Rademacher Observations, Private Data, and Boosting

Richard Nock
NICTA & the Australian National University
richard.nock@nicta.com.au

Giorgio Patrini
NICTA & the Australian National University
giorgio.patrini@anu.edu.au

Arik Friedman
NICTA & the University of New South Wales
arik.friedman@nicta.com.au

April 3, 2015

Abstract

The minimization of the logistic loss is a popular approach to batch supervised learning. Our paper starts from the surprising observation that, when fitting linear (or kernelized) classifiers, the minimization of the logistic loss is equivalent to the minimization of an exponential rado-loss computed (i) over transformed data that we call Rademacher observations (rados), and (ii) over the same classifier as the one of the logistic loss. Thus, a classifier learnt from rados can be directly used to classify observations. We provide a learning algorithm over rados with boosting-compliant convergence rates on the logistic loss (computed over examples). Experiments on domains with up to millions of examples, backed up by theoretical arguments, display that learning over a small set of random rados can challenge the state of the art that learns over the complete set of examples. We show that rados comply with various privacy requirements that make them good candidates for machine learning in a privacy framework. We give several algebraic, geometric and computational hardness results on reconstructing examples from rados. We also show how it is possible to craft, and efficiently learn from, rados in a differential privacy framework. Tests reveal that learning from differentially private rados can compete with learning from random rados, and hence with batch learning from examples, achieving non-trivial privacy vs accuracy tradeoffs.

1 Introduction

This paper deals with the following fundamental question:

What information is sufficient for learning, and what guarantees can it bring that regular data cannot?

By “regular”, we mean the usual inputs provided to a learner. In our context of batch supervised learning, this is a training set of examples, each of which is an observation with a class, and learning means inducing in reduced time an accurate function from observations to classes, a classifier. It turns out that we do not need the detail of classes to learn a classifier (linear or kernelized): an aggregate, whose size is the dimension of the observation space, is minimally sufficient, the mean operator [24].
But do we need examples?

This perhaps surprising and non-trivial question is becoming crucial now that the nature of stored and processed signals intelligence data is heavily debated in the public sphere [19, 28]. In the context of machine learning (ML), the objective of being accurate is more and more frequently subsumed by more complex goals, sometimes involving challenging tradeoffs in which accuracy does not ultimately appear in the topmost requirements. Privacy is one such crucial goal [10, 14, 15]. There are various models to capture the privacy requirement, such as secure multi-party computation and differential privacy (DP, [12]). The former usually relies on cryptographic protocols, which can be heavy even for bare classification and simple algorithms [4]. The latter usually relies on the power of randomization to ensure that any “local” change cannot be spotted from the output delivered [13, 12]. In a ML setting, randomization can be performed at various stages, from the examples to the output of a classifier. We focus on the upstream stage of the process, i.e. the input to the learner, which grants the benefits that all subsequent stages also comply with differential privacy. Randomization has its power: it also has its limits in this case, as it may significantly degrade the performance of learners.

The way we address this problem starts from a surprising observation, whose relevance to supervised ML goes beyond learning with private data: learning a linear (or kernelized) classifier over examples throughout the minimization of the expected logistic loss is equivalent to learning the same classifier by minimizing an exponential loss over a complete set of transformed data that we call Rademacher observations, rados. Each rado is the sum of edge vectors over examples (edge = observation × label). We also show that efficient learning from all rados may also be achieved when carried out over subsets of all possible rados.

This is our first contribution, and we expect it to be useful in several other areas of supervised learning. In the context of learning with private data, our other contributions can be summarized as showing how rados may yield new privacy guarantees — not limited to differential privacy — while authorising boosting-compliant rates for learning. More precisely, our second contribution is to propose a rado-based learning algorithm, which has boosting-compliant convergence rates over the logistic loss computed over the examples. Thus, we learn an accurate classifier over rados, and the same classifier is accurate over examples as well.

The fact that efficient learning may be achieved through subset of rados is interesting because it opens the problem of designing this particular subset to address domain-specific requirements that add to the ML accuracy requirement. Among our other contributions, we provide one important design example, showing how to build differentially private mechanisms for rado delivery, such as when protecting specific sensitive features in data. Experiments confirm in this case that learning from differentially private rados may still be competitive with learning from examples. We provide another design which pairs to our rado-based boosting algorithm, with the crucial property that when examples have been DP-protected by the popular Gaussian mechanism [12], the joint pair (rado delivery design, boosting algorithm) may achieve convergence rates comparable to the noise-free setting with high probability, even over strong DP protection regimes. Our last contribution is to show that rados may protect the privacy of the original examples not only in the DP framework, but also from several algebraic, geometric and even computational-complexity theoretic standpoints.

The remainder of this paper is organized as follows. Section §2 presents Rademacher observations, shows the equivalence between learning from examples and learning from rados, and how learning from subsets of rados may be sufficient for efficient learning; §3 presents our rado-based boosting algorithm, and §4 presents experiments with this algorithm; §5 presents our results in DP models, §6 presents related experiments; §7 provides results on the hardness of reconstructing examples from rados from algebraic, geometric and computational standpoints. To keep a readable paper, proofs and additional experiments are given in two separate appendices available in Section
Rados and supervised learning

Let \( [n] = \{1, 2, ..., n\} \). We are given a set of \( m \) examples \( S = \{(x_i, y_i), i \in [m]\} \), where \( x_i \in \mathcal{X} \subseteq \mathbb{R}^d \) is an observation and \( y_i \in \{-1, 1\} \) is a label, or class. \( \mathcal{X} \) is the domain. A linear classifier \( \theta \in \Theta \) for some fixed \( \Theta \subseteq \mathbb{R}^d \) gives a label to \( x \in \mathcal{X} \) equal to the sign of \( \theta^\top x \in \mathbb{R} \). Our results can be lifted to kernels (at least with finite dimension feature maps) following standard arguments [26]. We let \( \Sigma_m = \{-1, 1\}^m \).

**Definition 1** For any \( \sigma \in \Sigma_m \), the Rademacher observation \( \pi_\sigma \) with signature \( \sigma \) is \( \pi_\sigma \). The simplest way to randomly sample rados is to pick \( \sigma \) as i.i.d. Rademacher variables, hence the name. Reference to \( S \) is implicit in the definition of \( \pi_\sigma \). A Rademacher observation sums edge vectors (the terms \( y_i x_i \)), over the subset of examples for which \( y_i = \sigma_i \). When \( \sigma = y \) is the vector of classes, \( \pi_\sigma = m \mu_S \) is \( m \) times the mean operator [26] [24]. When \( \sigma = -y \), we get the null vector \( \pi_\sigma = 0 \). A popular approach to learn \( \theta \) over \( S \) is to minimize the surrogate risk \( F_{\log}(S, \theta) \) built from the logistic loss (logloss):

\[
F_{\log}(S, \theta) = \frac{1}{m} \sum_i \log \left( 1 + \exp \left( -y_i \theta^\top x_i \right) \right) .
\]

We define the exponential rado-risk \( F_{\exp}(S, \theta, U) \), computed on any \( U \subseteq \Sigma_m \) with cardinal \( |U| = n \), as:

\[
F_{\exp}(S, \theta, U) = \frac{1}{n} \sum_{\sigma \in U} \exp \left( -\theta^\top \pi_\sigma \right) .
\]

It turns out that \( F_{\log} = g(F_{\exp}) \) for some continuous strictly increasing \( g \); hence, minimizing one criterion is equivalent to minimizing the other and *vice versa*. This is stated formally in the following Lemma.

**Lemma 2** The following holds true, for any \( \theta \) and \( S \):

\[
F_{\log}(S, \theta) = \log(2) + \frac{1}{m} \log F_{\exp}(S, \theta, \Sigma_m) .
\]

(Proof in the Appendix, Subsection 10.1). Lemma 2 shows that learning with examples via the minimization of \( F_{\log}(S, \theta) \), and learning with all rados via the minimization of \( F_{\exp}(S, \theta, \Sigma_m) \), are essentially equivalent tasks. Since the cardinal \( |\Sigma_m| = 2^m \) is exponential, it is unrealistic, even on moderate-size samples, to pick that latter option. This raises however a very interesting question: if we replace \( \Sigma_m \) by subset \( U \) of size \( \ll 2^m \), what does the relationship between examples and rados in eq. (3) become? We answer this question under the setting that:

(i) instead of \( \Sigma_m \), we consider a predefined \( \Sigma_r \subseteq \Sigma_m \);

(ii) instead of considering \( U = \Sigma_r \), we sample uniformly i.i.d. \( U \sim \Sigma_r \) for \( n \geq 1 \) rados.
While (ii) is directly targeted at reducing the number of rados, (i) is an upper-level strategic design to tackle additional constraints, such as differential privacy. We now need the following definition of the logistic rado-risk:

\[ F_{\log}^r(S, \theta, U) = \log(2) + \frac{1}{m} \log F_{\exp}^r(S, \theta, U), \]

for any \( U \subseteq \Sigma_m \), so that \( F_{\log}(S, \theta) = F_{\log}^r(S, \theta, \Sigma_m) \). We also define the open ball \( \mathcal{B}(0, r) = \{ x \in \mathbb{R}^d : \| x \|_2 < r \} \).

**Theorem 3** Assume \( \Theta \subseteq \mathcal{B}(0, r_\theta) \), for some \( r_\theta > 0 \). Let:

\[ q = \sup_{\theta' \in \Theta} \max_{\pi_\sigma \in \Sigma_r} \exp(-\theta'^T \pi_\sigma) \]

\[ q' = \frac{F_{\exp}^r(S, \theta, \Sigma_r)}{F_{\exp}^r(S, \theta, \Sigma_m)} \]

where \( \Sigma_r \) follows (i) above. Then \( \forall \eta > 0 \), there is probability \( \geq 1 - \eta \) over the sampling of \( U \) in (ii) above that:

\[ F_{\log}(S, \theta) \leq F_{\log}^r(S, \theta, U) + Q - \frac{1}{m} \cdot \log \left( 1 - \frac{q}{\sqrt{n}} \right), \]

with

\[ q = \Omega \left( \sqrt{r_\theta \max_{\Sigma_r} \| \pi_\sigma \|_2 + d \log \frac{2en}{d} + \log \frac{1}{\eta}} \right) \]

and \( Q \) satisfies \( Q = 0 \) if \( \Sigma_r = \Sigma_m \) and

\[ Q \leq r_\theta \left( \| \nabla_\theta F_{\log}^r(S, \theta, \Sigma_m) \|_2 + \pi_r \right) \]

otherwise, letting \( \pi_r = \| \mathbb{E}_{\sigma \sim \Sigma_r} (1/m) \cdot \pi_\sigma \|_2 \). Furthermore, \( \forall 0 \leq \beta < 1/2 \), if \( m \) is sufficiently large, then letting \( \pi^*_r = \max_{\Sigma_r} \| (1/m) \cdot \pi_\sigma \|_2 \), ineq. (7) becomes:

\[ F_{\log}(S, \theta) \leq F_{\log}^r(S, \theta, U) + Q + O \left( \frac{\theta}{m^3} \cdot \sqrt{\frac{r_\theta \pi^*_r}{n} + \frac{d \log 2en}{dn}} \right). \]

(Proof in the Appendix, Subsection 10.2) Theorem 3 does not depend on the algorithm that learns \( \theta \). The right-hand side of ineq. (5) shows two penalties. \( Q \) arises from the choice of \( \Sigma_r \) and is therefore structural. Regardless of \( \Sigma_r \), when the classifier is reasonably accurate over all rados and expected examples edges in \( \Sigma_r \) average to a ball of reduced radius, the upperbound on \( Q \) in ineq. (7) can be very small. The other penalty, which depends on \( q \), is statistical and comes from the sampling in \( \Sigma_r \). Theorem 3 shows that when \( \Sigma_r = \Sigma_m \), even when \( n \ll m \), the minimization of \( F_{\log}^r(S, \theta, U) \) may still bring, with high probability, guarantees on the minimization of \( F_{\log}(S, \theta) \). Thus, a lightweight optimization procedure over a small number of rados may bring guarantees on the minimization of the expected logloss over examples for the same classifier. The following Section exhibits one such algorithm.
Algorithm 1 Rado boosting (RadoBoost)

**Input** set of rados \( S^r = \{\pi_1, \pi_2, ..., \pi_n\} \); \( T \in \mathbb{N}^* \);
Step 1 : let \( \theta_0 \leftarrow 0, w_0 \leftarrow (1/n)1 \);
Step 2 : for \( t = 1, 2, ..., T \)
  Step 2.1 : 
  \[ \text{let } r_t \leftarrow \frac{1}{\pi_{s(t)}} \sum_{j=1}^{n} w_{tj} \pi_{jk}(t) \]  
  (9) 
  \[ \alpha_t \leftarrow \frac{1}{2\pi_{s(t)}} \log \frac{1 + r_t}{1 - r_t} \]  
  (10) 
Step 2.2 : let 
Step 2.3 : for \( j = 1, 2, ..., n \)
  \[ w_{(t+1)j} \leftarrow w_{tj} \cdot \left( 1 - \frac{r_t \pi_{jk}(t)}{1 - r_t^2} \right) \]  
  (11)

**Return** \( \theta_T \) defined by \( \theta_{Tk} = \sum_{t_{i(t)} = k} \alpha_t, \forall k \in [d] \);

3 Boosting using rados

Algorithm [1] provides a boosting algorithm, RADOBOOST, that learns from a set of Rademacher observations \( S^r = \{\pi_1, \pi_2, ..., \pi_n\} \). Their (unknown) Rademacher assignments are denoted \( U = \{\sigma_1, \sigma_2, ..., \sigma_n\} \subseteq \Sigma_m \). These rados have been computed from some sample \( S \), unknown to RADOBOOST. In the statement of the algorithm, \( \pi_{jk} \) denotes coordinate \( k \) of \( \pi_j \), and \( \pi_{sk} = \max_j |\pi_{jk}| \).

More generally, the coordinates of some vector \( z \in \mathbb{R}^d \) are denoted \( z_1, z_2, ..., z_d \). Step 2.1 gets a feature index \( i(t) \) from a weak feature index oracle, WF1. In its general form, WF1 returns a feature index maximizing \( |r_t| \) in (9). The weight update was preferred to AdaBoost’s because rados can have large feature values and the weight update prevents numerical precision errors that could otherwise occur using AdaBoost’s exponential weight update. We now prove a key Lemma on RADOBOOST, namely the fast convergence of the exponential rado-risk \( F_{\exp}(S, \theta, U) \) under a weak learning assumption (WLA). We shall then obtain the convergence of the logistic rado-risk (4), and, via Theorem 3, the convergence with high probability of \( F_{\log}(S, \theta) \).

(WLA) \( \exists \gamma > 0 \) such that \( \forall t \geq 1 \), the feature returned by WF1 in Step 2.2 (9) satisfies \( |r_t| \geq \gamma \).

**Lemma 4** Suppose the (WLA) holds. Then after \( T \) rounds of boosting in RADOBOOST, the following upperbound holds on the exponential rado-loss of \( \theta_T \):

\[ F_{\exp}(S, \theta_T, U) \leq \exp \left( -T \gamma^2 / 2 \right) \]  
(12)

(Proof in the Appendix, Subsection 10.3) We now consider Theorem 3 with \( \Sigma_r = \Sigma_m \), and therefore \( Q = 0 \). Blending Lemma 4 and Theorem 3 using (4) yields that, under the (WLA), we may observe with high probability (again, fixing \( \Sigma_r = \Sigma_m \), so \( Q = 0 \) in Theorem 3):

\[ F_{\log}(S, \theta_T) \leq \log(2) - \frac{T \gamma^2}{2m} + Q' \]  
(13)
Table 1: Comparison of RadoBoost (n random rados), AdaBoost \[27\] (full training fold) and AdaBoost(n) (n random examples in training fold); domains ranked in increasing d · m value. Column “n/m” (resp. “n/2^m”) for AdaBoost(n) (resp RadoBoost) is proportion of training data with respect to fold size (resp. full set of rados). Notation [a: b] is shorthand for a \times 10^b. Column “100σ” is the number of features with outlier values distant from the mean by more than 100σ in absolute value. Column p (resp. p’) is p-value for a two-tailed paired t-test on AdaBoost (resp. AdaBoost(n)) vs RadoBoost. ε means < 0.01.

where Q’ is the rightmost term in ineq. (5) or ineq. (8). So provided n \ll 2^m is sufficiently large, minimizing the exponential rado-risk over a subset of rados brings a classifier whose average logloss on the whole set of examples may decrease at rate Ω(γ^2/m) under a weak learning assumption made over rados only. This rate competes with those for direct approaches to boosting the logloss \[23\], and we now show that our weak learning assumption is also essentially equivalent to the one done in boosting over examples \[27\]. Let us rewrite r_t(w) as the normalized edge in \[9\], making explicit the dependence in the current rado weights. Let

\[
r_t^{\text{ex}}(w) = \frac{1}{x_{\star\tau(t)}} \sum_{i=1}^{m} w_i x_{i\tau(t)}
\]

be the normalized edges for the same feature \(i(t)\) as the one picked in step 2.1 of RadoBoost, but computed over examples using some weight vector \(\hat{w} \in \mathbb{P}^m\); here, \(\mathbb{P}^m\) is the m-dim probability simplex and \(x_{\star\tau(t)} \equiv \max_{i} |x_{ik}|\).

Lemma 5 ∀ \(w_t \in \mathbb{P}^m, \forall γ > 0\), there exists \(\bar{w} \in \mathbb{P}^m\) and \(γ^{ex} > 0\) such that \(|r_t(w_t)| \geq γ\) iff \(|r_t^{\text{ex}}(\bar{w})| \geq γ^{ex}\).

(Proof in the Appendix, Subsection 10.4) The proof of the Lemma gives clues to explain why the presence of outlier feature values may favor RadoBoost.

4 Basic experiments with RadoBoost

We have compared RadoBoost to its main contender, AdaBoost \[27\], using the same weak learner; in AdaBoost, it returns a feature maximizing \(|r_t|\) as in eq. (14). In these basic experi-
ments, we have deliberately not optimized the set of rados in which we sample \( U \) for RadoBoost; hence, we have \( \Sigma_r = \Sigma_m \).

We have performed comparisons with 10 folds stratified cross-validation (CV) on 16 domains of the UCI repository [2] of varying size. For space considerations, Table 1 presents the results. Each algorithm was ran for a total number of \( T = 1000 \) iterations; furthermore, the classifier kept for testing is the one minimizing the empirical risk throughout the \( T \) iterations; in doing so, we also assessed the early convergence of algorithms. We fixed \( n = \min\{1000, \text{train fold size}/2\} \). Table 1 displays that RadoBoost compares favourably to AdaBoost, and furthermore it tends to be all the better as \( m \) and \( d \) increase. On some domains like Hardware and Twitter, the difference is impressive and clearly in favor of RadoBoost. As discussed for Lemma 5, we could interpret these comparatively very poor performances of AdaBoost as the consequence of outlier features that can trick AdaBoost in picking the wrong sign in the leveraging coefficient \( \alpha_t \) for a large number of iterations if we use real-valued classifiers (see column 100\( \sigma \) in Table 1). This drawback can be easily corrected (Cf Appendix, Subsection 11.1) by enforcing minimal \( |r_t| \) values. This significantly improves AdaBoost on Hardware and Twitter. The improvements observed on RadoBoost are even more favorable.

5 Rados and differential privacy

We now discuss the delivery of rados to comply with several DP constraints and their eventual impact on boosting. We thus address both levels (i+ii) of rado delivery in [2]. Our general model is the standard DP model [12]. Intuitively, an algorithm is DP compliant if for any two neighboring datasets, it assigns similar probability to any possible output \( O \). In other words, any particular record has only limited influence on the probability of any given output of the algorithm, and therefore the output discloses very little information about any particular record in the input. Formally, a randomized algorithm \( A \) is \((\epsilon, \delta)\)-differentially-private [11] for some \( \epsilon, \delta > 0 \) iff:

\[
P_A(O|S) \leq \exp(\epsilon) \cdot P_A(O|S') + \delta, \forall S \approx S', O,
\]

where the probability is over the coin tosses of \( A \). This model is very strong, especially when \( \delta = 0 \), and in the context of ML, maintaining high accuracy in strong DP regimes is generally
5.1 A feature-wise DP mechanism for rados

In this Subsection, we consider a relaxation of differential-privacy, namely feature-wise differential privacy, where the differential privacy requirement applies to $j_*$-neighboring datasets: we say that two samples $S, S'$ are $j_*$-neighbors, noted $S \approx j_* S'$, if they are the same except for the value of the $j^{th}$ observation feature of some example. We further assume that the feature is boolean. For example, we may have a medical database containing a column representing the HIV status of a doctor’s patients (1 row = a patient), and we do not wish that changing a single patient HIV status significantly changes the density of that feature’s values in rados. This setting would also be very useful in genetic applications to hide in rados gene disorders that affect one or few genes. Feature-wise DP is analogous to the concept of $\alpha$-label privacy [7], where differential privacy is guaranteed with respect to the label. Algorithm $A$ in ineq. (15) is given in Algorithm 2. It relies on the following subset $\Sigma_\beta \subseteq \Sigma_m$:

$$\Sigma_\beta^{j_*} = \left\{ \sigma \in \Sigma_m : \pi_{\sigma j_*} \in \left[ \left| \{i : y_i x_{ij_*} = +1\} \right| - \frac{m}{2} \pm \Delta_\beta \right] \right\}, \quad (16)$$

a tricky tradeoff [10]. Because rados are an intermediate step between training sample $S$ and a rado-based learner, there are two ways to design rados with respect to the DP framework: crafting DP-compliant rados from unprotected examples, or crafting rados from DP-compliant examples with the aim to improve the performance of the rado-based learner (Figure 5.2). These scenarii can be reduced to the design of $\Sigma_\beta$. 
with \( \Delta_\beta = (m/2) - \beta(m + 1) \). The key feature of this mechanism is that it does not alter the examples in the sense that DP-compliant rados belong to the set of cardinal \( 2^m \) that can be generated from \( S \). Usual data-centered DP mechanisms would rather alter data, e.g. via noise injection \[15\]. Algorithm \[2\] exploits the fact that it is the tails of feature \( j_* \) that leak sensitive information about the feature in rados (see Figure \[3\]). The following Theorem is stated so as we can pick small \( \delta \), typically \( \delta \ll 1/m \). Other variants are possible that bring different tradeoffs between \( \epsilon \) and \( \delta \).

**Theorem 6** Assume \( \epsilon \) is chosen so that \( \epsilon = o(1) \) but \( \epsilon = \Omega(1/m) \). In this case, DP-Feat maintains \((n \cdot \epsilon, n \cdot \delta)\)-differential privacy on feature \( j_* \) for some \( \delta > 0 \) such that \( \epsilon \cdot \delta = O(m^{-5/2}) \).

(Proof in the Appendix, Subsection \[10.5\]) We have implemented Step 2 in Algorithm DP-Feat in the simplest way, using a simple Rademacher rejection sampling where each \( \sigma_j \) is picked i.i.d. as \( \sigma_j \sim \sum_m \) until \( \sigma_j \in \Sigma_m^{j_*} \). The following Theorem shows its algorithmic efficiency.

**Theorem 7** For any \( \eta > 0 \), let \( n^*_\eta = \eta(1 - \exp((2\beta - 1))/((4\beta)) \), and let \( n_R \) denote the total number of rados sampled in \( \sum_m \) until \( n \) rados are found in \( \Sigma_m^{j_*} \). Then for any \( \eta > 0 \), there is probability \( \geq 1 - \eta \) that

\[
n_R \leq n \cdot \left\{ \frac{1}{mD_{BE}(1 - \beta \| 1/2 \}) \log \frac{n}{n^*_\eta} \right\} \quad \text{if} \quad n \leq n^*_\eta
\]

\[
\text{or otherwise}
\]

where \( D_{BE} \) is the bit-entropy divergence: \( D_{BE}(p \| q) = p \log(p/q) + (1 - p) \log((1 - p)/(1 - q)) \), for \( p, q \in (0, 1) \).

(Proof in the Appendix, Subsection \[10.6\]) Remark that replacing \( \Sigma_m \) by \( \Sigma_r = \Sigma_m^{j_*} \) would not necessarily impair the boosting convergence of RADOBOOST trained from rados samples from DP-Feat (Lemma \[3\]). The only systematic change would be in ineq. \[13\] where we would have to integrate the structural penalty \( Q \) from Theorem \[3\] to further upperbound \( F_{T \log} (S, \theta_T) \). In this case, the upperbound in \[7\] reveals that at least when the mean operator in \( \Sigma_m^{j_*} \) has small norm — which may be the case even when some examples in \( S \) have large norm — and the gradient penalty is small, then \( Q \) may be small as well.

We end up with several important remarks, whose formal statements and proofs are left out due to space constraints. First, the tail truncation design exploited in DP-Feat can be fairly simply generalized in two directions, to handle (a) real-valued features, and/or (b) several sensitive features instead of one. Second, we can do DP-compliant design of rado delivery beyond feature-wise privacy, e.g. to protect “rado-wide” quantities like norms.

### 5.2 Boosting from DP-compliant examples via rados

We now show how to craft rados from DP-compliant examples so as to approximately keep the convergence rates of RADOBOOST. More precisely, since edge vectors are sufficient to learn (eq. \[1\]), we assume that edge vectors are DP-compliant (neighbor samples, \( S \approx S' \) would differ on one edge vector). A gold standard to protect data in the DP framework is to convolute data with noise. One popular mechanism is the Gaussian mechanism \[12\] \[16\], which convolutes data with independent Gaussian random variables \( N(0, \varsigma^2 I) \), whose standard deviation \( \varsigma \) depends on the DP requirement \((\epsilon, \delta)\). Strong DP regimes are tricky to handle for learning algorithms. For example, the approximation factor \( \rho \) of the singular vectors under DP noise of the noisy power method roughly behaves as \( \rho = \Omega(\varsigma/\Delta) \) \[16\] (Corollary 1.1) where \( \Delta = O(d) \) is a difference between two
s

Table 2: Left table: RADOBOOST on feature-wise DP compliant rados (Subsection 5.1 showing standard deviations) vs RADOBOOST on plain random rados baseline and ADABOOST baseline (trained with complete fold). Center: test error of RADOBOOST minus ADABOOST’s (also showing ADABOOST error on right axis, dotted line), for rados with fixed support $s = \mu m$, in green, red, blue) and plain random rados (dotted grey). Right: test error of RADOBOOST using fixed support $s$ rados and a prudential learner, minus RADOBOOST using plain random rados and “strong” learner of Section 4 (See Table 3 through Table 11).

singular values. When $\varsigma$ is small, this is a very good bound. When the DP requirement blows up, the bound remains relevant if $d$ increases, which may be hard to achieve in practice — it is easier in general to increase $m$ than $d$, which requires to compute new features for past examples.

We consider ineq. (15) with neighbors $I$ and $I'$ being two sets of $m$ edge vectors differing by one edge vector, and $\hat{O}$ is a noised set of $m$ edge vectors generated through the Gaussian mechanism [12] (Appendix A). We show the following non-trivial result: provided we design another particular $\Sigma_r$, the convergence rate of RADOBOOST, as measured over non-noisy rados, essentially survives noise injection in the edge vectors through the Gaussian mechanism, even under strong noise regimes, as long as $m$ is large enough. The intuition is straightforward: we build rados summing a large number of edge vectors only (this is the design of $\Sigma_r$), so that the i.i.d. noise component gets sufficiently concentrated for the algorithm to be able to learn almost as fast as in the noise-free setting. We emphasize the non-trivial fact that convergence rate is measured over the non-noisy rados, which of course RADOBOOST does not see. The result is of independent interest in the boosting framework, since it makes use of a particular weak learner ($w_f$), which we call prudential, which picks features with $|r_t| < \frac{1}{2}$ upperbounded.

We start by renormalizing coefficients $\alpha_t$ (eq. 10) in RADOBOOST by a parameter $\kappa \geq 1$ given as input, so that we now have $\alpha_t \leftarrow \left(\frac{1}{2\kappa \pi \cdot e_t(t)}\right) \log((1 + r_t)/(1 - r_t))$ in Step 2.2. It is not hard
to check that the convergence rate of RADOBOOST now becomes, prior to applying the (WLA)

$$F^r_{\log}(S, \theta_T, U) \leq \log(2) - \frac{1}{2km} \sum_t r^2_t .$$

(17)

We say that $w_{fi}$ is $\lambda_p$-prudential for $\lambda_p > 0$ iff it selects at each iteration a feature such that $|r_t| \leq \lambda_p$. Edges vectors have been DP-protected as $y_i(x_i + x^r_i)$, with $x^r_i \sim N(0, \varsigma^2 I)$ (for $i \in [m]$).

Let $m_{\sigma} = |\{i : \sigma_i = y_i\}|$ denote the support of a rado, and ($m_* > 0$ fixed):

$$\Sigma_r = \Sigma_{m_*} = \{\sigma \in \Sigma_m : m_\sigma = m_*\} .$$

(18)

**Theorem 8** $\forall U \subseteq \Sigma_r, \forall \tau > 0$, if $\sqrt{m_*} = \Omega (\varsigma \ln(1/\tau))$, then $\exists \lambda_p > 0$ such that RADOBOOST having access to a $\lambda_p$-prudential weak learner returns after $T$ iteration a classifier $\theta_T$ which meets with probability $\geq 1 - \tau$:

$$F^r_{\log}(S, \theta_T, U) \leq \log(2) - \frac{1}{4km} \sum_t r^2_t .$$

(19)

The proof, in the Appendix (Subsection 10.7), details parameters and dependencies hidden in the statement. The use of a prudential weak learner is rather intuitive in a noisy setting since $\alpha_t$ blows up when $|r_t|$ is close to 1. Theorem 8 essentially yield that a sufficiently large support for rados is enough to keep with high probability the convergence rate of RADOBOOST within noise-free regime. Of course, the weak learner is prudential, which implies bounded $|r_t| < 1$, and furthermore the leveraging coefficients $\alpha_t$ are normalized, which implies smaller margins. Still, Theorem 8 is a good theoretical argument to rely on rados when learning from DP-compliant edge vectors.

### 6 Experiments on differential privacy

Table 2 presents a subset of the experiments carried out with RADOBOOST and ADABOOST in the contexts of Subsections 5.1 and 5.2 (see Section 11 for all additional experiments). Unless otherwise stated, experimental settings (cross validation, number of rados for learning, etc.) are the same as in Section 4.

In a first set of experiments, we have assessed the impact on learning of the feature-wise DP mechanism: on each tested domain, we have selected at random a binary feature, and then used Algorithm DP-Feat to protect the feature for different values of DP parameter $\epsilon$, in a range that covers usual DP experiments [18] (Table 1). The main conclusion that can be drawn from the experiments is that learning from DP-compliant rados can compete with learning from random rados, and even learning from examples (ADABOOST), even for rather small $\epsilon$.

We then have assessed the impact on learning of examples that have been protected using the Gaussian mechanism [12], with or without rados, with or without a prudential weak learner for boosting, and with or without using a fixed support for rado computation. The Appendix provides extensive results for all domains but the largest ones (Twitter, SuSy, Higgs). In the central column (and Tables 4 through 7 in the Appendix), computing the differences between RADOBOOST’s error and ADABOOST’s reveals that, on domains where it is beaten by ADABOOST when there is no noise, RADOBOOST almost always rapidly become competitive with ADABOOST as noise increases. Hence, RADOBOOST is a good contender from the boosting family to learn from differentially private (or noisy) data. Second, using a prudential weak learner which picks the median feature (instead of the
more efficient weak learner that picks the best as in Section 4) can have RadoBoost with fixed support rados compete or beat RadoBoost with plain random rados, at least for small noise levels (see Transfusion and Magic in the right column of Table 2). Replacing the median-prudential weak learner by a strong learner can actually degrade RadoBoost’s results (see the Appendix, Tables 10 and 11). These two observations advocate in favor of the theory developed in Subsection 5.2. Finally, using rados with fixed support instead of plain random rados (Section 4) can significantly improve the performances of RadoBoost (see the Appendix, Tables 10 and 11).

7 From rados to examples: hardness results

The problem we address here is how we can recover examples from rados, and when we cannot recover examples from rados. This last setting is particularly useful from the privacy standpoint, as this may save us costly obfuscation techniques that impede ML tasks [3].

7.1 Algebraic and geometric hardness

For any \( m \in \mathbb{N}^* \), we define matrix \( G_m \in \{0, 1\}^{m \times 2^m} \) as:

\[
G_m = \begin{bmatrix}
0^{2m-1}_\top & 1^{2m-1}_\top \\
G_{m-1} & G_{m-1}
\end{bmatrix}
\]

if \( m > 1 \), and \( G_1 = [0 \ 1] \) otherwise (\( z_d \) denotes a vector in \( \mathbb{R}^d \)). Each column of \( G_m \) is the binary indicator vector for the edge vectors considered in a rado. Hereafter, we let \( E \in \mathbb{R}^{d \times m} \) the matrix of columnwise edge vectors from \( S \), \( \Pi \in \mathbb{R}^{d \times n} \) the columnwise rado matrix and \( U \in \{0, 1\}^{2^m \times n} \) in which each column gives the index of a rado computed in \( S^r \). By construction, we have:

\[
\Pi = EG_m U ,
\]

and so we have the following elementary results for the (non) reconstruction of \( E \) (proof omitted).

**Lemma 9** (a) when recoverable, edge-vectors satisfy: \( E = \Pi U^{\top} G_m^\top (G_m U U^{\top} G_m^{\top})^{-1} \); (b) when \( U, \Pi, m \) are known but \( n < m \), there is not a single solution to eq. (21) in general.

Lemma 9 states that even when \( U, \Pi \) and \( m \) are known, elementary constraints on rados can make the recovery of edge vectors hard — notice that such constraints are met in our experiments with RadoBoost in Sections 4 and 6.

But this represents a lot of unnecessary knowledge to learn from rados: RadoBoost just needs \( \Pi \) to learn. We now explore the guarantees that providing this sole information brings in terms of (not reconstructing \( E \). \( \forall M \in \mathbb{R}^{a \times b} \), we let \( C(M) \) denote the set of column vectors, and for any \( \mathcal{C} \subseteq \mathbb{R}^d \), we let \( \mathcal{C} \oplus \epsilon = \cup_{\mathbf{z} \in \mathcal{B}(\mathbf{z}, \epsilon)} \mathcal{C} \). We define the Hausdorff distance, \( D_H(E, E') \), between \( E \) and \( E' \):

\[
D_H(E, E') = \inf \{ \epsilon : \mathcal{C}(E) \subseteq \mathcal{C}(E') \oplus \epsilon \wedge \mathcal{C}(E') \subseteq \mathcal{C}(E) \oplus \epsilon \}.
\]

The following Lemma shows that if the only information known is \( \Pi \), then there exist samples that bring the same set of rados \( \mathcal{C}(\Pi) \) as the unknown \( E \) but who are at distance proportional to the “width” of the domain at hand.
Lemma 10 For any \( \Pi \in \mathbb{R}^{d \times n} \), suppose eq. (21) holds, for some unknowns \( m > 0 \), \( E \in \mathbb{R}^{d \times m} \), \( U \in \{0, 1\}^{2m \times n} \). Suppose \( \mathcal{C}(E) \subset B(0, R) \) for some \( R > 0 \). Then there exists \( E' \in \mathbb{R}^{d \times (m+1)} \), \( U' \in \{0, 1\}^{2m+1 \times n} \) such that

\[
\mathcal{C}(E') \subset B(0, R) \quad \text{and} \quad \Pi = E'G_{m+1}U' ,
\]

but

\[
D_H(E, E') = \Omega \left( \frac{R \log d}{\sqrt{d \log m}} \right)
\]

if \( m \geq 2^d \), and \( D_H(E, E') = \Omega(R/\sqrt{d}) \) otherwise.

(Proof in the Appendix, Subsection 10.8) Hence, without any more knowledge, leaks, approximations or assumptions on the domain at hand, the recovery of \( E \) pays in the worst case a price proportional to the radius of the smallest enclosing \( B(0, R) \) ball for the unknown set of examples. We emphasize that this inapproximability result does not rely on the computational power at hand.

7.2 Computational hardness

In this Subsection, we investigate two important problems in the recovery of examples. The first problem addresses whether we can approximately recover sparse examples from a given set of rados, that is, roughly, solve (21) with a sparsity constraint on examples. The first Lemma we give is related to the hardness of solving underdetermined linear systems for sparse solutions \( [9] \). The sparsity constraint can be embedded in the compressed sensing framework \( [8] \) to yield finer hardness and approximability results, which is beyond the scope of our paper. We define problem “Sparse-Approximation” as:

(Instance) : set of rados \( S' = \{ \pi_1, \pi_2, ..., \pi_n \} \), \( m \in \mathbb{N}_* \), \( r, \ell \in \mathbb{R}_+ \), \( \| \cdot \|_p \), \( L_p \)-norm for \( p \in \mathbb{R}_+ \);

(Question) : Does there exist set \( S = \{(x_i, y_i), i \in [m]\} \) and set \( \ell = \{\sigma_1, \sigma_2, ..., \sigma_n\} \in \{-1, 1\}^m \) such that:

\[
\|x_i\|_p \leq \ell, \forall i \in [m], \quad \text{(Sparse examples)}
\]

\[
\|\pi_j - \pi_{\sigma_j}\|_p \leq r, \forall j \in [n]. \quad \text{(Rado approximation)}
\]

Lemma 11 Sparse-Approximation is NP-Hard.

(Proof in the Appendix, Subsection 10.9) In the context of rados, the second problem we address has very large privacy applications. Suppose entity \( A \) has a huge database of people (e.g. clients), and obtains a set of rados emitted by another entity \( B \). An important question that \( A \) may ask is whether the rados observed can be approximately constructed by its database, for example to figure out which of its clients are also its competitors'. We define this as problem “Probe-Sample-Subsumption”:

(Instance) : set of examples \( S \), set of rados \( S' = \{ \pi_1, \pi_2, ..., \pi_n \} \), \( m \in \mathbb{N}_* \), \( p, r \in \mathbb{R}_+ \).

(Question) : Does there exist \( S' = \{(x_i, y_i), i \in [m]\} \subseteq S \) and set \( \ell = \{\sigma_1, \sigma_2, ..., \sigma_n\} \in \{-1, 1\}^m \) such that:

\[
\|\pi_j - \pi_{\sigma_j}\|_p \leq r, \forall j \in [n]. \quad \text{(Rado approximation)}
\]
Lemma 12  \textit{Probe-Sample-Subsumption is NP-Hard.}

(Proof in the Appendix, Subsection 10.10) This worst-case result calls for interesting domain-specific qualifications, such as in genetics where the privacy of raw data, \textit{i.e.} individual genomes, can be compromised by genome-wise statistics \cite{17,21}.

8 Conclusion

We have introduced novel quantities that are sufficient for efficient learning, Rademacher observations. The fact that a subset of these can replace traditional examples for efficient learning opens interesting problems on how to craft these subsets to cope with additional constraints. We have illustrated these constraints in the field of efficient learning from privacy-compliant data, from various standpoints that include differential privacy as well as algebraic, geometric and computational considerations. In that last case, results rely on NP-Hardness, and thus go beyond the “hardness” of factoring integers on which rely some popular cryptographic techniques \cite{4}. Finally, rados are cryptography-compliant: homomorphic encryption schemes can be used to compute rados in the encrypted domain from encrypted edge vectors or examples — rado computation can thus be easily distributed in secure multiparty computation applications.

9 Acknowledgments

The authors are indebted to Tiberio Caétano for early discussions that brought the idea of Rademacher observations and their use in privacy related applications. Thanks are also due to Stephen Hardy and Hugh Durrant-Whyte for many stimulating discussions and feedback on the subject. NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Center of Excellence Program.

References

[1] R. Arratia and L. Gordon. Tutorial on large deviations for the binomial distribution. \textit{Bulletin of Mathematical Biology}, 51:125–131, 1989.

[2] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[3] J.-D. Boissonnat, F. Nielsen, and R. Nock. Bregman voronoi diagrams. \textit{DCG}, 44(2):281–307, 2010.

[4] R. Bost, R.-A. Popa, S. Tu, and S. Goldwasser. Machine learning classification over encrypted data. Cryptology ePrint Archive, Report 2014/331, 2014.

[5] S. Boucheron, G. Lugosi, and P. Massart. \textit{Concentration Inequalities: A Nonasymptotic Theory of Independence}. Oxford university press, 2013.

[6] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In \textit{Advanced Lectures on Machine Learning — ML Summer Schools 2003, Canberra, Australia}, pages 169–207, 2003.

[7] K. Chaudhuri and D. Hsu. Sample complexity bounds for differentially private learning. In \textit{Proc. of the 24 th COLT}, pages 155–186, 2011.
[8] D.-L. Donoho. Compressed sensing. *IEEE T. IT*, 52(4):1289–1306, 2006.

[9] D.-L. Donoho and J. Tanner. Sparse non-negative solution of underdetermined linear equations by linear programming. *PNAS*, 102:9446–9451, 2005.

[10] J.-C. Duchi, M.-I. Jordan, and M. Wainwright. Privacy-aware learning. *J. ACM*, 2014.

[11] C. Duchi, M.-I. Jordan, and M. Wainwright. Privacy-aware learning. *J. ACM*, 2014.

[12] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proc. of the 3rd TCC*, pages 265–284, 2006.

[13] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9:211–407, 2014.

[14] M. Enserink and G. Chin. The end of privacy. *Science*, 347:490–491, 2015.

[15] D.-L. Goroff. Balancing privacy versus accuracy in research protocols. *Science*, 347:479–480, 2015.

[16] M. Hardt and E. Price. The noisy power method: a meta algorithm with applications. In *NIPS* 27, pages 2861–2869, 2014.

[17] N. Homer, S. Szollosi, S. Redman, D. Duggan, W. Tembe, J. Muehling, J.-V. Pearson, D.-A. Stephan, S.-F. Nelson, and D.-W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genetics*, 4:e100167, 2008.

[18] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B.-C. Pierce, and A. Roth. Differential privacy: An economic method for choosing epsilon. In *Proc. of the 27th IEEE CSFS*, pages 398–410, 2014.

[19] S. Landau. Control use of data to protect privacy. *Science*, 347:504–506, 2015.

[20] C. McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 1–54. Springer Verlag, 1998.

[21] J.-J. Nietfeld, J. Sugarman, and J.-E. Litton. The bio-pin, a concept to improve biobanking. *Nature Reviews Cancer*, 11:303–308, 2011.

[22] R. Nock and F. Nielsen. A Real Generalization of discrete AdaBoost. *Artificial Intelligence*, 171:25–41, 2007.

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

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

[25] K. Pillaiapakkamnatt and V. Raghavan. On the limits of proper learnability of subclasses of DNF formulae. In *Proc. of the 7th COLT*, pages 118–129, 1994.

[26] N. Quadrianto, A.-J. Smola, T.-S. Caetano, and Q.-V. Le. Estimating labels from label proportions. *JMLR*, 10:2349–2374, 2009.
10 Appendix — Proofs

To simplify the proofs, we define the following quantity:

$$ \tilde{\pi}_\sigma = \sum_i \sigma_i x_i, \forall \sigma \in \Sigma_m. \quad (24) $$

so that each rado can be defined as: $$ \pi_\sigma = (1/2) \cdot (\tilde{\pi}_\sigma + \tilde{\pi}_y). $$ We recall that $$ y $$ is the label vector.

10.1 Proof of Lemma 2

We have

$$ F_{\log}(S, \theta) \equiv \frac{1}{m} \sum \log \left( \frac{1 + \exp \left( \frac{1}{2} \cdot y^\top \theta x_i \right)}{1 + \exp \left( \frac{1}{2} \cdot y^\top \theta x_i \right)} \right) $$

$$ = \frac{1}{m} \sum \log \left( \sum_{y \in \{-1,1\}} \exp \left( \frac{1}{2} \cdot y^\top \theta x_i \right) \right) - \frac{1}{m} \cdot \frac{1}{2} \cdot \theta^\top \tilde{\pi}_y $$

$$ = \frac{1}{m} \log \sum_{\sigma \in \Sigma_m} \exp \left( \frac{1}{2} \cdot \theta^\top \tilde{\pi}_\sigma \right) - \frac{1}{m} \cdot \frac{1}{2} \cdot \theta^\top \tilde{\pi}_y $$

$$ = \frac{1}{m} \log \sum_{\sigma \in \Sigma_m} \exp \left( \frac{1}{2} \cdot \theta^\top (\tilde{\pi}_\sigma - \tilde{\pi}_y) \right) $$

$$ = \frac{1}{m} \log \sum_{\sigma \in \Sigma_m} \exp \left( \frac{1}{2} \cdot \theta^\top (\tilde{\pi}_\sigma + \tilde{\pi}_y) \right) $$

$$ = \log(2) + \frac{1}{m} \log \frac{1}{2m} \sum_{\sigma \in \Sigma_m} \exp \left( \frac{1}{2} \cdot \theta^\top (\tilde{\pi}_\sigma + \tilde{\pi}_y) \right) $$

$$ = \log(2) + \frac{1}{m} \log \frac{1}{2m} \sum_{\sigma \in \Sigma_m} \exp \left( \theta^\top \pi_\sigma \right) $$

$$ = \log(2) + \frac{1}{m} \log F^r_{\exp}(S, \theta, \Sigma_m). $$

We refer to (24) (Lemma 1) for the proof of (25). Eq. (26) holds because $$ \Sigma_m $$ is closed by negation.
10.2 Proof of Theorem 3

Let us suppose that our set of rados \( \mathcal{U} \) satisfies:

\[
\mathcal{U} \subseteq \Sigma_r \subseteq \Sigma_m ,
\]

where \( \Sigma_r \) is a fixed reference subset of \( \Sigma_m \). We shall use the shorthand \( \mathbb{E}_U[f(U)] \) to denote uniform i.i.d. sampling of \( U \) in \( \Sigma_r \). Furthermore, we also let for short

\[
\ell = \sup_{\theta \in \Theta} \max_{\pi_\sigma \in \Sigma_r} \exp(-\theta^\top \pi_\sigma) .
\]

The proof relies on basic knowledge of VC theory and the “symmetrization trick”, which can be found e.g. in ([6]). Plugging eq. (28) into the proof of the symmetrization Lemma (Lemma 2 in ([6])) yields the following symmetrization Lemma for the exponential rado-loss. Notice that the assumption is the same as in Lemma 2 in ([6]).

**Lemma 13** For any fixed sample \( S \), for any \( t \) such that \( nt^2 \geq 2 \), the following holds over the Rademacher sampling of \( \sigma \) in \( \Sigma_m \):

\[
P \left[ \sup_{\theta \in \Theta} \left( \mathbb{E}_U[F^r_{\text{exp}}(S, \theta, U)] - F^r_{\text{exp}}(S, \theta, U) \right) \geq t \right] 
\leq 2\ell^2 \cdot P \left[ \sup_{\theta \in \Theta} (F^r_{\text{exp}}(S, \theta, U) - F^r_{\text{exp}}(S, \theta, U')) \geq \frac{t}{2} \right],
\]

where \( U, U' \) are two size-\( n \) i.i.d. samples.

Consider \( \mathcal{U}, \mathcal{U}' \subseteq \Sigma_r \), each of cardinal \( n \) and differing from one assignment only. Then it follows, for any \( \theta \in \Theta \) and from ineq. (29):

\[
|F^r_{\text{exp}}(S, \theta, U) - F^r_{\text{exp}}(S, \theta, U')| \leq \frac{2\ell}{n} .
\]

Applying the independent bounded differences inequality ([20]), we get, for any \( \theta \in \Theta \) and \( t > 0 \):

\[
P \left[ \mathbb{E}_U[F^r_{\text{exp}}(S, \theta, U)] - F^r_{\text{exp}}(S, \theta, U) \geq \frac{t}{4} \right] \leq \exp \left( -\frac{nt^2}{16\ell^2} \right) .
\]

Letting \( \Pi(n) \) denote the growth function for linear separators computed over rados, we still have the upperbound

\[
\Pi(n) \leq \left( \frac{en}{d+1} \right)^{d+1} .
\]
We thus get, for any $\theta \in \Theta$:

$$
\mathbb{P}
\left[
\sup_{\theta \in \Theta}
\left]E_U \left[F_{r,\exp}(S, \theta, U)\right] - F_{r,\exp}(S, \theta, U)\right) \geq t \right]
\leq 2\ell^2 \cdot \mathbb{P}
\left[
\sup_{\theta \in \Theta}
\left]F_{r,\exp}(S, \theta, U) - F_{r,\exp}(S, \theta, U')\right) \geq \frac{t}{2} \right)
\leq 2\Pi(2n)\ell^2 \cdot \mathbb{P}
\left[
\left]F_{r,\exp}(S, \theta, U) - F_{r,\exp}(S, \theta, U')\right) \geq \frac{t}{2} \right)
\leq 4\Pi(2n)\ell^2 \cdot \mathbb{P}
\left[
\left]E_U \left[F_{r,\exp}(S, \theta, U)\right] - F_{r,\exp}(S, \theta, U)\right) \geq \frac{t}{4} \right)
\leq 4\Pi(2n)\ell^2 \cdot \exp\left(-\frac{n\ell^2}{16\ell^2}\right)
\leq 4 \left(\frac{2\ell}{d + 1}\right)^{d + 1} \ell^2 \cdot \exp\left(-\frac{n\ell^2}{16\ell^2}\right)
$$

(32)

(33)

(34)

(35)

(36)

Ineq. (32) follows from Lemma 13, ineq. (33) follows from standard VC arguments (see e.g. (13)), ineq. (34) follows from the observation that event $a - b \geq u$ implies $(a - c \geq u/2) \lor (b - c \geq u/2)$, ineq. (35) follows from (30), and finally ineq (36) follows from ineq. (31). Picking

$$
t = t_* \doteq 16\ell \cdot \sqrt{\frac{1}{n} \log \ell + \frac{d}{n} \log \frac{2en}{d} + \frac{1}{n} \log \frac{1}{\eta}}
$$

(37)

yields that the right hand-side of ineq. (36) is not more than $\eta$, for any $\eta > 0$. So with probability $\geq 1 - \eta$, any classifier $\theta \in \Theta$ will enjoy $E_U \left[F_{r,\exp}(S, \theta, U)\right] \leq F_{r,\exp}(S, \theta, U) + t_*$, and so we shall have:

$$
F_{r,\exp}(S, \theta, U)
\doteq \log(2) + \frac{1}{m} \cdot \log F_{r,\exp}(S, \theta, U)
\geq \log(2) + \frac{1}{m} \cdot \log \left(E_U \left[F_{r,\exp}(S, \theta, U)\right] - t_*\right)
\doteq \log(2) + \frac{1}{m} \cdot \log \left(E_U \left[F_{r,\exp}(S, \theta, U)\right]\right)
\doteq \log(2) + \frac{1}{m} \cdot \log \left(1 - 16\varrho \cdot \sqrt{\frac{1}{n} \log \ell + \frac{d}{n} \log \frac{2en}{d} + \frac{1}{n} \log \frac{1}{\eta}}\right)
\doteq \log(2) + \frac{1}{m} \cdot \log F_{r,\exp}(S, \theta, \Sigma_m)
\doteq F_{\log}(S, \theta) + \frac{1}{m} \cdot \log \left(F_{r,\exp}(S, \theta, \Sigma_r) / F_{r,\exp}(S, \theta, \Sigma_m)\right)
\doteq F_{\log}(S, \theta) + \frac{1}{m} \cdot \log \left(1 - 16\varrho \cdot \sqrt{\frac{1}{n} \log \ell + \frac{d}{n} \log \frac{2en}{d} + \frac{1}{n} \log \frac{1}{\eta}}\right)
$$

(38)

(39)

In eq. (38), we use the fact that $\varrho = \ell / E_U \left[F_{r,\exp}(S, \theta, U)\right]$ and $E_U \left[F_{r,\exp}(S, \theta, U)\right] = F_{r,\exp}(S, \theta, \Sigma_r)$. Hence, reordering the expression yields that with probability $\geq 1 - \eta$, the final classifier $\theta$ will
satisfy:

\[
F_{\log}(S, \theta) \leq F_{\log}^r(S, \theta, U) - \frac{1}{m} \log \left( F_{\exp}^r(S, \theta, \Sigma_r) - \frac{F_{\exp}^r(S, \theta, \Sigma_m)}{F_{\exp}^r(S, \theta, \Sigma_m)} \right) - \log \left( 1 - 16 \cdot \frac{\theta}{\sqrt{n}} \cdot \sqrt{\log \ell + d \log \frac{2en}{d} + \log \frac{1}{\eta}} \right). \tag{40}
\]

There remains to use the fact that \( \ell \leq \exp(r_\theta \max_{\Sigma_r} \| \pi_{\sigma} \|_2) \) to complete the proof of ineq. (5) in Theorem 3. To prove ineq. (8), let us call \( 1 - z \) the quantity inside the log in ineq. (40). We clearly have to have \( 0 \leq z < 1 \), and so for any value of \( z \) and for any \( 0 \leq \alpha < 1 \), there exists a value \( m^* > 0 \) such that

\[
m^{1-\alpha} \geq \frac{1}{z} \log \frac{1}{1 - z} \geq 0, \tag{41}
\]

for any \( m \geq m^* \). In this case, we get after reordering, since \( 1 - z' \leq \exp z' \),

\[
1 - \frac{z}{m^\alpha} \leq \exp \left( - \frac{z}{m^\alpha} \right) \leq \exp \left( \frac{1}{m} \log (1 - z) \right), \tag{42}
\]

and so, taking logs and using ineq. (39), we obtain that for any \( 0 \leq \beta < 1/2 \), there exists \( m^* > 0 \) such that for any \( m \geq m^* \):

\[
F_{\log}(S, \theta) \leq F_{\log}^r(S, \theta, U) - \frac{1}{m} \log \left( F_{\exp}^r(S, \theta, \Sigma_r) - \frac{F_{\exp}^r(S, \theta, \Sigma_m)}{F_{\exp}^r(S, \theta, \Sigma_m)} \right) - \log \left( 1 - 16 \cdot \frac{\theta}{m^\beta} \cdot \sqrt{\frac{r_\theta}{n} \cdot \max_{\Sigma_r} \| \frac{1}{m} \cdot \pi_{\sigma} \|_2 + \frac{d}{nm} \log \frac{2en}{d} + \frac{1}{nm} \log \frac{1}{\eta}} \right). \tag{43}
\]

Calling \( 1 - z' \) the quantity inside the log, there remains to use \( \log (1 - z') \geq -Kz' \) for some \( K > 0 \) when \( z' \) is sufficiently close to 0 (hence, \( m \) sufficiently large again). This proves ineq. (8) and completes the proof of Theorem 3. Remark that provided \( n \) is sufficiently large, the right hand-side of ineq (41) admits the following equivalent:

\[
\frac{1}{z} \log \frac{1}{1 - z} \sim 1 + \frac{z}{2}, \tag{44}
\]

with \( z = \Omega(1/\sqrt{n}) \) (omitting the dependences in the other parameters). Hence, ineq (41) can be ensured as long as \( m \) is large enough with respect to \( r_\theta, \max_{\Sigma_r} \| (1/m) \cdot \pi_{\sigma} \|_2 \) (which cannot exceed the maximum norm of an observation in \( S \), \( d \) and \( \log (1/\eta) \)).

So, when we apply this last result to RADOBOOST, it says that for a large enough sample, we can indeed pick an \( n \) sufficiently large but small compared to \( m \) so that we shall observe with high probability a decay rate of the expected logistic loss computed over \( S \), \( \mathbb{E}[F_{\log}(S, \theta_T)] \), of order \( \Omega(\gamma^2/m) \) (expectation is measured with respect to the sampling of \( U \)).
We are now left with proving ineq. (7), and so we study:

\[-Q = \frac{1}{m} \cdot \log \left( \frac{\frac{1}{|\Sigma_r|} \sum_{\sigma' \in \Sigma_r} \exp(-\theta^\top \pi_{\sigma'})}{\frac{1}{|\Sigma_m|} \sum_{\sigma \in \Sigma_m} \exp(-\theta^\top \pi_{\sigma})} \right)\]

\[= \frac{1}{m} \cdot \log \left( \frac{|\Sigma_m| \sum_{\sigma' \in \Sigma_r} \exp(-\theta^\top \pi_{\sigma'})}{|\Sigma_r| \sum_{\sigma \in \Sigma_m} \exp(-\theta^\top \pi_{\sigma})} \right)\]

\[= \frac{1}{m} \cdot \log \left( \frac{\sum_{\sigma \in \Sigma_m} \sum_{\sigma' \in \Sigma_r} \exp(-\theta^\top \pi_{\sigma}) \cdot \exp(-\theta^\top (\pi_{\sigma'} - \pi_{\sigma}))}{\sum_{\sigma' \in \Sigma_r} \sum_{\sigma \in \Sigma_m} \exp(-\theta^\top \pi_{\sigma})} \right)\]

\[= \frac{1}{m} \cdot \log \left( \mathbb{E}_{(\sigma, \sigma') \sim D} \left[ \exp(-\theta^\top (\pi_{\sigma'} - \pi_{\sigma})) \right] \right),\]

with $D(\sigma, \sigma') \propto \exp(-\theta^\top \pi_{\sigma})$. Jensen’s inequality yields:

\[-Q \geq \frac{1}{m} \cdot \mathbb{E}_{(\sigma, \sigma') \sim D} \left[ -\theta^\top (\pi_{\sigma'} - \pi_{\sigma}) \right]\]

\[= \frac{1}{m} \cdot \mathbb{E}_{(\sigma, \sigma') \sim D} \left[ \theta^\top \pi_{\sigma} \right] - \frac{1}{m} \cdot \mathbb{E}_{(\sigma, \sigma') \sim D} \left[ \theta^\top \pi_{\sigma'} \right].\]  \hspace{1cm} (45)

We now remark that

\[
\mathbb{E}_{(\sigma, \sigma') \sim D} \left[ \theta^\top \pi_{\sigma'} \right] = \frac{\sum_{\sigma' \in \Sigma_r} \sum_{\sigma \in \Sigma_m} \exp(-\theta^\top \pi_{\sigma}) \cdot \theta^\top \pi_{\sigma'}}{\sum_{\sigma' \in \Sigma_r} \sum_{\sigma \in \Sigma_m} \exp(-\theta^\top \pi_{\sigma})} = \theta^\top \left( \frac{\sum_{\sigma' \in \Sigma_r} \pi_{\sigma'}}{\sum_{\sigma \in \Sigma_m} \exp(-\theta^\top \pi_{\sigma})} \right) = \theta^\top \mathbb{E}_{\sigma' \sim \Sigma_r} \left[ \pi_{\sigma'} \right],\]  \hspace{1cm} (46)

and furthermore

\[
\frac{1}{m} \cdot \mathbb{E}_{(\sigma, \sigma') \sim D} \left[ \theta^\top \pi_{\sigma} \right] = \frac{1}{m} \cdot \sum_{\sigma' \in \Sigma_r} \sum_{\sigma \in \Sigma_m} \exp(-\theta^\top \pi_{\sigma}) \cdot \theta^\top \pi_{\sigma} = \frac{1}{m} \cdot \sum_{\sigma \in \Sigma_m} \exp(-\theta^\top \pi_{\sigma}) \cdot \theta^\top \pi_{\sigma} = \theta^\top \left( \frac{1}{m} \cdot \sum_{\sigma \in \Sigma_m} \exp(-\theta^\top \pi_{\sigma}) \cdot \pi_{\sigma} \right) = \theta^\top \nabla_{\theta} \frac{1}{m} \cdot \log F_{\text{exp}}^r (S, \theta, \Sigma_m) = \theta^\top \nabla_{\theta} F_{\text{log}}^r (S, \theta, \Sigma_m).\]  \hspace{1cm} (47)

Assembling eqs (46) and (47), we get from ineq. (45):

\[Q \leq r_{\theta} \left\| \nabla_{\theta} F_{\text{log}}^r (S, \theta, \Sigma_m) \right\|_2 - \mathbb{E}_{\sigma \sim \Sigma_r} \left[ \frac{1}{m} \cdot \pi_{\sigma} \right] \|_2\]

\[\leq r_{\theta} \left( \left\| \nabla_{\theta} F_{\text{log}}^r (S, \theta, \Sigma_m) \right\|_2 + \left\| \mathbb{E}_{\sigma \sim \Sigma_r} \left[ \frac{1}{m} \cdot \pi_{\sigma} \right] \right\|_2 \right),\]

as claimed.
10.3 Proof of Lemma 4

Theorem 1 in ([22]) immediately yields

\[
\frac{1}{n} \exp \left( -\theta^\top T \pi_j \right) \leq \prod_{t=1}^{T} \sqrt{1 - r^2_t} \cdot w_{(T+1)j}, \forall j \in [n] .
\]  

(48)

Since \( 1^\top w_{T+1} = 1 \), summing over \( j \in [n] \) yields:

\[
F^r_{\exp}(S, \theta_T, U) \leq \prod_{t=1}^{T} \sqrt{1 - r^2_t} \leq \exp \left( \frac{-1}{2} \sum_{t} r^2_t \right).
\]

Using the (WLA), this yields ineq. (12).

10.4 Proof of Lemma 5

Fix for short \( k = \iota(t) \). We rewrite \( r_t(w_t) \) as a function of the examples:

\[
\begin{align*}
r_t(w_t) &= \frac{1}{\pi_{sk}} \sum_{j=1}^{n} w_{tj} \pi_{jk} \\
&= \frac{1}{\pi_{sk}} \sum_{j=1}^{n} \sum_{i: \sigma_{ji} = y_i} w_{tj} y_i x_{ik} \\
&= \frac{1}{x_{sk} \pi_{sk}} \sum_{i=1}^{m} \left( \frac{x_{sk}}{\pi_{sk}} \cdot \sum_{j: \sigma_{ji} = y_i} w_{tj} \right) y_i x_{ik} .
\end{align*}
\]

(49)

Define \( \tilde{w} \in \mathbb{P}^m \) such that

\[
\tilde{w}_i = \frac{1}{\tilde{W}} \cdot \frac{x_{sk}}{\pi_{sk}} \cdot \sum_{j: \sigma_{ji} = y_i} w_{tj}, \forall i \in [m] ,
\]

(50)

with

\[
\tilde{W} = \frac{x_{sk}}{\pi_{sk}} \cdot \sum_{i=1}^{m} \sum_{j: \sigma_{ji} = y_i} w_{tj} \]

\[
= \frac{x_{sk}}{\pi_{sk}} \cdot \sum_{j=1}^{n} w_{tj} |\{i : \sigma_{ji} = y_i\}|
\]

(51)

the normalization coefficient. Because \( w_t \in \mathbb{P}^n \), \( x_{sk} > 0 \) and \( \pi_{sk} > 0 \), it comes that indeed \( \tilde{w} \in \mathbb{P}^m \), and \( \tilde{W} > 0 \) (unless \( S^r \) is reduced to the null rado). We thus have \( |r_t(w_t)| \geq \gamma \) iff

\[
|r^e_t(\tilde{w})| \geq \frac{\gamma}{\tilde{W}} .
\]

(52)

This proves the statement of the Lemma. Remark that

\[
\frac{x_{sk}}{\pi_{sk}} \leq \tilde{W} \leq \left( \frac{x_{sk}}{\max_{j \in [n]} |\{i : \sigma_{ji} = y_i\}|} \right)^{\frac{1}{\gamma}} ,
\]

(53)
so if we assume the weak learning assumption holds for the examples, \( |r^x(\tilde{w})| \geq \gamma^{ex} > 0 \), then the weak learning assumption over rados always holds for

\[
\gamma = \frac{x_{*k}}{\pi_{*k}} \cdot \gamma^{ex},
\]

and may holds for a value \( \gamma \) which can be as large as

\[
\gamma = \frac{x_{*k}}{\max_j \{\{i:\sigma_{ji} = y_i\}\}} \cdot \gamma^{ex}.
\]

These two bounds are data dependent (but they depend on data only), and whenever they are significant outlier values for feature \( k \), i.e. \( x_{*k} \) is achieved by few examples and all others have feature value of significantly smaller order, then the available \( \gamma \) can be significantly larger than \( \gamma^{ex} \).

Compared to the cases where no such outliers would exist, we thus may expect significantly better results for RadoBoost.

10.5 Proof of Theorem 6

To ease notations hereafter, we consider wlog that \( d = 1 \) and so \( j_* = 1 \). We also drop index notation \( j_* \) in related notations (so \( \Sigma_{m}^{\beta,j_*} \) becomes \( \Sigma_{m}^{\beta} \)).

We let \( S \) and \( S' \) denote two \( j \)-neighbors, so that \( S \approx_j S' \) holds and they differ by the value of one (boolean) feature. Algorithm DP-Feat selects uniformly at random the rados in sets

\[
\Sigma_{m}^{\beta}(S) = \{ \sigma \in \Sigma_{m} : \pi_{\sigma} \in \mathbb{I}(S) \},
\]

\[
\Sigma_{m}^{\beta}(S') = \{ \sigma \in \Sigma_{m} : \pi_{\sigma} \in \mathbb{I}(S') \},
\]

with

\[
\mathbb{I}(S) = \{-(m - m(+) + \beta(m + 1) \leq z \leq m(+) - \beta(m + 1)\},
\]

\[
\mathbb{I}(S') = \{-(m - m(+) + \beta(m + 1) + \zeta \leq z \leq m(+) - \beta(m + 1) + \zeta)\},
\]

since \( m'(+) = m(+) + \zeta \) for some \( \zeta \in \{-1, 0, 1\} \). To relate the sizes of these two sets, we first compute the size of \( \{\sigma : \pi_{\sigma} = r|S\} \), for \( r \in \mathbb{Z} \). Assuming first \( r \geq 0 \), we have:

\[
|\{\sigma : \pi_{\sigma} = r|S\}| = \min\{m(+) - r, m - m(+)\} \sum_{i=0}^{\min\{m(+) - r, m - m(+)\}} \binom{m(+) - i - r}{i} \binom{m - m(+) - r}{i}. \tag{60}
\]

If \( r < 0 \), then similarly:

\[
|\{\sigma : \pi_{\sigma} = r|S\}| = \min\{m(+) + m - m(+) + r\} \sum_{i=0}^{\min\{m(+) + m - m(+) + r\}} \binom{m(+) + i}{i} \binom{m - m(+) - r}{i - r}, \tag{61}
\]

which is the same expression as \(60\) with the substitutions \( r \mapsto -r \), \( m(+) \mapsto m - m(+) \), \( m - m(+) \mapsto m(+) \), so we have only to analyse the case \( r \geq 0 \). If \( m(+) + r > m - m(+) \), we have by Vandermonde identity:

\[
|\{\sigma : \pi_{\sigma} = r|S\}| = \sum_{i=0}^{m - m(+)} \binom{m(+)}{m(+) - r} \binom{m - m(+) - r}{i} = \binom{m}{m(+) - r}. \tag{62}
\]
If \( m(+) - r \leq m - m(+) \), then it is not hard to show that Vandermonde identity still brings (62). We thus have

\[
|\Sigma_m^{\beta}(S)| = \sum_{r=-(m-m(+) + \beta(m+1))}^{m(+) - \beta(m+1)} \binom{m}{m(+) - r}
\]

\[
= \left( \frac{m}{\beta(m+1)} \right) + \sum_{r=-(m-m(+) + \beta(m+1)+1)}^{m(+) - \beta(m+1)} \binom{m}{m(+) - r}
\]

\[
= \frac{m - \beta(m + 1) + 1}{\beta(m+1)} \cdot \left( \binom{m}{\beta(m+1) - 1} + \sum_{r=-(m-m(+) + \beta(m+1)+1)}^{m(+) - \beta(m+1)} \binom{m}{m(+) - r} \right)
\]

\[
\geq \left( \frac{m}{\beta(m+1) - 1} \right) + \sum_{r=-(m-m(+) + \beta(m+1)+1)}^{m(+) - \beta(m+1)} \binom{m}{m(+) - r}
\]

\[
= \sum_{r=-(m-m(+) + \beta(m+1)+1)}^{m(+) - \beta(m+1)+1} \binom{m}{m(+) - r}
\]

\[
= \sum_{r=-(m-m(+) + \beta(m+1)+1)}^{m(+) - \beta(m+1)+1} \binom{m(+) + 1 - r}{m - m(+) + r} \cdot \binom{m}{(m(+) + 1) - r}
\]

\[
(63)
\]

\[
\geq \sum_{r=-(m-m(+) + \beta(m+1)+1)}^{m(+) - \beta(m+1)+1} \frac{\beta(m+1)}{m - \beta(m+1) + 1} \cdot \binom{m}{(m(+) + 1) - r}
\]

\[
(64)
\]

\[
= \left( \frac{1}{\beta} - 1 \right)^{-1} \sum_{r=-(m-m(+) + \beta(m+1)+1)}^{m(+) - \beta(m+1)+1} \binom{m}{(m(+) + 1) - r}
\]

\[
(65)
\]

\[
= \left( \frac{1}{\beta} - 1 \right)^{-1} \cdot |\Sigma_m^{\beta}(S')|
\]

(66)
if $\zeta = 1$, and

$$|\Sigma^\beta_m(S)| = \sum_{r=-(m-m(+)) + \beta(m+1)}^{m(+)-\beta(m+1)} \binom{m}{m(+)-r}$$

$$= \frac{m - \beta(m+1) + 1}{\beta(m+1)} \cdot \left( \frac{m}{\beta(m+1) - 1} \right)^{-1} \sum_{r=-(m-m(+)) + \beta(m+1)}^{m(+)-\beta(m+1)-1} \binom{m}{m(+)-r}$$

$$= \frac{1}{\beta - 1} \cdot \left( \frac{m}{\beta(m+1) - 1} \right)^{-1} \sum_{r=-(m-m(+)) + \beta(m+1)}^{m(+)-\beta(m+1)-1} \binom{m}{m(+)-r}$$

$$\geq \left( \frac{1}{\beta - 1} \right)^{-1} \cdot |\Sigma^\beta_m(S')|$$

if $\zeta = -1$. The last inequality follows from the same chain of inequalities as in eqs. (63 – 66). We now bound the ratio of probabilities for the rado being equal to $r$, for both sets:

$$\frac{\mathbb{P}_{\sigma \sim \Sigma^\beta_m(S)}[\pi_{\sigma} = r | S]}{\mathbb{P}_{\sigma \sim \Sigma^\beta_m(S')}[\pi_{\sigma} = r | S']} = \frac{|\Sigma^\beta_m(S')|}{|\Sigma^\beta_m(S)|} \cdot \frac{\binom{m}{m(+)-r}}{\binom{m}{m(+)-\zeta-r}}$$

$$\leq \left( \frac{1}{\beta - 1} \right) \cdot \frac{m}{m(+)-\zeta-r} \cdot \frac{\binom{m}{m(+)-r}}{\binom{m}{m(+)-\zeta-r}}$$

$$= \left( \frac{1}{\beta - 1} \right) \cdot \frac{(m(+)+\zeta-r)! (m-m(+)-\zeta+r)!}{(m(+)-r)! (m-m(+)+r)!}$$

$$= \left( \frac{1}{\beta - 1} \right) \cdot \left\{ \begin{array}{ll} \frac{m(+)+1-r}{m-m(+)+r} & \text{if } \zeta = 1 \\ 1 & \text{if } \zeta = 0 \\ \frac{m-m(+)+1+r}{m(+)-r} & \text{if } \zeta = -1 \end{array} \right.$$

$$\leq \left( \frac{1}{\beta - 1} \right)^2.$$
as well. We finally get from ineq. (69):

\[
\frac{\mathbb{P}_{\sigma \sim \Sigma_m^\beta(S)}[\pi_{\sigma} = r|S]}{\mathbb{P}_{\sigma \sim \Sigma_m^\beta(S')}[\pi_{\sigma} = r|S']} \leq \exp(\epsilon),
\]

which holds for any \( r \in \Sigma_m^\beta(S) \cap \Sigma_m^\beta(S') \). Notice however that the symmetric difference of these two sets is not empty. To finish the proof, we need to take into account this symmetric difference. This is the data-dependent step in DP-Feat which may leak information about one feature and disclose its content, through the use of eq. (58). To see this, if we assume that one possesses all the data but the unknown feature value for one person, and knows how rados are computed using DP-Feat, then by observing the output \( \pi_{\sigma,j^*} \), he may guess the unknown value, as depicted by Figure 3. Let us denote \( A \) this event. When returning one rado from \( \Sigma_m^\beta(\cdot) \), if we consider without loss of generality a uniform distribution over examples, then, referring to the notations of Figure 3, we have:

\[
P[A] = P[A|S]P[S] + P[A|S']P[S']
\]

(73)

and

\[
P[A|S] = \frac{(m-\beta(m+1))}{\sum_{r=\beta(m+1)}^{m} \binom{m}{r}}
\]

(75)

and we obtain following the same reasoning, using the fact that \( m(+) \) increases by one in \( S' \),

\[
P[A|S'] = \frac{(\beta(m+1))}{\sum_{r=\beta(m+1)}^{m} \binom{m}{r}}
\]

(76)

The probability of hitting the symmetric difference of \( \Sigma_m^\beta(S) \cap \Sigma_m^\beta(S') \) is taken into account considering \( \delta = P[A] \) in the \((\epsilon, \delta)\)-differentially private release of one rado. We get:

\[
\delta < \frac{2^{\beta(m+1)}}{\sum_{r=\beta(m+1)}^{m} \binom{m}{r}}.
\]

(77)

The interplay between \( \epsilon \) and \( \delta \) can be appreciated throughout the use of the following properties:

\[
\sum_{r=0}^{\beta(m+1)-1} \binom{m}{r} \leq 2^m H(u),
\]

(78)

\[
\binom{m}{m/2} < \frac{1}{\sqrt{m}} \cdot 2^m,
\]

(79)

we have used

\[
H(z) = -z \log_2 z - (1 - z) \log_2(1 - z),
\]

\[
u = \beta - \frac{1 - \beta}{m}.
\]
We get
\[ \delta < \frac{2}{\sqrt{m}} \cdot \frac{1}{1 - 2^{m(H(u) - 1)}} \]  
(80)

Because \( H(u) \) is concave, it satisfies (fixing \( \epsilon' = \epsilon/2 \) for short):
\[
H(u) \leq H(\beta) + (u - \beta)H'(\beta) \\
= H(\beta) - \frac{1 - \beta}{m} \log_2 \frac{1 - \beta}{\beta} \\
= H(\beta) - \frac{(1 - \beta)\epsilon'}{m} \\
= \frac{1}{\log 2} \left( \log(1 + \exp(\epsilon')) - \left(1 + \frac{1}{m}\right) \cdot \frac{\epsilon' \exp \epsilon'}{1 + \exp \epsilon'} \right) \equiv f(\epsilon') .
\]  
(81)

We have:
\[
\frac{1}{1 - 2^{m(f(\epsilon') - 1)}} \sim_0 \frac{1}{2m^2 \log^2(2)\epsilon'} + \left(\frac{1}{2} - \frac{1}{4m^2 \log^3(2)}\right) + O(\epsilon') .
\]  
(82)

So, assuming \( \epsilon' = o(1) \), there exists \( m' > 0 \) and a constant \( K > 0 \) such that for any \( m > m' \),
\[ \delta < K \cdot \frac{1}{m^2 \epsilon} . \]  
(83)

Finally, we get that when \( \epsilon = \Omega(1/m) \), \((\epsilon, \delta)\)-differential privacy can be ensured on the delivery of \( n = 1 \) rado as long as \( \epsilon \cdot \delta = O(m^{-5/2}) \). Taking into account the fact that rados are generated independently and using Theorem 3.16 in [12] concludes the proof of Theorem 6 for arbitrary \( n \).

To finish the proof, we remark that \( \Sigma_m(\cdot) \neq \emptyset \). Indeed, since \( m \geq 1, \beta < m/(m + 1) \); furthermore, as long as \( m > 2 \), provided we also have
\[
\frac{1 + 2\beta}{1 - 2\beta} = O(m) ,
\]
we shall have \( I(S) \cap \mathbb{Z} \neq \emptyset \). This can easily be ensured if
\[ \frac{1}{\epsilon} + \epsilon = O(m) , \]  
(84)

i.e., provided \( \epsilon = o(1) \), \( \epsilon = \Omega(1/m) \).

10.6 Proof of Theorem 7

We keep the same notations as in the proof of Theorem 6. The Rademacher rejection sampling of \( \sigma \) has a probability to reject a single rado bounded by (a fraction of) the tail of the Binomial, as
indeed

\[ \mathbb{P}_{\sigma \sim \sum_m \left[ \sigma \notin \sum_{m+1}^{\beta} | S \right] } = \frac{1}{2m} \cdot \sum_{r < -(m-k+1)+\beta(m+1)}^{-(m-k+1)+\beta(m+1)-1} \left( \frac{m}{m+1-r} \right) \]

\[ = \frac{1}{2m} \cdot \sum_{r = -(m-k+1)}^{m} \left( \frac{m}{m+1-r} \right) + \frac{1}{2m} \cdot \sum_{r = m-k+1}^{m+1} \left( \frac{m}{m+1-r} \right) \]

\[ = 2 \cdot \frac{1}{2m} \cdot \sum_{r = m-k+1}^{m} \left( \frac{m}{r} \right) \]

\[ = 2 \cdot \frac{1}{2m} \cdot \sum_{r = (1-\beta)(m+1)}^{m+1} \left( \frac{m+1-r}{m+1} \cdot \left( \frac{m+1}{r} \right) \right) \]

\[ \leq 2\beta \cdot \frac{1}{2m} \cdot \sum_{r = (1-\beta)(m+1)}^{m+1} \left( \frac{m+1}{r} \right) \]

\[ \leq 2\beta \cdot \frac{1}{2m} \cdot \sum_{r = (1-\beta)(m+1)}^{m+1} \left( \frac{m+1}{r} \right) \]

\[ = 4\beta \cdot \sum_{r = (1-\beta)(m+1)}^{m+1} \left( \frac{m+1}{r} \right) \cdot \left( \frac{1}{2} \right)^{m+1-r} \cdot \left( \frac{1}{2} \right)^r \]

\[ \leq 4\beta \exp \left( -(m+1) \cdot D_{BE}(\beta||1/2) \right), \]

\[ (85) \]
where $D_{BE}$ is the bit-entropy divergence ([3]):

$$
D_{BE}(p\|q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}.
$$

(86)

The last equation follows \textit{e.g.} from Theorem 2 in ([1]). So the probability $p$ that there exists a rado, among the $n$ generated, that was rejected at least $T_r$ times for some $T_r \geq 1$ satisfies

$$
p \leq 4n\beta \sum_{t=T_r}^{\infty} \exp \left( -(m+1) \cdot t \cdot D_{BE}(1-\beta\|1/2) \right)
= 4n\beta \cdot \exp \left( -(m+1) \cdot T_r \cdot D_{BE}(1-\beta\|1/2) \right) \sum_{t=0}^{\infty} \exp \left( -(m+1) \cdot t \cdot D_{BE}(1-\beta\|1/2) \right).
$$

We now use the facts that (i) $m \geq (1 + 2\beta)/(1 - 2\beta)$ (Step 2 in Algorithm DP-Feat), and (ii) function

$$f(z) = 2 \frac{1}{1 - 2z} \cdot \left( \log(2) + (1 - z) \log(1 - z) + z \log z \right)
$$

is convex over $[0, 1/2]$ and has limit tangent $1 - 2\beta$ in $z = 1/2$, so

$$
\exp \left( -(m+1) \cdot D_{BE}(1-\beta\|1/2) \right) \leq \exp \left( -\frac{2}{1 - 2\beta} \cdot \left( \log(2) + (1 - \beta) \log(1 - \beta) + \beta \log \beta \right) \right)
\leq \exp(2\beta - 1) < 1,
$$

and it comes

$$
\sum_{t=0}^{\infty} \exp \left( -(m+1) \cdot t \cdot D_{BE}(1-\beta\|1/2) \right) \leq \frac{1}{1 - \exp(2\beta - 1)},
$$

and so

$$
p \leq \frac{4n\beta}{1 - \exp(2\beta - 1)} \cdot \exp \left( -(m+1) \cdot T_r \cdot D_{BE}(1-\beta\|1/2) \right)
$$

(90)

So, if $n, \beta, \eta$ are such that

$$
n \leq \frac{\eta(1 - \exp(2\beta - 1))}{4\beta},
$$

(91)

then there is probability $\geq 1 - \eta$ that no rado was rejected. Otherwise, with probability $\geq 1 - \eta$, each rado among the $n$ was rejected no more than

$$
T_r^* = \left\lceil \frac{1}{mD_{BE}(1-\beta\|1/2)} \log \frac{4\beta n}{\eta(1 - \exp(2\beta - 1))} \right\rceil
$$

(92)

times. There remains to multiply this bound by the number of rados to get an upperbound on the number of iterations of Rademacher rejection sampling, and we obtain eq. (17). This finishes the proof of Theorem 7.
Figure 4: Left: function $f(\beta)$ as depicted in eq. (93). Right: same function over smaller range, depicting the value of $f$ for $\epsilon = 0.1$ (thick dark line) and $\epsilon = 0.01$ (slim dark line).

Remarks: the actual dependence of eq. (92) on $\beta$ is such that unless $\epsilon$ is extremely close to $\lfloor 1 \rfloor$ in which case the requirement on differential privacy is the strongest, $T^*_r$ does not actually blow up. To see this, let us define

$$f(\beta) \doteq \frac{1}{D_{BE}(1 - \beta\|1/2)} \log \frac{4\beta}{1 - \exp(2\beta - 1)}.$$ (93)

Figure 4 displays $f(\beta)$ over different ranges. One sees that when $\epsilon = 0.1$, provided $m/\log n$ is in the order of thousands and $n \gg e$, then $T^*_r$ is in fact of the order $\log(1/\eta)$, which may be quite small indeed.

10.7 Proof of Theorem 8

Let us first remark that the DP-protection of vector edges by computing noisified example set

$$S^+ \doteq \{(x^+_i, y_i) \doteq (x_i + x^T_i, y_i), i \in [m]\},$$ (94)

where $x^+_i \sim N(0, \varsigma^2 I)$, is equivalent to noisifying edges because label $y \in \{-1, 1\}$ and the pdf of the Gaussian mechanism is invariant by multiplication by $y$.

The key quantity to prove the Theorem is, for any noisified rado $\pi^+_j \doteq (1/2) \cdot \sum_i (\sigma_{ji} + y_i)x^+_i$, the support $m_j \doteq |\{i : \sigma_{ji} = y_i\}|$ of the rado. We also renormalize the leveraging coefficient in RADOBOOST, replacing eq. (10) in RADOBoost pseudocode by:

$$\alpha_t \leftarrow \frac{1}{2\kappa \pi_{sl(t)}} \log \frac{1 + r_t}{1 - r_t},$$ (95)

for some fixed $\kappa \geq 1$.

\footnote{Recall that $\beta = 1/(1 + \exp(\epsilon/2))$ in Step 1 of Algorithm DP-Feat.}
We now embark in the proof of Theorem 8. Lemma 2 in \([22]\) yields
\[
\exp \left( -\theta_T^\top \pi_j \right) = \exp \left( -\theta_T^\top \pi_j + \frac{1}{2} \cdot \theta_T^\top \sum_i (\sigma_{ji} + y_i) x_i^r \right) \\
\leq \left( \prod_{i=1}^{T} \sqrt{1 - r_i^2} \cdot n w_{(T+1)j} \right)^{\frac{1}{\kappa}} \cdot \exp \left( \frac{1}{2} \cdot \theta_T^\top \sum_i (\sigma_{ji} + y_i) x_i^r \right), \forall j \in [n]. \tag{96}
\]
Averaging over \( j \in [n] \) yields:
\[
F_{\exp}^r(S, \theta_T, \mathcal{U}) \leq \left( \prod_{i=1}^{T} \sqrt{1 - r_i^2} \right)^{\frac{1}{\kappa}} \cdot \sum_{j=1}^{n} \frac{1}{n} \cdot w_{(T+1)j} \cdot \exp \left( \frac{1}{2} \cdot \theta_T^\top \sum_i (\sigma_{ji} + y_i) x_i^r \right), \tag{97}
\]
with \( \bar{w}_{(T+1)j} = \frac{1}{n} \sum_{j=1}^{n} w_{(T+1)j} \). The right-hand side of ineq. \([97]\) multiplies two separate quantities, \( A \) which quantifies the performances of \( \theta_T \) in RADOBOOST on the set of noisy rados on which it was trained, and \( B \) which is an expectation, computed over \( \mathbf{w}_T \), of the agreements between \( \theta_T \) and the noisy part of the rados. When rados are noise-free and \( \kappa \geq 1 \), we have \( x_i^r = 0 \), \( \forall i \) and
\[
\sum_{j=1}^{n} \bar{w}_{(T+1)j} = n^{\frac{1}{\kappa}} \cdot \frac{1}{n} \sum_{j=1}^{n} w_{(T+1)j} \leq n^{\frac{1}{\kappa}} \cdot \left( \frac{1}{n} \sum_{j=1}^{n} w_{(T+1)j} \right)^{\frac{1}{\kappa}} = n^{\frac{1}{\kappa}} \cdot n^{-\frac{1}{\kappa}} = 1
\]
(98)
because of the concavity of \( x^{1/\kappa} \), and so we return to the noise-free rado boosting bound with “penalty \( 1/\kappa \)” for renormalizing the leveraging coefficients in RADOBOOST (this proves ineq. \([17]\)). Assuming \( \theta_T \) output by RADOBOOST, we obtain, \( \forall S, \mathcal{U} \) such that support of all \( n \) rados is of the same size, \( i.e. \) \( m_j = m, \forall j \in [n] \),
\[
F_{\log}^r(S, \theta_T, \mathcal{U}) = \log(2) + \frac{1}{m} \log F_{\exp}^r(S, \theta_T, \mathcal{U}) \leq \log(2) - \frac{1}{2\kappa m} \sum_t r_t^2 + \frac{1}{m} \cdot \log \sum_{j=1}^{n} \bar{w}_{(T+1)j} \cdot \exp \left( \frac{1}{2} \cdot \theta_T^\top \sum_i (\sigma_{ji} + y_i) x_i^r \right) \leq \log(2) - \frac{1}{2\kappa m} \sum_t r_t^2 + \frac{m}{m} \cdot \log \sum_{j=1}^{n} \bar{w}_{(T+1)j} \cdot \exp \left( \frac{1}{2m} \cdot \theta_T^\top \sum_i (\sigma_{ji} + y_i) x_i^r \right) = \log(2) - \frac{1}{2\kappa m} \sum_t r_t^2 + \frac{m}{m} \cdot \log \sum_{j=1}^{n} \bar{w}_{(T+1)j} \cdot \exp \left( \frac{\zeta}{\sqrt{m}} \cdot \theta_T^\top \sum_i (\sigma_{ji} + y_i) x_i^r \right) \tag{99}
\]
We now study a sufficient condition for $C - D$ to be $\Omega((1/m) \sum_i r_i^2)$ with high probability over the noise mechanism, thereby ensuring a convergence rate over non-noisy rados that shall comply with the noise-free bounds of ineq. (13), up to the hidden factors. This shall be achieved through several Lemmata.

**Lemma 14** With probability $\geq 1 - \tau$ over the noise mechanism we shall have:

$$\left\| \sum_i \frac{\sigma_{ji} + y_i}{2\sqrt{m_s}} x_i^r \right\|^2 \leq \sqrt{2 \log \left( \frac{n}{\tau} \right)} , \forall j \in [n] .$$ (100)

**Proof** The Sudakov-Tsirelson inequality ([5], Theorem 5.6) states that if $x \sim N(0, I_d)$ and $f(x) : \mathbb{R}^d \to \mathbb{R}$ is $L$-Lipschitz, then

$$\mathbb{P} \left[ f(x) - \mathbb{E}[f(x)] \geq t \right] \leq \exp \left( - \frac{t^2}{2L^2} \right) .$$ (101)

Since function $f(x) = \|x\|_2$ is 1-Lipschitz by the triangle inequality and $\sum_i \frac{\sigma_{ji} + y_i}{2\sqrt{m_s}} x_i^r$ is a standard Gaussian random because the $x_i^r$ are sampled independently, ineq. (101) yields that we shall have simultaneously over the randomized part of the rados, with probability $\geq 1 - \tau$,

$$\left\| \sum_i \frac{\sigma_{ji} + y_i}{2\sqrt{m_s}} x_i^r \right\|^2 \leq \sqrt{2 \log \left( \frac{n}{\tau} \right)} , \forall j \in [n],$$

which proves the Lemma. 

**Lemma 15** Assume $\theta_T \in \mathbb{B}(0, r_0)$ for some $r_0 > 0$. Then with probability $\geq 1 - \tau$ over the noise mechanism we shall have

$$D \leq \frac{s r_0}{m} \sqrt{2m_s \log \left( \frac{n}{\tau} \right)} .$$ (102)

**Proof** We use Lemma 14. Cauchy-Schwartz inequality implies

$$\theta_T^\top \sum_i \frac{\sigma_{ji} + y_i}{2\sqrt{m_s}} x_i^r \leq \left\| \theta_T \right\|_2 \cdot \left\| \sum_i \frac{\sigma_{ji} + y_i}{2\sqrt{m_s}} x_i^r \right\|_2 \leq r_0 \sqrt{2 \log \left( \frac{n}{\tau} \right)} , \forall j \in [n] .$$ (103)

We thus get in this case

$$D \leq \frac{s r_0}{m} \sqrt{2m_s \log \left( \frac{n}{\tau} \right)} + \frac{m_s}{m} \log \sum_{j=1}^{n} \tilde{w}_{(T+1)j}$$

$$\leq \frac{s r_0}{m} \sqrt{2m_s \log \left( \frac{n}{\tau} \right)} .$$ (104)

because of ineq. (98).

We now prove a specific $r_0 > 0$ which makes use of the concentration of the randomized part of rados in Lemma 14.
Lemma 16 Suppose there exists $\mu > \mu' > 0$ such that it simultaneously holds:

\[
\mu \leq \frac{\min_k \max_j |\pi_{jk}|}{m_*},
\]

\[
\mu' \leq \mu - \varsigma \sqrt{\frac{1}{m_*} \log\left(\frac{n}{\tau}\right)},
\]

where $\pi_{jk} = (1/2) \sum_i (\sigma_{ji} + y_i)x_{ik}$ is the non-noisy part of rado $\pi_j^+$. Assume the existence of $\rho > 0$ such that the weak learner $wfi$ in RADOBoost is $\lambda_p$-prudential for

\[
\lambda_p = 1 - \frac{2}{\sqrt{1 - \rho \kappa' m_*}}.
\]

Then probability $\geq 1 - \tau$ over the noise mechanism we shall have

\[
\|\theta_T\|_2 \leq (1 - \rho) \sum_t r_t^2.
\]

Remarks: notice that ineq. (105) is equivalent to saying that each coordinate $k$ has at east one non-zero entry in the noise-free part of the rados. Unless coordinate $k$ is zero for all examples — in which case we can just discard this feature —, this assumption is easy to satisfy.

Proof We have

\[
\|\theta_T\|_2 - \sum_t r_t^2 = \sum_t \frac{1}{4\kappa^2 \pi_{st}(t)} \log^2 \frac{1 + r_t}{1 - r_t} - r_t^2.
\]

Assuming the existence of $z > 0$ such that $2\kappa \pi_{st}(t) \geq z$, $\forall t$, and using the fact that

\[
\log^2 \frac{1 + x}{1 - x} \leq \frac{4x^2}{(1 - |x|)^2}, \forall x \in (0, 1),
\]

we shall have

\[
\sum_t \frac{1}{4\kappa^2 \pi_{st}(t)} \log^2 \frac{1 + r_t}{1 - r_t} - r_t^2 \leq \sum_t \frac{1}{z^2} \log^2 \frac{1 + r_t}{1 - r_t} - r_t^2
\]

\[
\leq \sum_t \frac{1}{z^2} \cdot \frac{4r_t^2}{(1 - |r_t|^2)^2} - r_t^2
\]

\[
= \sum_t r_t^2 \left(\frac{4 - z^2(1 - |r_t|^2)^2}{z^2(1 - |r_t|^2)^2}\right)
\]

\[
\leq -\rho \sum_t r_t^2,
\]

as long as

\[
|r_t| \leq 1 - \frac{2}{\sqrt{1 - \rho z}}, \forall t,
\]

where $\rho \in (0, 1)$. Since $\pi_{st}(t) \geq \min_k \max_j |\pi^+_{jk}|$, we can fix $z_\ast = 2\kappa \min_k \max_k |\pi^+_{jk}|$, but recall that $\pi^+_{jk}$ sums a random Gaussian part and a non random part. Ineq. (100) tells us that with high probability, the magnitude of the random part will satisfy

\[
\sum_i (\sigma_{ji} + y_i)x_i^* \leq \varsigma \sqrt{2m_* \log\left(\frac{n}{\tau}\right)}, \forall j \in [n].
\]
Thus, we shall have in this case, using ineqs. (105, 106) and given Lemma 14:
\[
\min_k \max_{j} |\pi_{jk}^*| \geq \left( \mu - \sqrt{\frac{1}{m} \log \left( \frac{n}{\tau} \right)} \right) \cdot m \ni \geq \mu' m,
\]
and we get the statement of the Lemma.

We now return to ineq. (99), and use Lemmata 14, 15 and 16, and obtain that with probability \( \geq 1 - \tau \), a sufficiently prudential weak learner shall imply:
\[
F^r_{\log}(S, \theta_T, U) \leq \log(2) - \frac{1}{2km} \sum_{t} r_t^2 + \frac{m}{m} \cdot \log \sum_{j=1}^{n} \tilde{\omega}_{(T+1)} \cdot \exp \left( \frac{\varsigma}{\sqrt{m}} \cdot \theta_T \sum_{i} \sigma_{ji} y_i \cdot x_i' \right)
\]
\[
\leq \log(2) - \frac{1}{m} \cdot \left( \frac{1}{2} - (1 - \rho) \cdot \sqrt{2m \log \left( \frac{n}{\tau} \right)} \right) \sum_{t} r_t^2.
\]

We want \( E \geq 1/(4\kappa) \). Equivalently, we want
\[
1 - \rho \leq \frac{1}{4\kappa \varsigma \sqrt{2m \log \left( \frac{n}{\tau} \right)}}
\]
and for the prudential weak learner to exist, we also need
\[
1 - \rho > \frac{4}{\kappa^2 \mu'^2 m^2}.
\]

Assuming ineqs (105) and (106), we thus get that if
\[
\kappa \geq \frac{4\kappa}{\mu'^2 m^2} \sqrt{2 \log \left( \frac{n}{\tau} \right)},
\]
then there exists a prudential weak learner for which, with probability \( \geq 1 - \tau \) over the noise mechanism, we shall have after \( T \) rounds of boosting of RADOBOOST, using the prudential weak learner and renormalizing the leveraging coefficients by \( \kappa \) as in (95),
\[
F^r_{\log}(S, \theta_T, U) \leq \log(2) - \frac{1}{4km} \sum_{t} r_t^2,
\]
which proves Theorem 8. Notice that the constraint \( \kappa \geq 1 \) can easily be enforced by picking \( \mu' \) sufficiently small.

Remarks: we finish by emphasizing the fact that ineq. (19) is computed over non-noisy rados. It is not hard to see that ineqs (105) and (106) shall be all the easier to meet as \( m \) is large compared to \( \log n, \log(1/\tau) \) and \( \varsigma \). So, provided rados have a sufficiently large support, the convergence rate of the logistic rado-risk of RADOBOOST over the non noisy rados may compete, up to a small constant factor, with the one that would be achieved by training RADOBOOST over non-noisy rados.
10.8 Proof of Lemma 10

Consider first that $m \geq 2^d$. A simple proof of the Lemma consists in considering the largest $d$-dim square, of edge length $\ell = 2R/\sqrt{d}$, shown with thick dashed line in Figure 5. We then pack this square with $m+1$ spheres, as shown. Since the edge length is covered by $\lceil \log(m)/\log(d) \rceil$ diameters of these spheres, we obtain that the radius $r$ of each such sphere satisfies:

$$r = \frac{2R}{\sqrt{d} \cdot \left\lceil \frac{\log(m+1)}{\log d} \right\rceil} \geq \frac{R \log d}{2\sqrt{d} \log(m+1)} ,$$

because $m \geq 2^d > d$. Because of the construction, at least one of these spheres does not contain an edge vector from $C(E)$ and is thus empty. Consider one such empty sphere whose center $e_*$ is the closest to 0, as shown in Figure 5, and consider one adjacent sphere, located no farther\footnote{If no such sphere exists, we can pick $e_* = 0$, the center of a sphere $B(0, r)$ which contains no example from $S$. In this case, there is no need to remove any example from $S$: the proof still holds by adding example $(0, y)$ to $S$, to} with one
edge vector \( e = yx \) from \( C(E) \) inside, with \( (x, y) \in S \), where \( S \) generates \( \Pi \). We create \( S' \) out of \( S \) by replacing \( (x, y) \) by two examples, \( (ye_*, y) \) and \( (e - ye_*, y) \). It is worthwhile remarking that

\[
C(E') \subset B(0, R)
\]

by construction, and furthermore any rado that can be created from \( S \) can also be created from \( S' \). Hence, any \( \Pi \) defined over \( S \) can also be obtained from \( S' \). There remains to remark that, by construction, \( e_* \) is distant from every edge vector of \( S \) from at least \( r \), and so:

\[
D(H(E, E')) = \Omega \left( \frac{R \log d}{\sqrt{d \log m}} \right);
\]

this proves Lemma 10 when \( m \geq 2^d \). When \( m < 2^d \), the construction of Figure 5 can still be done but with larger balls, for which

\[
r = \frac{R}{2\sqrt{d}}.
\]

Picking as \( e_* \) the center of any of these empty balls, we obtain

\[
D_H(E, E') \geq \frac{R}{2\sqrt{d}},
\]

as claimed.

10.9 Proof of Lemma 11

We make a reduction from the X3C3 ([25]) problem whose instance is a set \( S = \{s_1, s_2, ..., s_n\} \) and a set of 3-subsets of \( S \), \( C = \{c_1, c_2, ..., c_d\} \), and an integer \( m \). Each element of \( S \) belongs to exactly three subsets of \( C \). The question is whether there exists a cover of \( S \) using at most \( m \) elements from \( C \). The reduction is the following:

- to each feature corresponds an element of \( C \);
- to each element \( s_j \) of \( S \) we associate a boolean rado \( \pi_j \) which is 1 in coordinate \( k \) iff \( s_j \in c_k \), and zero otherwise:

\[
\pi_j = \mathbf{1}_{\{k : s_j \in c_k\}}.
\]

(\( \mathbf{1}_j \) is “1” in coordinate \( i_k \) for \( k \in \mathcal{I} \), and zero everywhere else)

- The number of examples is \( m \);
- Parameters \( r \) and \( \ell \) are fixed as follows:
  - if \( p \neq 0 \), the value of \( r \) is \( 2^{1/p} \). We also fix \( \ell = \epsilon \)-machine, where \( \epsilon \)-machine is the smallest \( \epsilon \) such that \( 1 - \epsilon < 1 \) in machine encoding;
  - else if \( p = 0 \), then \( r = 2 \) and \( \ell = 1 \);
Let us number the constraints of Sparse-Approximation, so that we want:
\[ \|x_i\|_p \leq \ell , \forall i \in [m] , \text{ (Sparse examples)} \] (125)
\[ \|\pi_j - \pi_{\sigma_j}\|_p \leq r , \forall j \in [n] . \text{ (Rado approximation)} \] (126)

Suppose there exists a solution to X3C3 with \( m \) subsets of \( C \), \( C^* = \{c_{k_1}^*, c_{k_2}^*, ..., c_{k_m}^*\} \). Create \( m \) positive examples \((y_i = 1)\) whose observation is \( x_i = 1_{\{k_i\}}\) (the all-0 vector with only one “1” in coordinate \( k_i \)). Clearly, the sparsity constraint on examples (125) is satisfied. We craft the rados following \( n \) Rademacher assignations, where \( \sigma_i \) is +1 only for \( x_{k_i} \), and −1 otherwise. Notice that
\[ \pi_j - \pi_{\sigma_j} = 1_{\{k_s_j \in c_k\}} - 1_{\{k_{s_j} \in c_{k_i}^*\}} = 1_{\{k_{s_j} \in c_k \land c_k \notin C^*\}} . \] (127)
It comes
\[ \|\pi_j - \pi_{\sigma_j}\|_p \leq 2^{1/p} = r , \forall j \in [n] , \] (129)
if \( p \neq 0 \), and
\[ \|\pi_j - \pi_{\sigma_j}\|_0 \leq 2 = r , \forall j \in [n] \] (130)
otherwise, since each element of \( S \) belongs to three sets in \( C \). Therefore, there exists a solution to Sparse-Approximation.

Now, suppose there exists a solution to Sparse-Approximation. Remark that we can remove wlog any example having null observation as this does not change the feasibility of the solution. Consider the case where \( p \neq 0 \). The Rado approximation constraint (126) of Sparse-Approximation makes that the following property (P) is satisfied:

(P) for each \( j \in [n] \), there exists \( i \in [m] \) and feature \( k \in [d] \) such that \( \pi_{\sigma_j} \) and example \( x_i \) have their coordinate \( k \) non-zero, and furthermore the coordinate in \( x_i \) has magnitude exactly \( \epsilon \): it cannot be less otherwise (126) is violated, and it cannot be more otherwise (125) is violated. Hence, each of these \( x_i \) have exactly one non-zero coordinate.

Because property (P) holds for all rados, we see that the corresponding indexes in the \( x_i \) (the corresponding non-zero coordinates for features for which (P) holds; there cannot be more than \( m \)) define a solution to X3C3. The case \( p = 0 \) is easier as (125) enforces the number of non-zero coordinates in each observation to be at most one, and therefore exactly one since there is no null observation.

We finally note that Sparse-Approximation trivially belongs to NP, so it is actually NP-Complete.

10.10 Proof of Lemma 12
We make the same reduction as for Sparse-Approximation. The set of examples \( S \) consists of all canonical basis vectors, associated to positive class.

11 Appendix — Experiments

11.1 Supplementary experiments to Table 1
Table 3 is obtained under the same experimental setting as that of Table 1 with an important modification in how the normalized edge is computed. More specifically, the computation of \( r_t \) in
Table 3: Comparison of RadoBoost to AdaBoost (27) and AdaBoost trained with a random subset of training of the same size as $\delta_s$ (AdaBoost($n$)). The symbol “∗” indicates algorithms are run with the replacement of eq. (131) for the normalized edge $r_t$. Conventions are the same as in Table 1. The symbols $Y$, $N$, —, respectively indicate whether the new version performs better than (resp. worse than, similarly to) the non-modified version.

Step 2.2 of RadoBoost (see (9)) is completed by the following step:

$$r_t \leftarrow \text{sign}(r_t) \cdot \max\{0.1, |r_t|\}$$  \hspace{1cm} (131)

The same modification is also carried out in AdaBoost (27) (Corollary 1). This aims to prevent the fact that domains with outlier feature values could trick AdaBoost in picking the wrong sign for $\alpha_t$ for a large number of iterations, due to values of $r_t$ with a very small magnitude (but with the wrong sign). Experiments display that this corrects AdaBoost’s bad results on Twitter, but on other domains like Fertility, Haberman, Sonar, Abalone, the change happens to give worse results for AdaBoost and/or AdaBoost($n$). RadoBoost’s results, on the other hand, tend to improve with sparse exceptions.

11.2 Supplementary experiments to Section 5 — I / III

Tables 4, 5, 6, 7 present results comparing AdaBoost, RadoBoost with random rados and RadoBoost with fixed support size rados ($m_s$). Unless otherwise stated in Tables, the following experimental setup holds:

- **RadoBoost** is trained with $n = \min\{1000, \text{train fold size}/2\}$ rados;
- **AdaBoost** is trained using the complete training fold;
- for each standard deviation $\sigma$, we generate 10 noisy domains; each is then processed following 10 folds stratified cross-validation. Thus, each dot on the colored curves is the average of ten experiments;
- **RadoBoost** is trained with two types of rados: random rados as in Section 4 — this gives the grey dashed curves —, or rados with fixed support $m_s$ (noted $s$ on the plots) as in Subsection 5.2 — this gives the colored curves —;
11.3 Supplementary experiments to Section 5 — II / III

Tables 8 and 9 compare RADOBOOST trained with rados of fixed support and using a “prudential” weak learner (which picks the median feature according to $|r_t|$), to RADOBOOST trained with plain random rados and using the “strongest” possible weak learner which picks the best feature according to $|r_t|$.

11.4 Supplementary experiments to Section 5 — III / III

Tables 10 and 11 compare two different rado generation mechanisms with respect to RADOBOOST: the random generation of arbitrary rados (Section 4), and the random generation of rados with fixed support (Subsection 5.2). In both Tables, the weak learner is always the same (contrary to Tables 8 and 9), i.e. the “strong” weak learner that picks the best feature according to $|r_t|$, at each iteration.
Table 4: Learning from examples that have been noisified using the Gaussian mechanism $N(0, \sigma^2I)$ (See Section 10.7), as a function of $\sigma$. In each plot, the right axis gives AdaBOOST’s (27) test error, related to the big dotted curve. All other curves are related to the left axis, which gives the difference of test errors ($\Delta perr$) between RADOBOOST and AdaBOOST. The grey dashed curve is for rados picked uniformly at random in $\Sigma_m$, following Section 3. The colored curves (green, red, blue) correspond to rados with fixed support $s (= m_e)$ such that $s/m \in \{0.25, 0.5, 0.75\}$, generated with the mechanism of Section 5.2. $m$ refers to the size of a training fold. Range of $\sigma$ is not the same on the left and right plots. The horizontal dashed black line indicates $\Delta perr = 0$: colored lines below this line indicate runs of RADOBOOST that are better than AdaBOOST’s.
Table 5: Learning from examples that have been noisified using the Gaussian mechanism $N(0, \sigma^2 I)$ (See Section 10.7), as a function of $\sigma$. Conventions follow Table 4.
Table 6: Learning from examples that have been noisified using the Gaussian mechanism \(N(0, \sigma^2I)\) (See Section 10.7), as a function of \(\sigma\). Conventions follow Table 4.
Table 7: Learning from examples that have been noisified using the Gaussian mechanism \(N(0, \sigma^2 I)\) (See Section 10.7), as a function of \(\sigma\). Conventions follow Table 4.
Table 8: Test error of RADOBOOST trained with rados with fixed support and Median-prudential weak learner (Subsection 5.2), minus test error of RADOBOOST trained with random rados and the “Strong” weak learner of Section 4 (i.e. the one that picks the best feature at each iteration), as a function of the Gaussian mechanism’s standard deviation $\sigma$. Horizontal dashed line correspond to $\Delta perr = 0$. Points below this line denote better performances over the rados with fixed support and with the prudential weak learner. $s$ is the support size ($m$ relates to the size of the training fold), for three values, $s/m = 0.25$ (green), $s/m = 0.5$ (red) and $s/m = 0.75$ (blue).
Table 9: Test error of RADOBOOST trained with rados with fixed support and Median-prudential weak learner, minus test error of RADOBOOST trained with random rados and the “Strong” weak learner of Section 4. Conventions follow Table 8.
Table 10: Test error of RADOBOOST trained with rados with fixed support minus test error of RADOBOOST trained with plain random rados, as a function of the Gaussian mechanism’s standard deviation $\sigma$. Points below the $\Delta \text{perr} = 0$ line indicate smaller errors for the training with rados of fixed support. $s$ is the support size ($m$ relates to the size of the training fold), for three values, $s/m = 0.25$ (green), $s/m = 0.5$ (red) and $s/m = 0.75$ (blue).
Table 11: Test error of RADOBOOST trained with rados with fixed support minus test error of RADOBOOST trained with plain random rados (continued). Conventions follow Table 10.