in addition, if $\hat{\mathsf{P}}$ is α-bounded then

$$\begin{aligned} \frac{\xi^{\frac{1}{4}}}{m}\cdot T^2 \quad &\leq \quad \frac{\xi^{\frac{1}{4}}}{m}\cdot\left(\frac{m}{\xi^{1/4}}\right)^{1-\alpha} \\ &= \left(\frac{\xi^{1/4}}{m}\right)^\alpha, \end{aligned} \tag{132}$$

which we factor in (131). This yields the proof of Theorem M. ■

Theorem M easily yields Theorem 3.

I.3 Proof of Theorem 5

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

$$\begin{aligned} |y(\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*)^\top \boldsymbol{x}| = |(\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*)^\top \boldsymbol{x}| &\leq \|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*\|_2 \|\boldsymbol{x}\|_2 \\ &\leq \frac{\xi^{\frac{1}{4}}}{m} \cdot T^2 \cdot \left(\xi^{\frac{3}{4}} \|\boldsymbol{\theta}_0^*\|_2 + \frac{\rho |b| L_\xi}{|c| X_*}\right) \cdot X_* \\ &= \frac{\xi^{\frac{1}{4}}}{m} \cdot T^2 \cdot \left(\xi^{\frac{3}{4}} \|\boldsymbol{\theta}_0^*\|_2 X_* + \frac{|b| L_\xi}{|c|} \cdot \rho\right) \quad . \end{aligned} \tag{133}$$

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

$$\frac{m}{\xi^{\frac{1}{4}} T^2} > \frac{\xi^{\frac{3}{4}} \|\boldsymbol{\theta}_0^*\|_2 X_*}{\kappa} + \frac{|b| L_\xi}{\kappa |c|} \cdot \rho \ ; \tag{134}$$

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

$$\begin{aligned} y(\boldsymbol{\theta}_T^*)^\top \boldsymbol{x} &= y(\boldsymbol{\theta}_0^*)^\top \boldsymbol{x} + y(\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*)^\top \boldsymbol{x} \\ &\geq y(\boldsymbol{\theta}_0^*)^\top \boldsymbol{x} - |y(\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_0^*)^\top \boldsymbol{x}| \\ &> \kappa - \kappa = 0 \ , \end{aligned} \tag{135}$$

and we get the statement of (12). If furthermore $\hat{\mathsf{P}}$ is α-bounded, it comes

$$\frac{m}{\xi^{\frac{1}{4}} T^2} \geq \left(\frac{m}{\xi^{\frac{1}{4}}}\right)^\alpha, \tag{136}$$

from which we get (13).

I.4 Proof of Theorem 6

For any $\boldsymbol{\theta}^* \in \mathcal{C}$, we let $\mathcal{N}(\boldsymbol{\theta}^*)$ denote an open neighborhood of $\boldsymbol{\theta}^*$ over which ℓ_F is convex, which is guaranteed to be non empty by the assumptions on F. We proceed in two steps, first assuming that F is convex and then relaxing the assumption.

Case 1 — F convex. We consider any convex Ridge regularized loss $\ell_F(\hat{S}, \boldsymbol{\theta}; \Gamma_F) = L + R$ with $R \doteq \boldsymbol{\theta}^\top \Gamma_F \boldsymbol{\theta}$ and

$$L \doteq \frac{1}{m} \cdot \sum_i F(y_i \boldsymbol{\theta}^\top \hat{\boldsymbol{x}}_i) \ , \tag{137}$$

for some convex twice differentiable F. We first focus on the approximation of L via a Taylor loss. We perform a local Taylor-Lagrange expansion of each $F(y_i \boldsymbol{\theta}^\top \hat{\boldsymbol{x}}_i)$ in eq. (137) around 0 and obtain that there exists $c_1, c_2, ..., c_m \in F''(\mathbb{I}) \subseteq \mathbb{R}_+$ such that

$$L = F(0) + \frac{F'(0)}{m} \cdot \sum_i y_i \boldsymbol{\theta}^\top \hat{\boldsymbol{x}}_i + J \ , \tag{138}$$

where $\mathbb{I} \doteq [-\hat{X}_* \Theta_*, \hat{X}_* \Theta_*]$ (since $|y_i \boldsymbol{\theta}^\top \hat{\boldsymbol{x}}_i| \leq \hat{X}_* \Theta_*$ by Cauchy-Schwartz inequality) and $J \doteq (1/2m) \cdot \sum_i c_i (y_i \boldsymbol{\theta}^\top \hat{\boldsymbol{x}}_i)^2$. Here, we have assumed that there exists some $\Theta_* > 0$ such that $\|\boldsymbol{\theta}\|_2 \leq \Theta_*$; we shall see that such a bound Θ_* indeed exists for the $\boldsymbol{\theta}$ which interests us. Let

$$c' \doteq \frac{\sum_i c_i (y_i \boldsymbol{\theta}^\top \hat{\boldsymbol{x}}_i)^2}{\sum_i (y_i \boldsymbol{\theta}^\top \hat{\boldsymbol{x}}_i)^2} . \tag{139}$$

It trivially follows that $c' \in F''(\mathbb{I})$ and

$$J = \frac{c'}{2m} \cdot \sum_i (y_i \boldsymbol{\theta}^\top \hat{\boldsymbol{x}}_i)^2 . \tag{140}$$

What we thus get is that for any $\forall \hat{S}, \boldsymbol{\theta}$, there exists $c \in (1/2) \cdot F''(\mathbb{I}) \subseteq \mathbb{R}_+$ such that

$$L = F(0) + \frac{F'(0)}{m} \cdot \sum_i y_i \boldsymbol{\theta}^\top \hat{\boldsymbol{x}}_i + \frac{c}{m} \cdot \sum_i (y_i \boldsymbol{\theta}^\top \hat{\boldsymbol{x}}_i)^2 , \tag{141}$$

and we also observe that L is convex. We now consider the choice $c \doteq c^*$ obtained for

$$\boldsymbol{\theta}^* \doteq \arg\min_{\boldsymbol{\theta}} \ell_F(\hat{S}, \boldsymbol{\theta}; \Gamma_F) . \tag{142}$$

Let us denote ℓ_P the particular Taylor loss obtained, which therefore matches $\ell_F(\hat{S}, \boldsymbol{\theta}; \Gamma_F)$ for the choice $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. We now design the regularizer of the Taylor loss to ensure that its *optimum* is also achieved by $\boldsymbol{\theta}^*$. It is not hard to check that the optimum of the Ridge regularized Taylor loss $\ell_P(\hat{S}, \boldsymbol{\theta}; \Gamma_T)$, $\boldsymbol{\theta}^\circ$, satisfies:

$$c^* \hat{\mathsf{X}} \hat{\mathsf{X}}^\top \boldsymbol{\theta}^\circ + 2m\Gamma_T \boldsymbol{\theta}^\circ = -F'(0) \boldsymbol{\mu}_{\hat{S}} , \tag{143}$$

where $\boldsymbol{\mu}_{\hat{S}} \doteq \sum_i y_i \hat{\boldsymbol{x}}_i$ is the mean operator (Patrini et al., 2014). Let us find the equivalent expression for loss ℓ_F via a series of Taylor-Lagrange expansions, letting $z_i \doteq y_i \boldsymbol{\theta}^{*\top} \hat{\boldsymbol{x}}_i$ for short:

$$\forall i \in [m], \exists c'_i \in F''(\mathbb{I}) : F(z_i) = F(0) + F'(0) z_i + \frac{c'_i}{2} z_i^2 . \tag{144}$$

Define $\boldsymbol{c} \in \mathbb{R}_+^m$ the vector with $c_i \doteq c'_i/2$. It follows that because of eq. (142), $\boldsymbol{\theta}^*$ satisfies $\sum_i c_i (\boldsymbol{\theta}^{*\top} \hat{\boldsymbol{x}}_i) \hat{\boldsymbol{x}}_i + 2m\Gamma_F \boldsymbol{\theta}^* = -F'(0) \boldsymbol{\mu}_{\hat{S}}$, or more concisely,

$$\hat{\mathsf{X}} \mathrm{Diag}(\boldsymbol{c}) \hat{\mathsf{X}}^\top \boldsymbol{\theta}^* + 2m\Gamma_F \boldsymbol{\theta}^* = -F'(0) \boldsymbol{\mu}_{\hat{S}} . \tag{145}$$

Now, we want $\boldsymbol{\theta}^\circ = \boldsymbol{\theta}^*$, which imposes from eqs (143) and (145), $c^* \hat{\mathsf{X}} \hat{\mathsf{X}}^\top \boldsymbol{\theta}^* + 2m\Gamma_T \boldsymbol{\theta}^* = \hat{\mathsf{X}} \mathrm{Diag}(\boldsymbol{c}) \hat{\mathsf{X}}^\top \boldsymbol{\theta}^* + 2m\Gamma_F \boldsymbol{\theta}^*$, or equivalently, after simplifying,

$$\Gamma_T \boldsymbol{\theta}^* = \kappa \boldsymbol{\theta}^* + \Gamma_F \boldsymbol{\theta}^* , \tag{146}$$

where

$$\kappa \doteq \hat{\mathsf{X}} \left(\frac{1}{2m} (\mathrm{Diag}(\boldsymbol{c} - c^* \mathbf{1})) \right) \hat{\mathsf{X}}^\top \tag{147}$$

is symmetric but not necessarily positive definite. We clearly have $\boldsymbol{\theta}^\top \mathrm{K} \boldsymbol{\theta} \geq -\hat{X}_*^2 \sup F''(\mathbb{I})/2$ for any unit $\boldsymbol{\theta}$. So, if we fix

$$\Gamma_T \doteq \mathrm{K} + \Gamma_F \quad (148)$$

after picking Γ_F such that its smallest eigenvalue satisfies, for some fixed $\lambda^\circ > 0$,

$$\lambda^\circ(\Gamma_F) \geq \lambda^\circ + \frac{\hat{X}_*^2}{2} \sup F''(\mathbb{I}) \ , \quad (149)$$

then we shall have eq. (146) ensured with Γ_T symmetric positive definite with $\lambda^\circ(\Gamma_T) \geq \lambda^\circ$. We can also remark that eq. (145) yields, because $\hat{\mathsf{X}}\mathrm{Diag}(\boldsymbol{c})\hat{\mathsf{X}}^\top$ is positive semi-definite,

$$\|\boldsymbol{\theta}^*\|_2 \leq \frac{|F'(0)|\hat{X}_*}{2\lambda_1^\uparrow(\Gamma_F)} \ , \quad (150)$$

so we can posit $\Theta_* \doteq |F'(0)|\hat{X}_*/(2\lambda_1^\uparrow(\Gamma_F))$ and in fact we can pick

$$\mathbb{I} \doteq \frac{|F'(0)|\hat{X}_*^2}{2\lambda_1^\uparrow(\Gamma_F)} \cdot [-1, 1] \ . \quad (151)$$

For any finite $\lambda^\circ, \hat{X}_*$, let us define

$$\mathbb{J}(\lambda^\circ, \hat{X}_*) \doteq \left\{ z \in \mathbb{R} : z \geq \lambda^\circ + \frac{\hat{X}_*^2}{2} \sup F'' \left(\lim_{z' \to z} \frac{|F'(0)|\hat{X}_*^2}{2z'} \cdot [-1, 1] \right) \right\} \cap \mathbb{R}_+ \ . \quad (152)$$

Picking $\lambda^\circ(\Gamma_F)$ in $\mathbb{J}(\lambda^\circ, \hat{X}_*)$ guarantees that it satisfies eq. (149). Let us denote for short $\mathbb{J}'$ to be the leftmost set in the intersection in eq. (152). Because the argument of F'' is the same for any $\pm z$, if there exists any $z < 0$ in $\mathbb{J}'$, then $-z$ is also in $\mathbb{J}'$. We remark that because $|F'(0)| \ll \infty$, the argument set of $F''(.)$ converges to $\{0\}$ with $z \to \pm\infty$; since F'' is continuous and $|F''(0)| = F''(0) \ll \infty$ by assumption, we get that $\mathbb{J}'$ is non-empty, and so $\mathbb{J}' \cap \mathbb{R}_+$ is non-empty, thus

$$\mathbb{J}(\lambda^\circ, \hat{X}_*) \neq \emptyset \ . \quad (153)$$

So, let us define

$$\lambda^* \doteq \inf \mathbb{J}(\lambda^\circ, \hat{X}_*) \ (\geq 0) \ , \quad (154)$$

removing the dependence of λ^* in $\lambda^\circ, \hat{X}_*$ for clarity.

To summarize, for any $\lambda^\circ > 0$ and any Ridge regularized loss $\ell_F(\hat{S}, \boldsymbol{\theta}; \Gamma_F)$ satisfying $F \in C^2$, $|F'(0)|, F''(0) \ll \infty$ and $\lambda^\circ(\Gamma_F) \geq \lambda^*$ where λ^* is finite and defined in eq. (154), there exists a Taylor loss $\ell_P(\hat{S}, \boldsymbol{\theta}; \Gamma_T)$ such that

1. $\ell_F(\hat{S}, \boldsymbol{\theta}^*; \Gamma_F) = \ell_P(\hat{S}, \boldsymbol{\theta}^*; \Gamma_T)$ where $\boldsymbol{\theta}^* \doteq \arg\min_{\boldsymbol{\theta}} \ell_F(\hat{S}, \boldsymbol{\theta}; \Gamma_F)$;

2. $\arg\min_{\boldsymbol{\theta}} \ell_F(\hat{S}, \boldsymbol{\theta}; \Gamma_F) = \arg\min_{\boldsymbol{\theta}} \ell_P(\hat{S}, \boldsymbol{\theta}; \Gamma_T)$;

3. $\lambda^\circ(\Gamma_T) \geq \lambda^\circ$.

We also check that $a = F(0), b = F'(0)$, and we get the statement of the Theorem when F is convex.

Case 2 — F not convex. When F is not convex, we still have for any $\boldsymbol{\theta}^* \in \mathcal{C}$ because F is twice differentiable,

$$\ell_F(\hat{S}, \boldsymbol{\theta}^*; \Gamma_F) = \ell_F(\hat{S}, \mathbf{0}; \Gamma_F) + \boldsymbol{\theta}^{*\top} \nabla_{\boldsymbol{\theta}} \ell_F(\hat{S}, \boldsymbol{\theta}; \Gamma_F)_{|\boldsymbol{\theta}=\mathbf{0}} + \frac{1}{2} \cdot \boldsymbol{\theta}^{*\top} \nabla\nabla_{\boldsymbol{\theta}} \ell_F(\hat{S}, \boldsymbol{\theta}; \Gamma_F)_{|\boldsymbol{\theta}=\boldsymbol{u}}, \tag{155}$$

for some $\boldsymbol{u} = t \cdot \boldsymbol{\theta}^*$ with $t \in [0, 1]$, where $\nabla\nabla$ denote the Hessian, given by

$$\nabla\nabla_{\boldsymbol{\theta}} \ell_F(\hat{S}, \boldsymbol{\theta}; \Gamma_F)_{|\boldsymbol{\theta}=\boldsymbol{u}} = \sum_i F''(y_i \boldsymbol{u}^\top \hat{\boldsymbol{x}}_i) \cdot \hat{\boldsymbol{x}}_i \hat{\boldsymbol{x}}_i^\top + 2\Gamma_F \ , \tag{156}$$

positive semi-definite since $\boldsymbol{\theta}^* \in \mathcal{C}$. F being C^2, F'' being continuous, $\boldsymbol{\theta}^*$ is a local minimum of the loss in an open neighborhood $\mathcal{N}(\boldsymbol{\theta}^*)$ of $\boldsymbol{\theta}^*$. We still can build the equivalent Taylor loss and first its L part as in eq. (138). However, L is *not* necessarily convex this time. The Hessian of the Taylor loss *regularized* is now

$$\nabla\nabla_{\boldsymbol{\theta}} \ell_P(\hat{S}, \boldsymbol{\theta}; \Gamma_T) = c^* \sum_i \hat{\boldsymbol{x}}_i \hat{\boldsymbol{x}}_i^\top + 2\Gamma_T \ , \tag{157}$$

and so to obtain a convex regularized Taylor loss, it is sufficient to ensure, for some fixed $\lambda^\circ > 0$,

$$2\lambda^\circ(\Gamma_T) \geq \lambda^\circ + \hat{X}_*^2 \sup F''(\mathbb{I}) \ , \tag{158}$$

which is exactly ineq. (149) with its argument λ° halved. So, the regularized Taylor loss is in fact convex, and the only other modification is to now ensure $|F''(0)| \ll \infty$ since F can be concave in 0.

Remark: throughout all results, we have assumed without loss of generality that the mean operator $\boldsymbol{\mu}_{\hat{S}} \neq \mathbf{0}$, which implies, from eqs (143) and (145) that $\mathbf{0}$ cannot be a critical point of the losses.

I.5 Strength of the regularisation imposed on the loss

The set of losses we consider, which we define as regular symmetric proper losses (RSPL), are essentially proper, strictly convex and have no class-dependent misclassification costs Nock & Nielsen (2008, 2009). For any such loss there exists a *permissible* ψ such that $F \doteq F_\psi$ with

$$F_\psi(z) \doteq \frac{\psi(0) + \psi^\star(-z)}{\psi(0) - \psi(1/2)} \doteq a_\psi + \frac{\psi^\star(-z)}{b_\psi} \ , \tag{159}$$

where $\star$ is the convex conjugate (Nock & Nielsen, 2009). A permissible ψ satisfies $\mathrm{dom}(\psi) \supseteq [0, 1]$, ψ strictly convex, differentiable and symmetric with respect to $1/2$. We add the condition that ψ' is concave on $[0, 1/2]$ and denote this set of losses as *regular symmetric proper losses* (RSPL). Popular examples of RSPLs include the square, logistic and Matsushita losses (Nock & Nielsen, 2009), the square loss also being a Taylor loss.

Lemma P *If ℓ_F is* RSPL *in Theorem 6 (main file), we can pick $\lambda^* \doteq \lambda^\circ + F''(0)\hat{X}_*^2/2$, where $\hat{X}_* \doteq \max_i \|\hat{\boldsymbol{x}}_i\|_2$.*

Remark: as examples, $F''(0)$ is respectively $1/4$, $1/2$ for the logistic and Matsushita losses Nock & Nielsen (2009), which results in a relatively small value for λ^*.

Proof Since ψ is strictly convex differentiable, its convex conjugate is $\psi^\star(z) = z\psi'^{-1}(z) - \psi(\psi'^{-1}(z))$, from which we easily get $F''_\psi(z) = 1/(b_\psi\psi''(\psi'^{-1}(-z)))$. Because ψ' is concave on $[0, 1/2]$, ψ'' is decreasing on $[0, 1/2]$ and therefore increasing on $[1/2, 1]$, achieving its minimum for $\psi'^{-1}(-z) = 1/2$, which gives $-z = \psi'(1/2) = 0$ and $z = 0$ for the arg max of $F''_\psi(z)$. Hence, $\mathbb{J}(\lambda^\circ, \hat{X}_*)$ becomes more explicit:

$$\mathbb{J}(\lambda^\circ, \hat{X}_*) \doteq \left\{z \in \mathbb{R}_+ : z \geq \lambda^\circ + \frac{F''_\psi(0)\hat{X}_*^2}{2}\right\} , \tag{160}$$

so we can just pick

$$\lambda^* \doteq \lambda^\circ + \frac{F''_\psi(0)\hat{X}_*^2}{2} , \tag{161}$$

as claimed. ■

II Supplementary material on experiments

II.1 Detailed setting

We consider the setting in which there exists a small set of features that is present in both peers, A and B. We call them the *shared features* and use them for record linkage. This setting is realistic considering, for example, that many businesses or government bodies would share basic information about their customers (such as gender, postal code, age, contact number, etc.) (Patrini et al., 2016). We then put noise in those shared features as a slider to vary the hardness of the task. We adopt a simple noise injection process, inspired by thorough analyses in the area (Christen & Pudjijono, 2009). Let p be the noise probability. Each shared value is replaced with probability p by a *neighbor* in the feature's domain, *i.e.* if we assume a total order in the feature values (which is available for most: binary, real or ordinal), we replace with probability p the feature value by a *neighbor* in the order: if the feature is binary, then it is replaced by the other value; otherwise, we pick uniformly at random a value in the set of neighboring $\pm u$ indexes, clamped to the observed set of values — *i.e.* we do not generate unobserved feature values. If there are more than 20 recorded values for the feature, then $u = 10$; otherwise, $u = 2$. Such a neighbor noise process follows the observed pattern that errors in the real world often generate neighboring values, for a neighbor relationship that can belong to the phonetic, typographic, OCR or just keyboard spaces (Christen & Pudjijono, 2009).

II.2 Detailed algorithms and baselines for record linkage

The max-weighted matching problem and the GREEDY routine — there is a particularly interesting routine that we call GREEDY, which delivers a fast approximation to a problem that

Algorithm 1 Greedy($\mathcal{I}$)

```
Input: set $[m]^2 \supset \mathcal{I} \doteq \{(i_{\mathsf{A}}, i_{\mathsf{B}})\}$, where $i_{\mathsf{A}}$ (resp. $i_{\mathsf{B}}$) belongs to indexes of $\mathsf{A}$ (resp. $\mathsf{B}$)
$\mathcal{I}_{\mathrm{g}} \leftarrow \emptyset$
repeat
    let $(i^*_{\mathsf{A}}, i^*_{\mathsf{B}}) \doteq \arg\max_{(i,i') \in \mathcal{I}} \widetilde{\cos}(\mathrm{shared}(\boldsymbol{x}_{\mathsf{A}i}), \mathrm{shared}(\boldsymbol{x}_{\mathsf{B}i'}))$
    $\mathcal{I}_{\mathrm{g}} \leftarrow \mathcal{I}_{\mathrm{g}} \cup \{(i^*_{\mathsf{A}}, i^*_{\mathsf{B}})\}$
    delete $i^*_{\mathsf{A}}$ from $\mathcal{I}$
    delete $i^*_{\mathsf{B}}$ from $\mathcal{I}$
until $\mathcal{I} = \emptyset$
return $\mathcal{I}_{\mathrm{g}}$
```

generalizes ours for record linkage: maximum weighted matching for balanced bipartite graphs (Avis, 1983). The instance of this problem is a balanced complete bipartite graph with non-negative weights, a feasible solution is a subset of edges covering all vertices, in which each vertex appears once. The criterion to be maximized is the sum of weights. If we take as the total (sum-of) cosine similarity the criterion to be maximized for record linkage and note that maximizing the criterion for the cosine similarities is equivalent to maximizing the same criterion for (1+cosine similarity)*es*, which is non-negative, then Greedy, provided in Algorithm 1, provides a fast approximation to record linkage, namely $O(|\mathcal{I}|^2 \log |\mathcal{I}|)$ for a non-optimized implementation. Let us denote C^* the optimal value of the total cosine similarity. There exists a long-known method, the Hungarian algorithm, that provably achieves the optimum (Kuhn, 1955), yet it requires a significantly more sophisticated implementation to even reach $O(|\mathcal{I}|^3)$ time complexity. We stick to the greedy algorithm Greedy not just for computational reasons and its straightforwardness of implementation: Greedy does provide a guaranteed good constant approximation to C^*.

Lemma Q *Avis (1983) (Theorem 4) Let us denote C_{Greedy} as the total record linkage similarity retrieved by* Greedy. *Then $C_{\text{Greedy}} \geq C^*/2$.*

It is also believed that the actual worst-case approximation provided by Greedy is even better (Deligkas et al., 2017). In our experiments, we test and compare several algorithms for record linkage in various environments. We now present those algorithms.

B **does not use classes: GreedyRL** — In this case, peer B does not have the knowledge of classes and does not use the knowledge of classes for record linkage: linking proceeds from a straightforward use of routine Greedy, as explained in the boxed algorithm below, where $\mathcal{I} \doteq \{(i, i'), i \in [m], i' \in [m]\}$.

Algorithm GreedyRL($\mathcal{I}$) — Let

$$\mathcal{I}_{\mathrm{g}} \quad \leftarrow \quad \text{Greedy}(\mathcal{I}) \ . \tag{162}$$

Link all data following $\mathcal{I}_{\mathrm{g}}$, return $\hat{S}$.

B **has classes: GreedyRL+C**[2] — This approach can be implemented when both A and B have the knowledge of the true class for their respective observations, which is the setting of Patrini

[2] In the main file, we have merged this case with GreedyRL+$\tilde{\mathrm{C}}$ for the sake of saving space. In this SM, we chose

et al. (2016). The algorithm simply consists in running GREEDY over the positive class only, then GREEDY over the negative class only and finally linking the datasets according to the outputs of GREEDY. More formally, if we denote for short $S_{\mathsf{A}} \doteq \{(\boldsymbol{x}_{\mathsf{A}i}, y_{\mathsf{A}i}) : i = 1, 2, ..., m\}$ the sample from A, and $S_{\mathsf{B}} \doteq \{(\boldsymbol{x}_{\mathsf{B}i}, y_{\mathsf{B}i}) : i = 1, 2, ..., m\}$ the sample from B, then the algorithm can be summarized as follows, with $\mathcal{I}^+ \doteq \{(i, i') : y_{\mathsf{A}i} = y_{\mathsf{B}i'} = +1\}$ and $\mathcal{I}^- \doteq \{(i, i') : y_{\mathsf{A}i} = y_{\mathsf{B}i'} = -1\}$.

Algorithm GREEDYRL+C($\mathcal{I}^+, \mathcal{I}^-$) — Let

$$\mathcal{I}^+_{\mathrm{g}} \leftarrow \text{GREEDY}(\mathcal{I}^+) \ , \tag{163}$$

$$\mathcal{I}^-_{\mathrm{g}} \leftarrow \text{GREEDY}(\mathcal{I}^-) \ . \tag{164}$$

Link the datasets following $\mathcal{I}^+_{\mathrm{g}}$ and $\mathcal{I}^-_{\mathrm{g}}$, return $\hat{S}$.

B **does not have classes but *learns* them: GREEDYRL+$\overline{\text{C}}$** — In this case, peer B does not have the knowledge of classes but *computes* classes using a simple four-steps practical approach relying on shared features:

(i) we run GREEDY as in GREEDYRL and then *discard* couples in $\mathcal{I}_{\mathrm{g}}$ whose similarity is below the median similarity. We then assign a label to the observations of B still appearing in $\mathcal{I}_{\mathrm{g}}$, by using the correspondence with A in $\mathcal{I}_{\mathrm{g}}$;

(ii) to complete labelling in B, we use a simple k-NN algorithm inside B which gives a label to the remaining observations based on the labels computed from step (i) only. At this stage, all observations in B are given a class;

(iii) We then run GREEDYRL+C using the predicted classes for B. *Notice that we have no guarantee that the class proportions in* B *will be the same as in* A. For that reason, we end up in general with a subset of observations in A and B being not linked;

(iv) to complete linkage, we just run GREEDYRL in the subset of remaining observations.

The overall algorithm is sketched in the box below.

to keep them apart to clearly distingush between the case where there is no noise from the case where there is some noise in RL.

Algorithm GreedyRL+$\overline{\text{c}}$— Let $\mathcal{I} \doteq \{(i, i'), i \in [m], i' \in [m]\}$, and

$$\mathcal{I}_{\mathrm{g}} \leftarrow \textsc{Greedy}(\mathcal{I}) . \tag{165}$$

Let ς be the median similarity in $\mathcal{I}_{\mathrm{g}}$. Discard from $\mathcal{I}_{\mathrm{g}}$ all couples with similarity below ς and affect classes to observations of B using the remaining couples:

$$\forall (i, i') \in \mathcal{I}_{\mathrm{g}}, y_{\mathsf{B}i'} \leftarrow y_{\mathsf{A}i} . \tag{166}$$

Let $S^{\emptyset}_{\mathsf{B}}$ denote the subset of observations of B without a label, and $S^{\mathrm{c}}_{\mathsf{B}}$ denote the subset of observations of B with a label (the total set of observations of B is $S^{\emptyset}_{\mathsf{B}} \cup S^{\mathrm{c}}_{\mathsf{B}}$). Use a k-NN rule to give a label to observations from $S^{\emptyset}_{\mathsf{B}}$:

$$\forall \boldsymbol{x}_{\mathsf{B}i'} \in S^{\emptyset}_{\mathsf{B}}, y_{\mathsf{B}i'} \leftarrow k\text{-NN}(S^{\mathrm{c}}_{\mathsf{B}}) . \tag{167}$$

Let $\mathcal{I}^{+} \doteq \{(i, i') : y_{\mathsf{A}i} = y_{\mathsf{B}i'} = +1\}$ and $\mathcal{I}^{-} \doteq \{(i, i') : y_{\mathsf{A}i} = y_{\mathsf{B}i'} = -1\}$. Run GreedyRL+c($\mathcal{I}^{+}, \mathcal{I}^{-}$). Let $\mathcal{I}_{\mathsf{A}} \subseteq [m]$ and $\mathcal{I}_{\mathsf{B}} \subseteq [m]$ denote (indexes of) the subsets of observations not linked in A and B (we have $|\mathcal{I}_{\mathsf{A}}| = |\mathcal{I}_{\mathsf{B}}|$). Run GreedyRL($\mathcal{I}_{\mathsf{A}} \times \mathcal{I}_{\mathsf{B}}$), link all data, return $\hat{S}$.

B **has *noisy* classes: GreedyRL+$\tilde{\text{c}}$** — This corresponds to running GreedyRL+c in an environment where A has the knowledge of the true class but B has a knowledge of *noisy* classes. To conform with the vertical partition setting, we simulate permutation noise over classes in B by the following process: starting from setting GreedyRL+c / true classes, given a proportion p', we permute a random positive class and a random negative class for $[mp']$ iterations in B, where $[.]$ gives integer rounding. We then run GreedyRL+c as in the noise-free setting. To distinguish with the noise-free environment, we call this approach GreedyRL+$\tilde{\text{c}}$(p'). We consider $p' \in \{0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.15, 0.2\}$. Remark that $p' = 0.2$ can be considered a fairly large noise proportion.
‘Ideal’ — because we use simulated domains, we are able to compute the performances of the ideal record linkage. In this case, $\hat{S} = S$.

Algorithm ‘Ideal’ — return S.

‘Ideal’ gives our ‘optimal’ baseline to compare against the practical approaches to record linkage developed thereafter.

II.3 Domains

To have reliable baselines against which to compare our algorithms, we have used UCI domains (Blake et al., 1998) from which we have generated our distributed data using the following process: given a set of shared features, split randomly the remaining features between A and B. The shared features of B are then noisified using the process describes above in Subsection II.1. A always has access to the classes. Remark that since only the shared features of B are noisified, this guarantees

Domain	m	d	s	shared	linear correlations wrt class	C.Err
magic	19020	10	4	0, 1, 2, 3	$0.29, 0.25, 0.11, -0.02$	10^{-4}
page	5473	10	3	0, 1, 2	$-0.12, -0.03, -0.09$	1.83
sonar	208	60	3	0, 1, 2	$0.27, 0.23, 0.19$	3.69
winered	1599	11	2	7, 8	$-0.15, -0.003$	6.02
eeg	14980	14	4	0, 1, 2, 3	$0.01, -0.08, 0.04, -0.08$	6.08
phishing$_H$	11055	30	5	5, 6, 7, 13, 25	$0.34, 0.30, 0.71, 0.69, 0.34$	7.39
winewhite	4898	11	3	0, 1, 2	$-0.08, -0.21, -0.0007$	8.57
breast-wisc	699	9	2	0, 1	$-0.68, -0.78$	9.21
fertility	100	9	3	2, 3, 4	$-0.02, -0.09, 0.03$	12.22
banknote	1372	4	1	0	-0.72	13.14
creditcard	14599	23	4	1, 2, 3, 4	$-0.02, 0.01, -0.02, 0.004$	14.96
qsar	1055	41	4	2, 5, 8, 9	$-0.28, -0.16, -0.05, 0.16$	16.67
transfusion$_H$	748	4	1	0	-0.24	17.36
transfusion$_L$	748	4	1	3	-0.03	17.80
firmteacher	10800	16	2	0, 1, 2, 3, 4	$-0.22, 0.29, -0.25, 0.18, 0.10$	19.78
ionosphere	351	33	1	0	0.45	20.57
phishing$_L$	11055	30	4	0, 1, 2, 3	$0.09, 0.05, -0.06, 0.05$	24.35

Table A1: UCI domains used (Blake et al., 1998). For each domain, we indicate the total number of examples (m), total number of features (d) and the number of shared features used in our simulations (s). We then indicate the list of shared features (indexes as recorded in the UCI) and the list of linear correlations with the class for each of them. We finally indicate the average class errors in record linkage **for GREEDYRL** (C.Err), *i.e.* the proportion of examples from one class matched by GREEDYRL with examples from the other class. Domains are listed in increasing value of C.Err.

that the final observation matrix, $\hat{\mathsf{X}}$, obtained after record linkage indeed meets the decomposition

$$\hat{\mathsf{X}} \doteq \left[\begin{array}{c} \mathsf{X}_{\mathsf{A}} \\ \hline \mathsf{X}_{\mathsf{B}}\hat{\mathsf{P}} \end{array}\right] \quad (168)$$

for some unknown $\hat{\mathsf{P}}$, where the shared features in the matrix are those of A. This guarantees that the differences between learning algorithms are not due to the (variable) effect of noise in features but to the errors of $\hat{\mathsf{P}}$ following mistakes in record linkage. Table A1 presents the domains we have used. For two of them (phishing, transfusion), we have considered two versions, one in which the shared attributes are highly correlated with the class (H) and one in which they are not (L). Remark that the proportion of linkage errors between classes ranges between ≈ 0 to $\approx 25\%$, which is very significant, and the proportion of shared features among all ranges from $\approx 3\%$ to $> 30\%$.

II.4 The class is key to optimizing record linkage

Results are displayed in Table A2. From those results, several observations come to the fore. First, the larger the number of errors of record linkage among classes for GREEDYRL (Table A1, C.Err), the more beneficial are the approaches using the class information for record linkage. On domains firmteacher, ionosphere, phishing, using the class information is almost always on par with or (significantly) better than GreedyRL. Second, the improvement can be extremely significant as witnessed by domains creditcard or firmteacher, with almost 20 % improvement when using (even noisy) classes on creditcard, and still up to 6% improvement when using predicted classes (GreedyRL+$\overline{\mathrm{c}}$) on creditcard. This is very good news because the shared features we used on creditcard — sex, education, marriage, age — are typically those that would be shared in a federated learning setting.

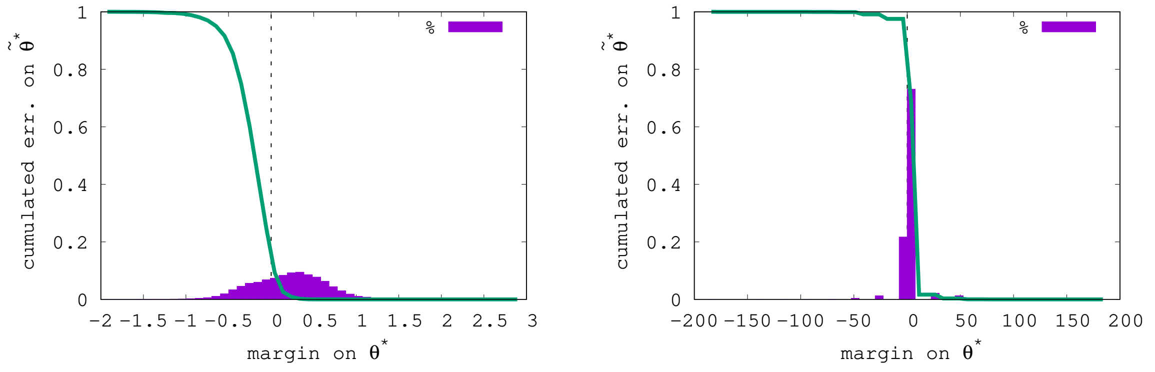
Another observation may be made: on all domains but one (banknote), carrying record linkage is susceptible to compete against 'Ideal'. On the majority of domains, there exists a version of GreedyRL[as is — +$\overline{\mathrm{c}}$ — +c — +$\tilde{\mathrm{c}}$] which *beats* 'Ideal' — even when not statistically in most cases —. On few domains, page, sonar, transfusion (both H and L), using class information yields results that almost always beat the 'Ideal' baseline. One explanation comes from the fact that all learners, including 'Ideal', use AdaBoost for a limited number of iterations. On these domains, the models learned after record linkage tend to be slightly less sparse than for 'Ideal'. So, it seems reasonable that record linkage, when carefully used as may be the case with class information, may force the spread of AdaBoost's feature leveraging to a larger number of relevant features, compared to 'Ideal' which focuses on a smaller set during the thousand iterations allocated and thus comes up with a model that can less accurate. Also, considering phishing, we see that having shared features that are more correlated with the class (phishing$_H$ vs phishing$_L$) certainly helps to compete against 'Ideal', in particular when one peer does not have classes. The same observation can be made for transfusion, even when the gains with more correlated features are less important in this case, which can be due to the small number of shared features.

If we now compare the two approaches of GreedyRL using class information (with classes, even noisy, vs without), then it is apparent that having noisy classes — with up to 20% noise — can very significantly help against GreedyRL compared to carrying out record linkage without ground class information (but learning classes) as in GreedyRL+$\overline{\mathrm{c}}$. Our approach that learns classes in GreedyRL+$\overline{\mathrm{c}}$ is simple but still manages to deliver significant improvements in some cases, typically high noise for shared features (winewhite, creditcard) or shared features sufficiently correlated with class (transfusion$_H$).

Finally, we keep in mind that these results are obtained for simulations that include in general a small number of shared features (2.8 on average) and a shared feature noise that ranges up to p =30%, which would correspond to relatively challenging practical settings. This suggests that if we exclude pathological domains like banknote in our benchmark, there would be for most domains good reasons to carry out tailored approaches to record linkage for learning with the ambition to challenge the unknown learner having access to the ideally linked data. This is not surprising: it is known that the sufficient statistics for the class is very simple for many relevant losses (Patrini et al., 2014), so we should not expect perfect record linkage to be necessary to improve learning performance.

II.5 Observation of immunity of large margin classification to record linkage mistakes

In Section 3 (main file), we essentially show that all examples receiving large margin classification on $\boldsymbol{\theta}_0^*$ are given the right class by $\boldsymbol{\theta}_T^*$. To our knowledge, such a result has never been documented,



winewhite, GreedyRL+$\overline{\text{C}}$(5)

creditcard, GreedyRL+$\overline{\text{C}}$(10)

Figure 3: Margin distribution on two domains with shared attribute noise $p = 0.3$. The magenta histogram displays the distribution of margins of $\boldsymbol{\theta}^*= \boldsymbol{\theta}^*_0$ on training. The green curve is the cumulated relative error of $\tilde{\boldsymbol{\theta}}^*= \boldsymbol{\theta}^*_T$ *above* some margin x. For example, on winewhite, less than 20% of the errors on training happen on examples with positive margin, and approximately *no* error happens on examples with positive margin above 0.5 — in other words, *all* examples with margin above 0.5 on $\boldsymbol{\theta}^*_0$ receive the right class from $\boldsymbol{\theta}^*_T$ and so, following Definition 4, $\boldsymbol{\theta}^*_T$ happens to be **immune** to record linkage at margin 0.5. Since the maximal margin recorded for $\boldsymbol{\theta}^*_0$ is ≈ 3.0, we see in this example that immunity occurs for a comparatively small positive margin (best viewed in color, see text for details).

even experimentally, but it would represent a significant support for federated learning since one can hope, by joining diverse databases, to increase not just the accuracy of classifiers but in fact the optimal margins over examples, thereby bringing immunity to the mistakes of record linkage for examples that would attain sufficiently large margins. But how ‘large’ a margin is necessary? On each domain, we have computed the margin distributions of $\boldsymbol{\theta}^*_0$ — approximated by the output of AdaBoost ran on the training sample S for twice the usual number of iterations, that is, 2000 (we do this for all cross validation folds). We then compute, for all examples, whether they are given the right class by $\boldsymbol{\theta}^*_T$. We finally compute the cumulative error distribution, in between 0 and 1, of $\boldsymbol{\theta}^*_T$. For any $x \in [\kappa_m, \kappa_M]$ (the interval of observed margins), the cumulative error on x is just the proportion of errors occurring for margins in the interval $[x, \kappa_M]$. When $x = \kappa_m$, this is just 1. Figure 3 provides two examples of curves obtained, which does not just validate immunity: on winewhite, it shows that it can happen for a quite small margin (≈ 0.5) with respect to the maximal margin ($\kappa_M \approx 3.0$), which reinforces the support for federated learning. On creditcard, we have $\kappa_M \approx 188$ while immunity happens at margin ≈ 100. Less than 1% of mistakes have margin larger than 30.

Finally, in table A3, we provide the minimal immunity margin on one domain for GreedyRL+$\tilde{\text{C}}$, for different values of p', that is, the minimal κ for which there is no error on examples with margin $\geq \kappa$ on $\boldsymbol{\theta}^*_0$. We can see that this minimal margin largely increases with noise, and so increasing noise in the record linkage process degrades the margin picture, which is also consistent with the fact that the error of $\boldsymbol{\theta}^*_T$ also significantly increases.

II.6 RL impact on ML via ξ or T

According to Theorem 3 (main file), the size of P (T) and parameter ξ should be a reasonable handle on how the training and test errors behave. For several datasets, we have extensively computed an upperbound of ξ which amounts to consider $\varepsilon = 0$ in (6). In this case, we get $\xi \leq B$ for

$$B \doteq \max_t \max\{\|\boldsymbol{a}_t - \boldsymbol{b}_t\|_2, \|\boldsymbol{a}'_t - \boldsymbol{b}'_t\|_2\}, \tag{169}$$

where $\boldsymbol{a}_t, \boldsymbol{b}_t, \boldsymbol{a}'_t, \boldsymbol{b}'_t$ are the four observations involved in P_t (Section 3, main file), and the size of P_t is (under)estimated using the linear time algorithm before (6) (main file). If our theory holds, then diagrams giving errors as a function of ξ or T should have a commonpoint: the spread of the error(s) should be more important as we move alongside the x axis, remembering that Theorem 3 gives one-sided inequalities (*i.e.* not lowerbounds). If we were to plot results, we then should reasonably expect datapoint to spread in a cone whose apex should appear at the left of the plot, among the smallest values of ξ or T. Tables A4 and A5 summarize a series of results for a subset of domains and a subset of algorithms considered, on which we have approximated the corresponding cone for breast-wisc (ξ), displaying that it indeed does fit the formal shape that the theory suggests. The general cone pattern is clearly visible for most domains, for both ξ and T *and* on training and testing. Furthermore, on almost all domains, a larger amount of noise clearly 'pushes' the results towards the right of the plot, *i.e.* towards the degradation of the parameter ξ or T. That it also fits test errors is not unexpected but given the uncertainty on estimating test errors, the accuracy of the fit and picture on testing is good news. Indeed, it suggests that it could be possible to estimate the impact of record linkage on *testing* ML models, which is quite downstream the pipeline considering how RL is upstream in the process. That such criteria could survive not just one (learning ML model) but two (evaluating the ML model on test data) major sources of uncertainty is good news. Our plots suggest it is crucial to estimate the range of parameters ξ or T – they can fluctuate very significantly depending on the domain. Provided this can be done and good estimation is available for ξ or T, one could not just be able to understand where the current ML model stands, but eventually where could any model stand by fitting this conic region. Some particular cases emerge from our experiments: domain ionosphere does not display the expected shape for ξ, but this could be due to the exceptional small range of ξ of $\approx .1$ width. The plots for T clearly display the expected shape. Domain page display, for a few runs, an improvement of the test error as p' increases. However, this appears for a moderate value of the RL noise ($p = 0.1$) for a domain in which greedy RL performs already very well without using the class information (Table A1).

References

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

Bhatia, R. *Matrix Analysis*. Springer, 1997.

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

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

Deligkas, A., Mertzios, G.-B., and Spirakis, P.-G. The computational complexity of weighted greedy matching. In *AAAI'17*, pp. 466–474, 2017.

Kuhn, H.-W. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83—-97, 1955.

Nock, R. and Nielsen, F. On the efficient minimization of classification-calibrated surrogates. In *NIPS*21*, pp. 1201–1208, 2008.

Nock, R. and Nielsen, F. Bregman divergences and surrogates for learning. *IEEE Trans.PAMI*, 31: 2048–2059, 2009.

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

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

Domain	Noise p	'Ideal'	GreedyRL[as is — $+\overline{\text{C}}$ — +C — $+\tilde{\text{C}}$]														#
			as is	$+\overline{\text{C}}(k)$				+C	$+\tilde{\text{C}}(p')$								beats
				1	2	5	10		0.01	0.02	0.03	0.04	0.05	0.10	0.15	0.20	ideal
magic	0.05	21.14	21.15	21.08	21.08	21.06	21.11	21.15	21.04	21.10	21.08	21.19	21.21	21.65	***22.28**	***22.70**	7
	0.1	21.14	21.19	21.42	***21.53**	21.21	21.17	21.08	21.21	21.08	21.17	21.18	21.33	***21.82**	***22.57**	***23.58**	2
	0.3	21.16	21.14	21.26	21.33	21.14	21.16	21.14	21.06	21.16	21.24	***21.62**	***21.72**	****22.34**	****23.75**	****25.31**	4
page	0.05	27.62	25.16	25.31	25.36	25.31	25.14	25.85	25.82	***27.26**	***27.28**	***27.66**	***28.02**	***29.23**	***30.31**	***31.04**	9
	0.1	27.17	26.11	26.03	25.43	***24.65**	***24.61**	25.63	26.63	26.31	26.55	26.48	27.39	27.96	***31.32**	***34.59**	10
	0.3	27.66	24.83	***26.43**	26.18	25.79	25.14	***25.87**	25.67	26.24	26.49	26.82	26.65	28.10	29.21	33.44	11
sonar	0.05	26.93	25.95	25.00	28.31	24.95	26.40	26.92	24.50	25.45	22.55	24.05	24.48	23.07	23.05	25.90	13
	0.1	26.88	26.45	25.45	28.38	24.52	27.45	26.88	24.98	27.40	25.48	26.50	28.40	25.02	29.33	25.55	8
	0.3	26.02	25.05	26.45	24.55	24.98	25.02	25.50	23.59	22.14	22.62	24.07	24.55	25.50	27.02	26.90	11
winered	0.05	26.08	26.58	26.20	26.32	26.45	26.58	26.77	26.45	26.83	26.20	26.45	26.33	26.64	28.20	***28.77**	0
	0.1	26.57	26.76	27.26	27.26	26.82	27.57	27.01	26.82	26.82	26.76	26.39	26.88	26.76	27.95	28.64	0
	0.3	26.58	27.58	27.01	27.20	27.64	26.89	26.89	26.83	26.70	27.01	26.64	26.83	26.76	26.82	27.08	0
eeg	0.05	45.18	45.05	44.43	44.99	43.88	44.16	45.16	45.29	45.59	45.72	45.62	45.84	46.46	46.47	46.50	6
	0.1	45.79	45.92	45.47	45.61	45.20	45.80	45.79	45.83	46.03	45.74	46.19	46.45	43.96	***43.42**	43.99	7
	0.3	45.19	46.10	46.08	46.58	46.68	46.60	45.18	45.40	45.45	45.96	45.02	45.61	45.07	45.07	45.04	5
phishing$_H$	0.05	8.03	8.40	8.08	8.18	8.70	8.62	8.17	8.05	8.05	8.14	8.23	***8.95**	***10.84**	****12.85**	****15.41**	0
	0.1	7.92	8.35	8.23	8.16	8.36	8.51	7.92	***7.76**	8.01	8.21	8.72	8.76	***9.61**	****12.86**	****15.29**	1
	0.3	7.96	8.90	9.15	8.91	9.01	8.86	8.39	***8.09**	***8.17**	8.46	8.67	9.05	***11.49**	****13.5**	****15.46**	0
winewhite	0.05	30.58	30.34	30.30	30.28	30.36	30.31	30.60	30.65	30.65	30.69	30.65	30.85	30.30	30.69	30.30	7
	0.1	30.99	31.11	30.91	30.97	30.97	30.85	30.75	30.95	31.03	31.26	31.03	31.11	31.58	31.60	31.95	6
	0.3	30.95	32.79	***31.44**	***31.36**	***31.31**	***31.16**	***30.97**	***31.19**	***31.17**	***30.89**	***31.07**	***30.99**	***31.48**	32.01	32.97	0
breast-wisc	0.05	3.00	3.71	3.43	3.57	3.57	3.43	***2.43**	2.57	3.29	3.29	3.14	3.57	3.43	3.57	3.00	2
	0.1	3.00	3.86	3.86	3.43	3.29	4.29	3.15	3.29	3.00	3.29	3.00	3.72	4.01	4.29	5.30	0
	0.3	2.71	6.29	5.58	5.72	5.57	***4.86**	***3.28**	***3.43**	***3.43**	***3.85**	***3.86**	***4.14**	***4.43**	4.86	6.01	0
fertility	0.05	43.00	49.00	43.00	33.00	41.00	44.00	42.00	43.00	52.00	50.00	52.00	52.00	43.00	48.00	55.00	3
	0.1	43.00	41.00	41.00	47.00	42.00	52.00	45.00	44.00	44.00	50.00	47.00	44.00	53.00	47.00	55.00	3
	0.3	46.00	50.00	55.00	58.00	49.00	59.00	49.00	49.00	54.00	50.00	54.00	54.00	53.00	55.00	43.00	0
banknote	0.05	2.77	13.26	13.71	13.92	12.83	13.92	***7.95**	***7.65**	***7.43**	***7.80**	***7.72**	***8.31**	***9.98**	12.82	14.93	0
	0.1	2.77	14.94	14.79	14.50	15.23	14.79	***11.88**	***11.51**	12.68	***12.31**	12.53	13.63	14.72	16.25	17.27	0
	0.3	2.91	12.89	13.84	12.74	12.39	12.97	10.06	10.64	11.15	10.78	11.73	12.03	13.55	14.69	16.91	0
creditcard	0.05	23.26	23.26	23.26	23.26	23.26	23.26	23.26	23.26	23.26	23.26	23.26	23.26	23.26	23.26	23.26	0
	0.1	23.26	41.87	40.66	41.46	***36.91**	36.89	****23.26**	****23.26**	****23.26**	****23.26**	****23.26**	***26.19**	42.65	43.08	44.36	0
	0.3	23.26	42.49	41.19	42.03	***38.82**	***36.51**	****23.26**	****23.26**	****23.26**	***24.72**	***25.01**	***32.28**	40.87	41.75	40.89	0
qsar	0.05	21.80	23.51	23.60	22.94	23.03	24.17	21.62	21.90	21.72	21.72	22.19	22.19	22.28	22.38	23.51	3
	0.1	21.51	23.22	23.02	23.40	23.78	23.31	22.27	21.79	21.70	21.99	21.79	21.89	22.36	22.75	22.75	0
	0.3	21.81	22.85	23.13	22.19	23.13	22.27	22.19	22.28	22.28	21.81	21.71	21.90	22.00	22.66	22.76	1
transfusion$_H$	0.05	39.57	36.10	***33.82**	***34.09**	***34.89**	***34.36**	***39.03**	***39.84**	***39.71**	***38.77**	***38.64**	37.44	***34.89**	***33.43**	35.03	12
	0.1	39.72	35.83	36.09	35.95	***33.55**	***33.28**	***40.38**	***38.92**	***37.57**	36.89	34.88	34.89	***33.16**	***33.82**	34.63	13
	0.3	38.37	35.55	35.83	34.08	34.89	34.89	***38.65**	***37.98**	***37.44**	37.04	37.17	35.97	35.62	35.03	34.89	13
transfusion$_L$	0.05	38.64	34.65	34.77	34.76	34.22	34.89	38.23	36.90	37.43	34.90	33.96	34.23	33.17	34.63	33.57	14
	0.1	39.02	35.15	34.75	35.29	35.16	34.48	37.16	38.09	38.09	37.17	36.36	33.68	33.55	33.55	33.41	14
	0.3	39.29	34.09	33.41	35.16	34.76	34.63	***39.29**	***38.77**	***37.82**	35.82	***37.02**	35.82	35.56	34.08	32.48	13
firmteacher	0.05	12.45	17.57	18.03	18.23	17.75	18.00	****12.71**	****12.68**	****12.71**	***13.06**	****13.02**	****13.35**	***14.81**	***15.90**	17.38	0
	0.1	12.39	21.03	21.06	21.29	21.51	21.54	****12.89**	****12.71**	****12.73**	****12.72**	****13.14**	****13.36**	****14.82**	***16.98**	***18.06**	0
	0.3	12.35	20.45	***21.12**	***21.16**	20.32	20.34	****12.54**	****12.45**	****12.42**	****12.73**	****12.81**	****13.00**	****14.54**	****16.05**	****17.44**	0
ionosphere	0.05	11.95	19.04	19.34	19.61	20.19	19.34	14.51	14.23	16.49	14.79	16.48	14.80	17.63	19.35	19.35	0
	0.1	10.28	16.25	14.54	15.13	15.96	16.54	13.42	14.56	15.41	15.68	15.39	15.68	15.39	15.97	16.26	0
	0.3	10.84	17.95	19.38	***22.50**	20.23	17.93	***13.97**	***13.40**	13.70	14.55	15.98	17.38	15.68	18.24	18.23	0
phishing$_L$	0.05	7.97	14.80	14.82	14.99	15.02	14.98	****7.91**	****8.27**	****8.44**	****8.45**	****8.61**	****8.83**	****9.94**	****10.16**	****11.18**	1
	0.1	7.89	11.11	11.11	11.11	11.11	11.11	****8.02**	****7.92**	****7.82**	****7.82**	****7.91**	****8.11**	****8.50**	****9.32**	***10.65**	2
	0.3	7.91	13.73	13.73	13.73	13.73	13.73	****8.29**	****8.51**	****8.47**	****8.44**	****8.60**	****8.80**	****9.16**	****9.81**	****10.54**	0

Table A2: Results (test errors) comparing, for three values of the shared features noise (p), the various approaches built on top of GreedyRL to 'Ideal'. Domains are listed in the same order as in Table A1. Grey shaded cells are the results of 'Ideal' and GreedyRL (indicated 'as is'). Red text denote results that are statistically outperformed by GreedyRL; green text denote results of GreedyRL[$+\overline{\text{C}}$ — +C — $+\tilde{\text{C}}$] statistically better than greedyER. One star ($*$) indicated p-value in $(10^{-6}, 10^{-2}]$, two stars ($**$) indicated p-value $\leq 10^{-6}$ (best viewed in color). The rightmost column ('# beats ideal') records the number of columns in which a version of GreedyRL is better (but not necessarily statistically better) than 'Ideal' (max = 14).

$p' = 0$	$p' = 0.01$	$p' = 0.02$	$p' = 0.03$	$p' = 0.04$	$p' = 0.05$	$p' = 0.1$	$p' = 0.15$	$p' = 0.2$
0.068	0.086	0.218	0.359	0.362	0.513	0.891	1.113	0.913

Table A3: Minimal immunity margin on domain magic ($p = 0.3$) for GreedyRL$+\tilde{\text{C}}$.

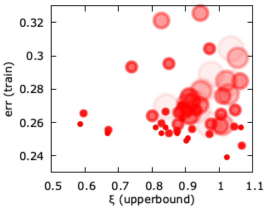
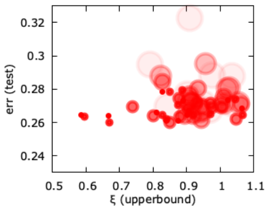
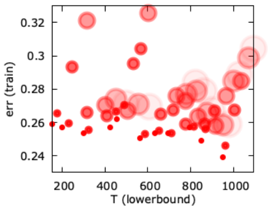
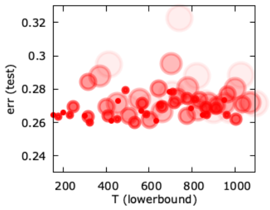
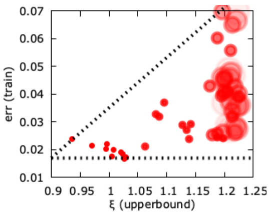
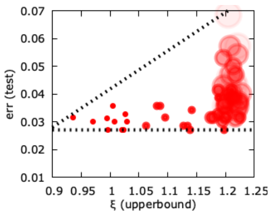
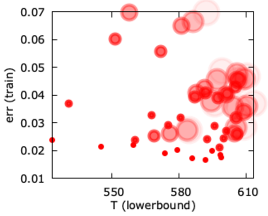
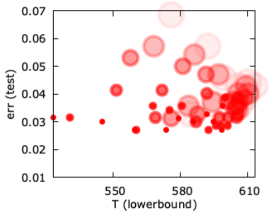
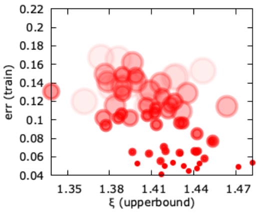
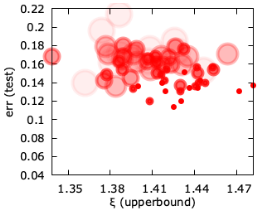
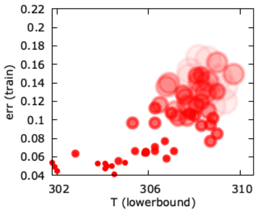
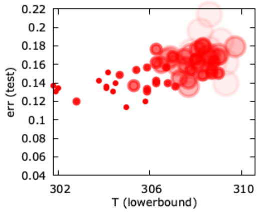
$x =$	ξ (upperbound)		T (lowerbound)	
$y =$	Train err.	Test err.	Train err.	Test err.
winered				
breast-wisc				
ionosphere				

Table A4: Train and test errors as a function of an upperbound on ξ or lowerbound on the permutation size T, for GREEDYRL+$\tilde{\text{C}}(p')$, for p' ranging in between 0 (GREEDYRL+C above) and $.5$. The size of the disks is correlated with the amount of noise p'. See text for the meaning of dashed axes on breasc-wisc.

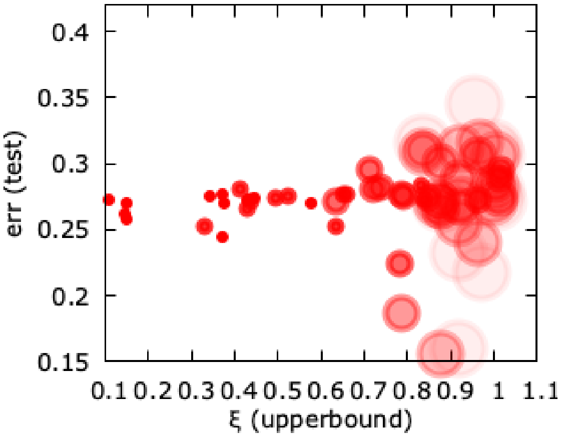
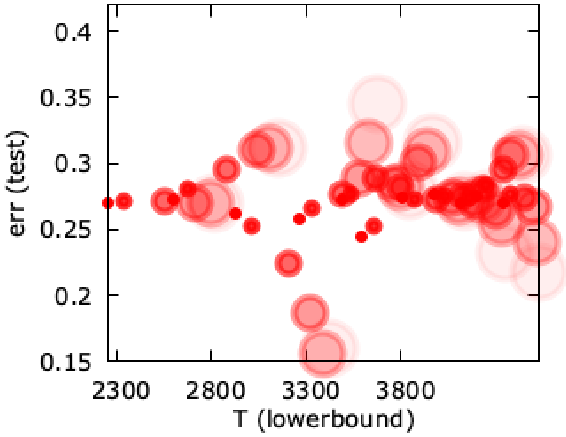
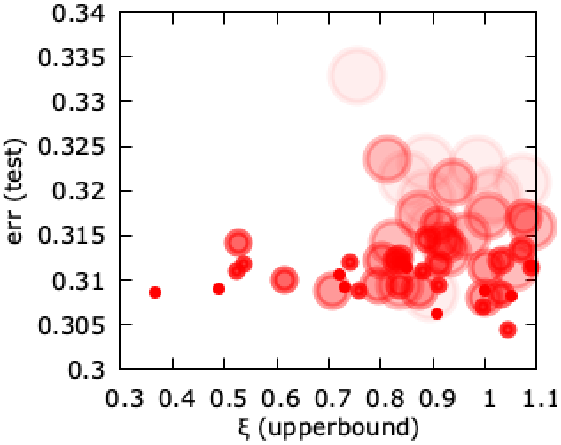
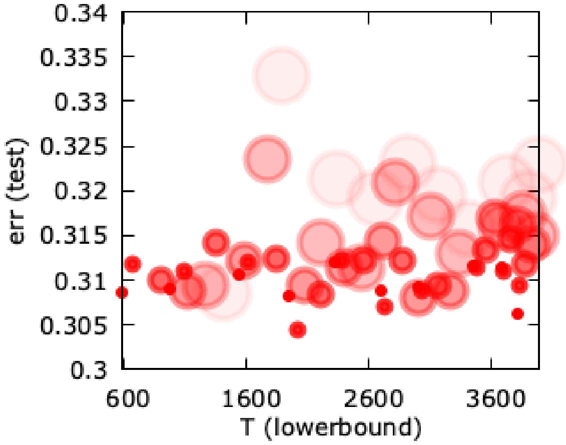
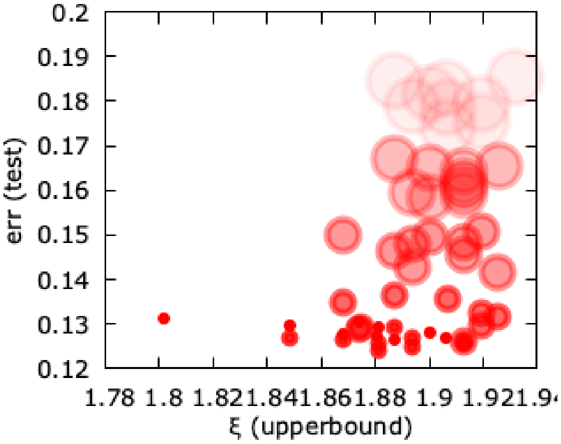
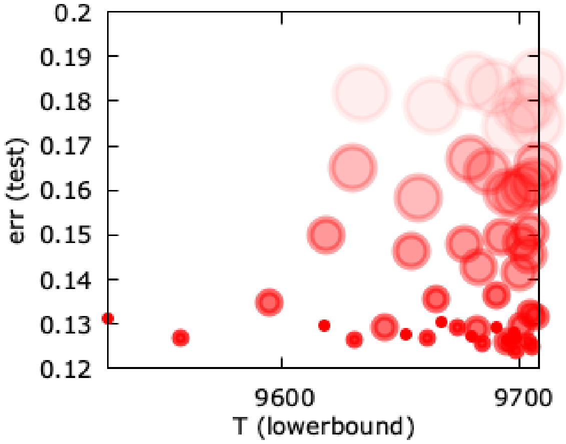
ξ (upperbound)	T (lowerbound)
	
page	
	
winewhite	
	
firmteacher	

Table A5: More results on *test errors*, conventions follow Table A4.