EXTREMELY PRIVATE SUPERVISED LEARNING

Armand Lacombe*, **Saumya Jetley*** **& Michèle Sebag**
AO/TAU, CNRS - INRIA - LISN - Univ. Paris-Saclay
Gif-sur-Yvette, France
{lacombe,jetley,sebag}@lri.fr

ABSTRACT

This paper presents a new approach called EXPRIL for learning from extremely private data. Iteratively, the learner supplies a candidate hypothesis and the data owner only releases the marginals of the error incurred by the hypothesis. Using the marginals as supervisory signal, the goal is to learn a hypothesis that fits the target data as best as possible. The privacy of the mechanism is provably enforced, assuming that the overall number of iterations is known in advance.

1 INTRODUCTION

In quite a few sensitive domains, the data owner is not willing to divulge any of their data, referred to as *target data*. Some approaches, aimed to learn from (very) limited information about the target data, have been proposed in generative modelling, privacy-preserving learning, and domain adaptation (more in section 1). All these approaches, to our best knowledge, assume that the learner has access to target data. The novelty of the proposed approach, called EXPRIL, is to only require some access to the marginals of the target data.

Formally, EXPRIL relies on two assumptions (section 2.1). Firstly, an (unlabelled) source dataset is assumed to be available, as an i.i.d. sample drawn after a source distribution overlapping with the target data distribution. Secondly, the data owner, referred to as *Oracle* in the following, is willing to provide i) the marginal distribution of the target data; ii) the marginals of the errors committed by any submitted candidate hypothesis on the target data, under the ϵ-privacy requirements Dwork et al. (2006). EXPRIL first estimates the importance weights associated with each source sample, in order to match the marginal distribution of the target data. It thereafter iteratively builds a sequence of candidate hypotheses, and uses their marginal errors to estimate the target label associated to each sample.

This scheme can be viewed as an active learning scheme Balcan & Feldman (2013), with two differences. Firstly, the proposed learner asks for the fraction of errors attached to a bag of unknown target samples, whereas an active learning asks for the label attached to a known target sample. Secondly, an active learner selects the most informative known sample for its query, whereas the proposed approach only queries the oracle based on its current best hypothesis.

On the one hand, the target data never leave the data owner, e.g. the hospitals. On the other hand, the differential privacy of the mechanism Dwork et al. (2006) can be established through perturbing the marginals supplied by *Oracle* by addition of Laplacian noise. The empirical validation of the approach shows that the EXPRIL differential privacy is obtained with a moderate loss of predictive accuracy for medium-dimensional problems (section 3.3).

RELATED WORK

The presented approach is at the crossroad of privacy preservation, generative modelling, and domain adaptation.

Privacy-preserving approaches A variety of real-world applications such as healthcare, customer analytics, financial reporting restrict the access to true data owing to privacy concerns. k-anonymisation for private data publishing works by blending a data point with $k-1$ (nearest)

*equal contribution

points to secure privacy, but suffers from attribute disclosure through homogeneity and inference attacks Machanavajjhala et al. (2007); the problem becoming worse for high-dimensional data Aggarwal (2005). Epsilon-differential privacy (ε-DP) is satisfied when any statistical query addressed to the dataset, with or without the inclusion of any particular data entry, yields outputs that are ϵ close Dwork et al. (2006). Thus, no adversary would be able to tell whether the dataset queried contains a particular data entry or not. The ε-DP definition has prompted a series of work on modelling synthetic datasets after a given target dataset through differentially private queries pertaining to data descriptions such as edge structure, conditional marginals Zhang et al. (2017b); Ping et al. (2017). The work most related to the presented approach is that of Zhang et al. (2017b), with the difference that our queries are simpler and only involve marginals along different feature axes.

Generative modelling Generative Adversarial Networks (GANs) Goodfellow et al. (2014) have been demonstrated to learn probability distributions of data living in high-dimensional spaces such that sampling from this distribution provides an artificial dataset, that is viewed as a synthetic version of the true data. GANs are being heralded as a privacy-preserving solution to making sensitive data available in the public domain Jordon et al. (2019); Yale et al. (2020). However, deep neural networks are prone to data memorisation Zhang et al. (2017a); the potential data leak among the target and the synthetic datasets is measured by the so-called *privacy loss* metric in ***Health-GAN***, which measures the relative resemblance of the synthetic data w.r.t the real training and test target data respectively. In particular, the synthetic data distribution must be indistinguishable from both the training and test distributions for the privacy loss to be 0. While this metric (indirectly) captures the dissimilarity between synthetic and real samples, critiques Stadler et al. (2020) question the alignment of this dissimilarity with formal/legal notions of privacy (i.e. how dissimilar is dissimilar enough?).

Along a different line, Papernot et al. (2017) propose learning a classification function through differentially private queries on an ensemble of teacher networks trained on disjoint subsets of a sensitive target dataset. The discriminators in Jordon et al. (2019) are modelled after the teacher architectures in Papernot et al. (2017) to guarantee differential privacy. All above methods depend on access to the target data, something that is unavailable to us.

Domain adaptation In domain adaptation (DA) Ben-David et al. (2007); Courty et al. (2014), two different distributions $P_s(X, Y)$ and $P_t(X, Y)$ with X the sample description and Y its label, respectively referred to as source and target distributions, are considered. DA, usually assuming that $P_s(Y|X)$ is not too different from $P_t(Y|X)$, aims to exploit the wealth of source data to build a better model on the target domain, usually including little data and even less labels. DA approaches often rely on designing embeddings, mapping source and target instances on some latent space, such that i/ this embedding preserves the discriminant information on the source data; ii/ it mixes the images of the source and target instances in such a way that the lack of target information is mitigated Ganin et al. (2016).
Extreme Domain Adaptation[1] exploits the source data together with minimal cues about the target data, expressed as the marginals of the label ($P(Y|X_i)$ with X_i a single descriptive feature). The model built from the source data is adapted to match the marginals. Another related approach is that of Hebert-Johnson et al. (2018), where a model is likewise adapted to achieve a calibrated prediction on all identifiable sub-groups within a given population.

2 THE EXPRIL ALGORITHM

EXPRIL achieves extremely private supervised learning. The contribution of the approach is that it only requires the target data to be known from its marginals. In the following, we restrict ourselves to the binary label case and to d-dimensional instance spaces.

Let $\mathcal{D}_s = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ denote the source dataset, with $\mathbf{x}_i \in \mathbb{R}^d$. The domain of each feature is partitioned in q bins. Let $\mathcal{B} = \{B_1, \ldots B_K\}$ denote the set of bins. Note that the total number of bins $K = qd$ linearly grows with dimension d.

[1] Uri Shalit, talk at ELLIS 2020

2.1 OVERVIEW

Formally, EXPRIL relies on two assumptions:
1. The **target dataset** is drawn after some distribution P_t on the source domain, satisfying the overlap assumption w.r.t the source distribution ($P_t(A) > 0 \implies P_s(A) > 0$ with P_s the source distribution for all $A \subset \mathbb{R}^d$ subset of the instance space). No labelling of the source dataset is required.
2. The interaction with the *Oracle* provides the marginals of i) the distribution $P_t(X)$, that is, the mass contained in a feature bin; ii) the error of any candidate hypothesis submitted by EXPRIL.

EXPRIL iteratively addresses two subproblems: *source reweighting* (referred to as Pb. 1) and *label estimation* (Pb. 2).

Pb. 1 aims to associate an importance weight to each source data sample, such that the weighted source dataset matches to the best possible extent the target marginals provided by the oracle in response to the first query. Pb. 2 aims to estimate the (target) label associated to each source data sample, based on the error marginals provided by the oracle in response to each submitted candidate hypothesis. Along the differential privacy protocol all the marginals supplied by the *Oracle* are perturbed using Laplacian noise of adequate standard deviation (section 2.5).

As will be shown below, both Pb. 1 and Pb. 2 can be formulated using a system of linear equations. For computational convenience, each system is solved by minimizing a convex optimization problem, using a stochastic gradient descent approach.

In the following, each sample is represented by concatenating the one-hot encodings associated to each feature, noting for each feature the bin it belongs to. Eventually, the dataset is encoded in binary (K, n) matrix R, where n is the number of samples, $K = qd$ is the total number of bins, and $R_{k,i} = 1$ iff sample i falls into bin k.[2]

2.2 PB. 1: IMPORTANCE SAMPLING

In this initial phase of the algorithm, one exploits the oracle output providing the marginals of the target dataset along every bin in $\mathcal{B}$. Formally, the *Oracle* yields $p_j = P_t(X \in B_j)$ for all bins in $\mathcal{B}$ (with p_j possibly perturbed for privacy). Let $\mathbf{p}$ denote the vector made of all p_j.

Let $\mathbf{w} \in \mathbb{R}^{+,n}$ denote the sought vector of importance weights associated to the source samples. Under the overlapping assumption and in the large sample limit, one has:

$$R\mathbf{w} = \mathbf{p} \tag{1}$$

The search for the optimal $\mathbf{w}$ considers Eq. 1 augmented with a regularisation term aimed to avoid weight collapse and distribute the weights as equally as possible over all points in a bin (see A.1). In the following, by abuse of notation and for simplicity, R denotes the data matrix weighted with these importance weights ($R := Rdiag(\mathbf{w})$).

2.3 PB. 2: LABEL ESTIMATION

Any binary hypothesis h defined on $\mathbb{R}^d$ defines a label in $\{0, 1\}$ for each data sample in the target data. By definition, letting $\mathbf{q}$ denote the vector yielding the fraction of errors of h compared to the ground truth label h^* on each bin in $\mathcal{B}$, with $\cdot$ the Hadamard product, and still under the overlapping and large sample limit assumptions, it comes:

$$R\,|\mathbf{h} - \mathbf{h}^*| = (R\,\mathbf{1}) \cdot \mathbf{q} \tag{2}$$

Likewise, after linearization (see A.2 for more details) this system with K equations and n unknowns is tackled as a convex optimization problem, relaxing the binary constraint and taking $\hat{y}_i$ in [0,1], and minimizing the squared difference of the right and left hand sides by gradient descent. Two variants of EXPRIL will be considered in the experimental validation: *Cumulative*-EXPRIL solves at iteration t the convex optimization problem derived from the full stacked system (with $K.t$ equations); *Instantaneous*-EXPRIL solves the convex optimization problem only derived from the current linear system (Eq. 2).

[2]More precisely, for $k = q\ell + j$, $R_{k,i} = 1$ iff the i-th sample falls in the j-th bin of the $(\ell + 1)$-th feature.

2.4 ALGORITHM

Learning h_t. Let $\hat{y}_i$ be the (relaxed) label of sample i estimated at iteration $(t-1)$ (Eq. 7): it is mapped to $\{0,1\}$, and h_t is straightforwardly learned to minimize the cross-entropy loss from the labelled dataset $(\mathbf{x}_i, \hat{y}_i)$ for $i = 1 \dots n$, with w_i the weight of the i-sample. The pseudocode of the algorithm is available in A.4.

It is emphasized that, when the answers to the queries (similar to that in Zhang et al. (2017b)) are perturbed by addition of Laplacian noise, the set of queries defines an ϵ-differentially private protocol.

2.5 ϵ-DIFFERENTIAL PRIVACY

Following the differential privacy protocol defined by Dwork et al. (2006), an i.i.d. noise sampled from $Lap(\lambda)$ is added to each entry of the query feedback. For ϵ privacy budget, it must be that $\lambda \geq S/\epsilon$, where the parameter S, called the sensitivity of a query Q, is defined as $\max_{\{\mathcal{D},\mathcal{D}'\}} ||Q(\mathcal{D}) - Q(\mathcal{D}')||_1$ for two neighbouring datasets $\mathcal{D}$ and $\mathcal{D}'$ s.t. $\mathcal{D} = \mathcal{D}' \cup \{\mathbf{x}\}$.

In our setup, a single query corresponds to soliciting marginals along $K = qd$ bins. Thus, the addition of a single data point between D and D' corresponds to an ℓ_1 distance of d; each data point belongs to one and only one bin per dimension d.[3] Stacking $T+1$ such queries, one for learning the importance weights and one for each hypothesis $h_t \; \forall \, t \in \{0, \dots T-1\}$, yields

$$\lambda \geq d(T+1)/\epsilon \tag{3}$$

The proof is detailed in A.3.

3 EXPERIMENTAL VALIDATION

The goal of the experiments is to assess (i) the performance of EXPRIL compared to that of a reference baseline, trained directly on the target data, and (ii) to compare the instantaneous and cumulative versions of EXPRIL, along with an evaluation of the sensitivity of the approach to the dimension d of the dataset.

3.1 BENCHMARKS

Besides a real-world problem (Cardiotocography dataset), five artificial problems noted A, B, C, D, E have been considered, with dimensions ranging in $2, 4, 10, 15, 25$. The datasets are described in detail in A.5, and the hyper-parameters of the generation process in A.6.

3.2 EXPERIMENTAL SETTING

The hypothesis space is made of neural nets, with architectures described in A.6. The (optimal) baseline is given by the average accuracy of a model h^* with the described neural architecture trained directly on the considered target dataset.

The target dataset is divided in two: the target validation set is used to compute the oracle feedback (marginals of the target distribution and error marginals of candidate hypotheses). The target test set is used to measure the reported performance of the approach.

Experiments are averaged over 100 independent runs by varying the pattern of perturbation of the target labels and the split of the target dataset into *validation* (used to answer the queries) and *test* sets. Parameters and settings of each experiment are detailed in A.6. The number of EXPRIL queries (governing the Laplacian noise for differential privacy) is set to 3.

The ratio indicator is evaluated as $\frac{acc(\hat{h}^*) - acc(ExPriL)}{acc(\hat{h}^*)}$, with $\hat{h}^*$ learned from the true target dataset.

[3] For ease of understanding, we derive the differential privacy requirements in terms of counts instead of proportions as in section 2.4.

Dataset	$\hat{h}^*$	C-ExPriL	I-ExPriL	1DP-C-ExPriL	1DP-I-ExPriL	h_0^*	ratio
A	94.3 ± 2.3	76.8 ± 4.6	74.9 ± 5.0	74.0 ± 4.8	71.9 ± 4.6	50.5 ± 7.9	1.8e-1
B	79.9 ± 2.0	58.5 ± 2.5	57.4 ± 2.4	53.8 ± 2.7	53.5 ± 2.7	50.8 ± 3.6	2.7e-1
C	96.2 ± 1.5	88.8 ± 3.3	88.7 ± 3.8	85.8 ± 4.0	85.6 ± 4.0	49.5 ± 6.8	7.7e-2
D	90.9 ± 2.1	82.2 ± 3.5	80.7 ± 3.9	75.1 ± 4.4	74.1 ± 4.7	50.3 ± 5.4	9.5e-2
E	80.3 ± 2.5	72.3 ± 3.4	70.4 ± 3.7	63.0 ± 4.1	60.1 ± 3.9	50.3 ± 4.7	1.0e-1
CTG A	92.0 ± 1.0	87.8 ± 1.5	86.9 ± 1.6	73.0 ± 3.7	70.5 ± 3.2	77.7 ± 1.2	4.5e-2
CTG B	91.2 ± 1.1	86.1 ± 1.4	84.9 ± 1.8	70.1 ± 3.8	65.1 ± 4.0	75.5 ± 1.4	5.5e-2

Table 1: Comparative performances of both models (average and standard deviation over 100 runs), with $\hat{h}^*$ the model learned from the validation target dataset and h_0^* the constant model predicting the majority class.

3.3 Experimental results

In cases where the learner is aware that the source and target instance distributions are the same, the IS step can be skipped. In our experiments, this pertains to datasets A, B and CTG-A, for which we accordingly bypass the IS step. For dataset C, D, E, and CTG-B, the instance distributions are different, and hence IS step is retained. Note that the learning model is always a neural network. The table in A.6 details the experiments settings, and the learning curves are available in A.7.

General comments. As shown in Table 1, in all experiments except B and C and in both the no-DP and the 1-DP cases, C-ExPriL significantly outperforms I-ExPriL, with confidence over 95% after Wilcoxon-Mann-Whitney signed test. The learning curves (subsection A.7) show that both C-ExPriL and I-ExPriL performance plateau after the first iteration.

As could have been expected the performances of the 1-DP algorithms are lower than that of their no-DP counterparts. The gap between those performances increases with number of dimensions of the instance space. This is explained by the fact that the level of differential noise λ is proportional to the number of dimensions d. Additionally, the performance of 1-DP version of both C-ExPriL and I-ExPriL falls below the baseline (h_0^*) for CTG dataset. A likely reason for this fact is that the level of differential noise is too high (compared to the signal) for the oracle feedback to be meaningful for CTG dataset, particularly in the CTG setting of higher dimensions combined with a low sample size ($n \sim 700$ compared to ~ 5000 in E).

4 Discussion and Perspectives

The contribution of this paper has been to show how to learn from a target dataset that is never released by their owner, in a differentially private way. The main limitation of the approach is that the level of noise required to enforce DP increases with the dimensionality d of the instance space, and with the overall number of queries T.

Further work is concerned with addressing this limitation, along the following ways. A first perspective is based on the remark that the queries for the error marginals could be distilled over the iterations. In basic mode, one could randomly subsample the features queries in each iteration, mechanically increasing the number of allowed iterations or decreasing the amplitude of the Laplacian noise. Along this same approach, one could select the most informative features (in terms of supervised feature selection, or considering the entropy of the errors in the bins). Finally, one could vary the bins to be queried in each iteration.

Another perspective consists in designing new informative features, and querying the oracle to provide the error marginals along these features.

A third perspective is to extend the approach to the regression case, where the query would return the average squared error in each bin.

REFERENCES

Charu Aggarwal. On k-anonymity and the curse of dimensionality. *VLDB 2005 - Proceedings of 31st International Conference on Very Large Data Bases*, 2005.

Maria-Florina Balcan and Vitaly Feldman. Statistical active learning algorithms. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 1295–1303, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/84117275be999ff55a987b9381e01f96-Abstract.html.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, 2007.

Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *ECML/PKDD 2014*, 2014.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2020. URL https://archive.ics.uci.edu/ml/datasets/cardiotocography.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, 2006.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 2016.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 2014.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

J. Jordon, Jinsung Yoon, and M. V. D. Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *Proceedings of the International Conference on Learning Representations*, 2019.

Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 2007.

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *Proceedings of the International Conference on Learning Representations*, 2017.

Haoyue Ping, Julia Stoyanovich, and Bill Howe. Datasynthesizer: Privacy-preserving synthetic datasets. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, 2017.

Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data – a privacy mirage, 2020.

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *Proceedings of the International Conference on Learning Representations*, 2017a.

Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Trans. Database Syst.*, 2017b.

A APPENDIX

A.1 PB 1 : IMPORTANCE SAMPLING DETAILS

The goal of Pb 1 consists in approximating the vector of importance weights associated to the source samples. As evoked in 2.2, under the overlapping assumption and in the large sample limit, one has:

$$R\mathbf{w} = \mathbf{p}$$

Letting $\mathbf{1}$ denote the 1-dimensional vector taking value 1 on every coordinate, $\mathbf{w}$ is sought as:

$$\mathbf{w} = \underset{\mathbf{w}\in\mathbf{R}^{+,n}}{\arg\min} \frac{1}{2}||R\mathbf{w} - \mathbf{p}||^2 + \alpha||\mathbf{w} - \frac{1}{n}\mathbf{1}||^2 \tag{4}$$

The above equation defines a quadratic optimization problem, that can be handled using standard non-negative least squares optimization methods. As said, stochastic gradient descent is used for convenience.

A.2 PB2 : LABEL ESTIMATION DETAILS

Under the assumptions of 2.3 and with the same notations, it comes:

$$R\,|\mathbf{h} - \mathbf{h}^*| = (R\,\mathbf{1}) \cdot \mathbf{q}$$

This equality is handled as an approximation to account for the fact that the weighted source dataset only approximately matches the target dataset. A differentiable optimization objective is obtained by rewriting the above as:

$$R\,diag(sign(\mathbf{h} - \mathbf{h}^*))\,(\mathbf{h} - \mathbf{h}^*) \approx (R\,\mathbf{1}) \cdot \mathbf{q} \tag{5}$$

The vector of prediction $\mathbf{h}$ is binary, so it follows that the ground truth $\mathbf{h}^*$ is necessarily solution to

$$R\,diag(sign(\mathbf{h} - 0.5))\,\mathbf{h}^* \approx -(R\,\mathbf{1}) \cdot \mathbf{q} + R\,\mathbf{h} \tag{6}$$

Precisely, the EXPRIL algorithm defines a sequence of candidate hypotheses h_t for $t = \{0 \dots T\}$. Each h_t is submited to the oracle, yielding the vector $\mathbf{q}_t$ of errors on each bin in $\mathcal{B}$. The first hypothesis h_0 is set to the constant hypothesis predicting the label 1 for every sample in the target dataset.

Two EXPRIL variants are considered:
Instantaneous-EXPRIL (referred to as I-EXPRIL) solves at iteration t the convex optimization problem derived from the current linear system (Eq. 6).
Cumulative-EXPRIL (referred to as C-EXPRIL) considers at iteration t the full stacked system with $K.t$ equations, stacking matrices $R\,diag(sign(\mathbf{h}_\ell - 0.5))$ for $\ell \leq t$ into a single matrix $\mathring{R} \in \mathcal{M}_{Kt,n}$ and the vectors $-(R\,\mathbf{1}) \cdot \mathbf{q}_\ell + R\,\mathbf{h}$ into a single vector $\mathring{\mathbf{q}}$. Likewise, an estimation $\hat{y}$ of the ground truth vector $\mathbf{h}^*$ is sought as solution of equation

$$\mathring{R}\mathbf{h}^* = \mathring{\mathbf{q}} \tag{7}$$

This linear system with Kt equations and n unknowns is tackled as a convex optimization problem, relaxing the binary constraint and taking $\hat{y}$ in $[0,1]^n$, and minimizing the squared difference of the right and left hand sides by gradient descent.

A.3 ϵ-DIFFERENTIAL PRIVACY PROOF

Let $\mathcal{A}$ be a randomized algorithm, with h a binary model defined on a 1-dimensional instance space and $\mathcal{B} = \{B_1, \dots B_q\}$ a set of bins on this instance space. Given a supervised dataset $\mathcal{D} = (x_i, y_i)_{i\in[\![1,n]\!]}$, $\mathcal{A}$ returns the clipped Laplacian-perturbed proportions of errors of h in each bin, with $clip(x)$ a function that returns 0 if x is negative, 1 if $x \geq 1$, x otherwise. Finally, $\mathcal{A}(\mathcal{D})$ returns a vector of instantiations of random variables $(C_1, C_2, \cdots, C_q)$ with

$$C_j \sim (clip(\frac{\mathcal{L}aplace(\lambda) + \sum_{x_i\ in\ B_j} \delta_{h(x_i)\neq y_i}}{\sum_{x_i\ in\ B_j} 1})) \tag{8}$$

Let us consider two datasets $\mathcal{D}$ and $\mathcal{D}'$ such that $\mathcal{D} = \mathcal{D}' \bigcup \{x\}$, and let us assume with no loss of generality that x belongs to bin B_1. Let us assess the probability for $\mathcal{A}$ to return the same output for $\mathcal{D}$ and $\mathcal{D}'$. For $(c_1, \ldots, c_q)$ be in $[0,1]^q$, let us define:

$$\exp(\eta) = \frac{\mathbb{P}(\mathcal{A}(\mathcal{D}) = (c_1, \ldots, c_q))}{\mathbb{P}(\mathcal{A}(\mathcal{D}') = (c_1, \ldots, c_q))} \tag{9}$$

$\mathcal{A}$ is ϵ-DP if $\eta \leq \epsilon$ for any $(c_1, \ldots, c_q)$. By independence of the noise term in each bin,

$$\exp(\eta) = \frac{\mathbb{P}(\mathcal{A}(\mathcal{D})_1 = c_1)}{\mathbb{P}(\mathcal{A}(\mathcal{D}')_1 = c_1)} \tag{10}$$

Let us now suppose that $h(x) = y$ (same result holds if $h(x) \neq y$). We note $U = \sum_{x_i \ in \ B_1} \delta_{h(x_i) \neq y_i}$, $V = \sum_{x_i \ in \ B_1} \delta_{h(x_i) = y_i}$ and L the Laplacian noise term. Then,

$$\exp(\eta) = \frac{\mathbb{P}(clip(\frac{U+L}{U+V}) = c_1)}{\mathbb{P}(clip(\frac{U+L}{U+1+V}) = c_1)} \tag{11}$$

Three cases arise.

If $c_1 \in]-1, 1[$,

$$\exp(\eta) = \frac{\exp\Big(-\frac{1}{\lambda}(|c_1(U+V) - U|)\Big)}{\exp\Big(-\frac{1}{\lambda}(|c_1(U+1+V) - U|)\Big)} \leq \exp(\frac{1}{\lambda}) \tag{12}$$

If $c_1 = 1$,

$$\exp(\eta) = \frac{\mathbb{P}(L > V)}{\mathbb{P}(L > V + 1)} = \exp(\frac{1}{\lambda}) \tag{13}$$

If $c_1 = 0$, the computations are identical.
Hence, it appears that $\mathcal{A}$ is $\frac{1}{\lambda}$-differentially private. In the context of EXPRIL, the *Oracle* acts in a similar way when returning the proportions of errors in each bin, and does it for each of the d features axis in parallel. Each query of this kind to the *Oracle* is thus d/λ-differentially private, which concludes the proof.

A.4 PSEUDO-CODE

The pseudo-code of EXPRIL illustrates the estimation of labels $\hat{y}_i \in [0, 1]$ of the weighted source samples, where hypothesis h is described by parameter vector θ.

Algorithm 1: Iterative marginals matching

Max. number of iterations T
Init: $t := 0$
  Learn **w** from Eq. 4
  $h_0(X; \theta := \theta_0)$ s.t. $h_0 = 1$ everywhere
for *t in [1, T]* **do**
  Query the *Oracle*:
    $\mathbf{q}_{t-1}$ = error marginals incurred by h_{t-1}
  Update($\mathring{R}, \mathring{\mathbf{q}}$):
    [per I-EXPRIL or C-EXPRIL]
  Solve $\mathbf{h}^*$ = $\arg\min ||\mathring{R}\hat{\mathbf{h}} - \mathring{\mathbf{q}}||^2$
  Learn $h_t(\theta := \theta_t)$ s.t. $\theta_t = \arg\min_\theta$ loss($\mathbf{h}^*$; $\mathbf{h_t}$)
end

A.5 DATASETS DESCRIPTION

Artificial datasets A and B Both the source and target instances distributions are sampled from a d-dimensional uniform law over the hypercube $[0, 1]^n$. A randomly initialized K-Means algorithm clusters the target. Half of the clusters are randomly assigned the label 0, and the remaining the label 1. All samples inherit their label from the value of the cluster they belong to. Finally, the label of all target samples x_i that are the solution to the equation $\frac{d}{2} - \lambda \leq \sum_{k=1}^{d} x_{i,k} \leq \frac{d}{2} + \lambda$ are flipped. The value of λ is chosen such that this set corresponds to approximately one third of all target samples.

Artificial datasets C, D and E The source $\mathcal{Z}_s$ and target $\mathcal{Z}_t$ latent distributions are sampled from two different but overlapping mixtures of Gaussians in a latent space of dimension d_l. A randomly initialized neural network f then maps the points in the latent space to the instance space yielding a source $f(\mathcal{Z}_s)$ and a target $f(\mathcal{Z}_t)$ distribution, of dimension d_i. The target labels are obtained by thresholding the output $f'(\mathcal{Z}_t)$ of another randomly initialised neural network f'.

Cardiotocography dataset The Cardio-Tocography (CTG) dataset Dua & Graff (2020) is a public medical dataset. After normalization and feature processing, two versions with 22 real-valued features are considered; CTG-A with same source and target instance distributions, and CTG-B where the two distributions are significantly different (but overlapping).

CTG-A: the original dataset is randomly split into a source and a target dataset ;
CTG-B: a specific feature, the heart beat rate, is considered. Depending on its value $f(x)$ normalized in [.1, .9]), sample x is selected as source sample with probability $f(x)$, otherwise, it is selected as target sample. Source and target distributions are thus different, though they satisfy the overlapping assumption. Source and target datasets have the same size.

A.6 EXPERIMENTAL SETTINGS

d_i : number of instance dimensions
d_l : number of latent dimensions

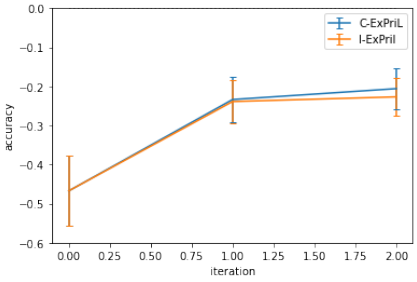
Dataset	#runs	#queries (inc. IS step)	IS step	#target samples	Parameters	[Hidden Layers]
A	100	2	No	2500	15 clusters, $d_i = 2$	[16,256,256,16]
B	100	2	No	2500	64 clusters, $d_i = 4$	[16,256,256,16]
C	100	3	Yes	5000	$d_l = 3$, $d_i = 10$	[32,32]
D	100	3	Yes	5000	$d_l = 5$, $d_i = 15$	[32,32]
E	100	3	Yes	5000	$d_l = 10$, $d_i = 25$	[32,32]
CTG-A	100	2	No	700	*n/a*	[16,256,256,16]
CTG-B	100	3	Yes	700	*n/a*	[16,256,256,16]

Table 2: Settings of the different experiments

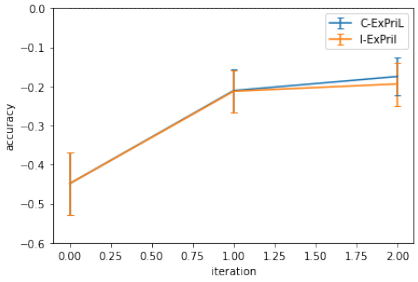
A.7 LEARNING CURVES

For the learning curves of Figure **1**, the averaged accuracy of a model h^* trained directly on the target dataset is taken as reference. A point in the graph at coordinate $[x, y]$ represents the performance of a model h after the x^{th} iteration where the performance is summarized in $y = acc(h) - acc(\hat{h}^*)$.

One sees from these learning curves that the performance plateaus very soon (except on dataset A); the magnitude of the Laplacian noise for guaranteeing privacy thus is over-dimensioned (expecting a total of 3 queries) and could have been reduced to 2.

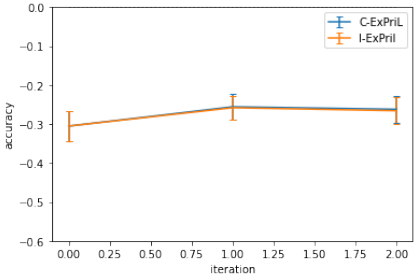


1-DP version

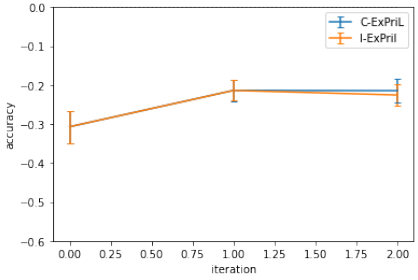


no-DP version

(a) Experiments over dataset A, error bars = $\pm\sigma$

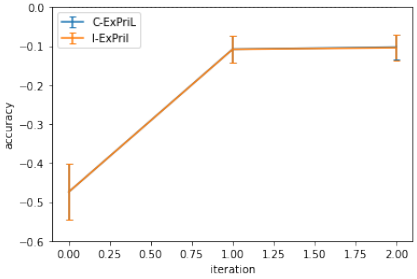


1-DP version

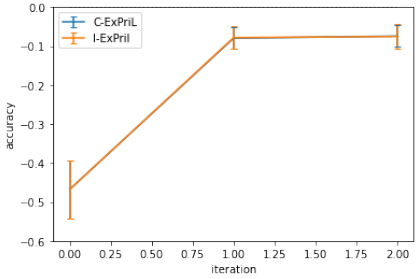


no-DP version

(b) Experiments over dataset B, error bars = $\pm\sigma$

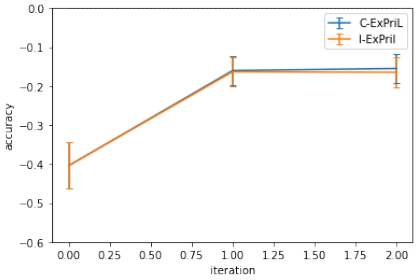


1-DP version

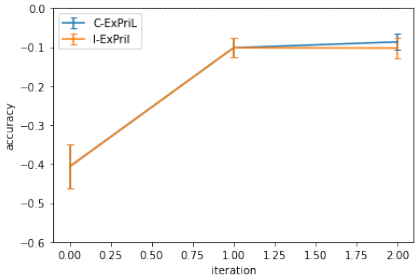


no-DP version

(c) Experiments over dataset C, error bars = $\pm\sigma$

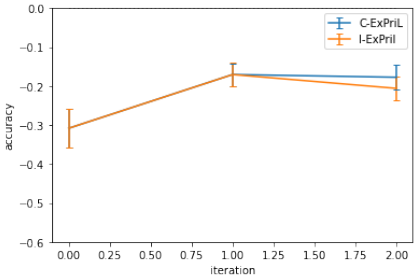


1-DP version

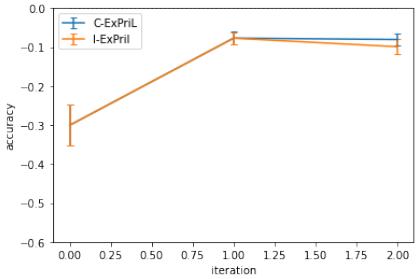


no-DP version

(d) Experiments over dataset D, error bars = $\pm\sigma$

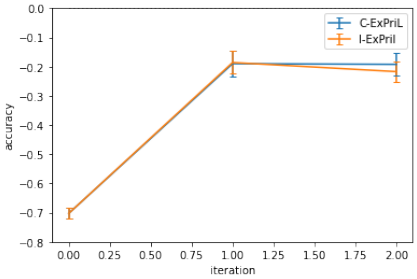


1-DP version

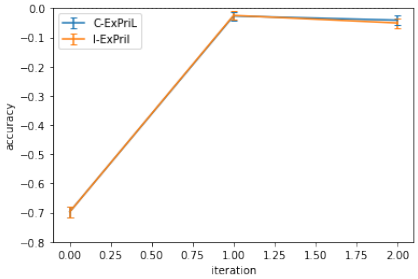


no-DP version

(e) Experiments over dataset E, error bars = $\pm\sigma$

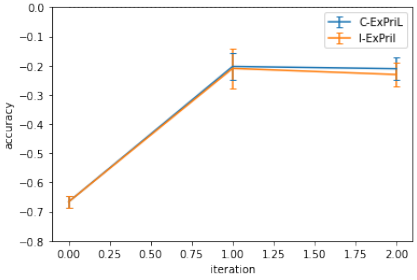


1-DP version

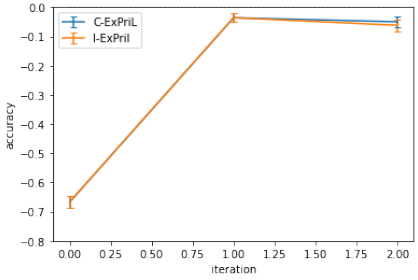


no-DP version

(f) Experiments over dataset CTG-A, error bars = $\pm\sigma$



1-DP version



no-DP version

(g) Experiments over dataset CTG-B, error bars = $\pm\sigma$

Figure 1: Learning curves