Bridging weak supervision and privacy aware learning via sufficient statistics

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

Frank Nielsen
École Polytechnique
Sony Computer Science Laboratories
Frank.Nielsen@acm.org

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

Abstract

We present a first attempt in connecting two areas of statistical learning that have not shared much common ground: weakly supervised learning and privacy aware learning. In the former, we aim to learn models of labeled data, when full information of the labels is not available; the latter concerns the design of algorithms with privacy guarantees for the protection of the data, while trading off utility for learning.

We focus on classification with linear separators. There exists a sufficient statistic that summarizes all information from the label variable, the mean operator. The fact is known for a broad set of loss functions. Learning algorithms have exploited this property and overcome the lack of label knowledge, learning with label proportions only. We extend the result with almost no structural assumptions on loss functions and regularizers, and show how the approach is potentially viable for any weakly supervised task. Further, we consider the label as the only sensitive variable to protect, while the rest of the data is of public domain. In this scenario, we propose a simple method based on the Laplacian mechanism that obfuscates the mean operator and feed it to a learning algorithm which (a) enjoys α-label differential privacy, (b) is characterized by a generalization bound under almost no structural assumptions and (c) can be integrated into a secure data-sharing protocol for learning. Remarkably, some known results are recovered with simplified proofs.

1 Introduction

This work develops around the concept of *statistical sufficiency*. Informally, a statistic –a function of data– is sufficient when it summarizes the data without information loss for a given task, such that we can fit a model without the need of accessing the original source anymore. The power of statistical sufficiency is therefore highly practical. The first reason is that a sufficient statistic offers a more concise representation of the data. Thus, reducing data to it is advantageous from memory and computational standpoints –although this is not always the case [15]. While in Statistics the notion is built on probability distributions, it comes more natural in Machine Learning to define it on a loss function, a central object of the learning theory. In Section 2 we present a factorization theorem analogous to the classical result on the exponential family due to Fisher and Neymann [14]:

a statistic called *mean operator*[1] is sufficient for the label variable for a broad set of losses utilized in classification. This result provides the fundamental tool for the rest of our discussion.

A second motivation concerns learning scenarios that are peculiar because some of the data is unknown. Following a terminology somehow settled in the field, we refer to those as *weakly supervised learning*[2]. Beyond the classical setting where we learn from a set examples $(\boldsymbol{x}, y)$, we assume that the labels y are in a form that carries poorer information that would be given by their full knowledge. For example, labels may be *noisy* [16, 22], or missing at all for some examples as in *semi-supervised learning* [3], or represented by aggregation as it happens in *multiple instance learning* [7] and *learning from label proportions* (LLP) [19, 18]. As the practical success of many solutions in this settings show, labels themselves are not strictly necessary for learning. Sufficiency provides a simple and principled way to frame those problems and to design their solutions: from the available weak supervision, estimate the mean operator; then, learn a model on it instead of using the (unknown) labels. This direction was explored by work in LLP [19, 18]; We discuss this approach and present an abstract meta-algorithm with known learning guarantees in Section 3.

A third reason stands in the realm of privacy preserving machine learning, an active area of research in the last decade. The objective is two-fold: learning model with quantifiable generalization performance *and* protecting the exposure of data [23]. Differential privacy has gained consensus as a solid theoretical framework for privacy [9]. Loosely speaking, a learning algorithm is said to be differentially private if it produces models that do not vary much when one individual example is added or deleted from the learning sample. This is a powerful concept: every person's attributes contained in a dataset are individually protected. At the same time, the protection is against every attack independently of side information [10]. We present an application on α-label privacy, in which one only wishes to protect the label component of the data. For instance, imagine a learner interested in classifying users of an open-access social network, for which some sensitive label attributes (credit card transactions, criminal records, sexual orientation, etc.) have to be requested to a third party; the data provider may be willing to share the information, upon some guarantees on the individual privacy of the surveyed people or customers. Again, the argument is built sufficient statistics, the only quantities that needs to be sanitized. This new perspective on the question recovers as special cases some known results from [5, 6] and frees them from formal requirements, in particular regularity conditions. The nature of this solution, which acts at the level of (a function of) the data and not on the learning algorithm, allows the implementation of a simple two-agent data-exchange protocol for private learning. See Section 4.

Release of noisy-fied sufficient statistics is not a novel idea itself. Observe for example how perturbed contingency tables [2, 11] and loss (sub)gradients [23] can be seen as similar. Our work is conceptually close to the release mechanism in [23], in the sense that the form of the perturbed data is tied to the learning process we want to solve by the loss function. A related but more ambitious model of protection is *local privacy* [8]. Data needs protection even before learning; the data provider may not trust even the learner. In this scenario the privacy is required on data independently of any future use, while we focus on the release of statistics for learning particular models.

On one hand, advances in learning under differential privacy systematically offers generalization guarantees for algorithms, in order to certificate their utility beyond privacy. On the other hand, researchers in weakly supervised learning often underline the potential implications of proposed techniques when applied to infer sensitive data. However, we are not aware of formal connections between the two areas. One of our objective is to show how both sides of the question can be framed and studied utilizing the same language.

2 Learning setting, mean operator and sufficiency

We denote vectors with boldfaces and the sequence $i = 1, \dots, m$ as $i \in [m]$. The notation $1\{p\}$ is the indicator function of a predicate p being true. Let be $\mathcal{X} \subseteq \mathbb{R}^d$ the feature space and $\mathcal{Y} \doteq \{-1, 1\}$

[1]The name *mean operator*, sometimes referred to as *mean map*, is borrowed from the theory of Hilbert space embedding in kernel methods. See for example [21, 19].

[2]We deliberately do not consider the situation of information missing on the features vectors, as in the case of missing feature values, for instance.

the label space. The learning sample of a binary classification problem is $\mathcal{S} = \{(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i \in [m]\}$. We classify instances $x \in \mathcal{X}$ with linear separators as $\text{sgn}\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle$.

A loss is any function $f : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ defined as $f(y\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)$. A loss in this form is named binary margin loss [20] and is intrinsically symmetric, *i.e.* it admits identical class misclassification costs. We interpret such function as giving a penalty when predicting the value $\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle$ and an observed label is y. Sometimes we will use a generic scalar argument $f(x)$ for lighter notation. We aim to learn models $\boldsymbol{\theta}$ that make the smallest number of mistakes as measured by the *0-1 loss* $f(x) = 1\{x < 0\}$. Due to the difficulty on optimizing the 0-1 loss directly, learning algorithms traditionally resort to surrogates losses, *e.g. logistic* $f(x) = \log(1+\exp(-x))$ and *square* $f(x) = (1-x)^2$; they all upper bound the true objective. We define the empirical (surrogate) risk as $L(\mathcal{S}, \boldsymbol{\theta}) \doteq \mathbb{E}_{\mathcal{S}}[f(y\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle)] = \frac{1}{m} \sum_i f(y_i \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle)$. The regularized version is denoted as $L_\lambda(\mathcal{S}, \boldsymbol{\theta}) \doteq L(\mathcal{S}, \boldsymbol{\theta}) + \lambda \Omega(\theta)$, where $\Omega : \mathbb{R} \to \mathbb{R}^+$ is a non-decreasing function of $\boldsymbol{\theta}$. Next we introduce the mean operator, a statistic that will naturally let us draw a link between weakly-supervised learning and differential privacy.

Definition 1 *The (empirical)* mean operator *of a learning sample* $\mathcal{S}$ *is* $\boldsymbol{\mu}(\mathcal{S}) \doteq \frac{1}{m} \sum_{i=1}^m y_i \boldsymbol{x}_i$.

We will drop dependence on $\mathcal{S}$ when clear by context. This quantity has been studied in [1, 19, 18] and is the pillar of our argument; the property that is relevant to us is its statistical sufficiency, that we now define.

Definition 2 *A statistic* $t(\mathcal{S})$ *computed on the learning sample* $\mathcal{S}$ *is said to be sufficient for a random variable* z *iff: for any* f*, any* $\boldsymbol{\theta}$ *and any two samples* $\mathcal{S}$ *and* $\mathcal{S}'$*, the quantity* $L(\mathcal{S}, \boldsymbol{\theta}) - L(\mathcal{S}', \boldsymbol{\theta})$ *does not depend on* z *iff* $t(\mathcal{S}) = t(\mathcal{S}')$.

The definition is given in analogy with Statistics. To see that, we recall the case of fitting a binary conditional exponential family with maximum likelihood. Its probability distribution is parametrized by $\boldsymbol{\theta}$ as $\exp(\langle \boldsymbol{\theta}, y\boldsymbol{x} \rangle - \log \sum_y \exp(\langle \boldsymbol{\theta}, y\boldsymbol{x} \rangle))$, where $y \in \{-1, 1\}, \boldsymbol{x} \in \mathbb{R}^d$. $y_i \boldsymbol{x}_i$ is a function of the data that fully summarizes one sample, and indeed is called *sufficient statistic*. The sufficient statistic of a set of independent random variables drawn is simply the sum of individual sufficient statistics, $\sum_i y_i \boldsymbol{x}_i$, *i.e* the mean operator. In fact, under the *i.i.d.* assumption, the log-likelihood of $\boldsymbol{\theta}$ is (the negative of)

$$\sum_{i=1}^m \log \sum_{y \in \mathcal{Y}} \exp(y\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle) - \sum_{i=1}^m \langle \boldsymbol{\theta}, y_i \boldsymbol{x}_i \rangle \tag{1}$$

$$= \sum_{i=1}^m \log \sum_{y \in \mathcal{Y}} \exp\left(y\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle\right) - \sum_{i=1}^m \log \exp\left(y_i \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle\right)$$

$$= \sum_{i=1}^m \log \left(\frac{\exp(\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle) + \exp(-\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle)}{\exp(y_i \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle)} \right)$$

$$= \sum_{i=1}^m \log\left(1 + \exp(-2y_i \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle)\right) \tag{2}$$

$$\propto \frac{1}{m} \sum_{i=1}^m \log\left(1 + \exp(-y_i \langle \boldsymbol{\theta}', \boldsymbol{x}_i \rangle)\right) \tag{3}$$

We obtain Step (2) by observing that the label always takes value in $\{-1, +1\}$. In the last step (3), we operate a change of variable by expressing the new model $\boldsymbol{\theta}' \leftarrow 2\boldsymbol{\theta}$, and normalizing the cost by $1/m$. Thus, maximizing this log-likelihood is equivalent to minimizing the empirical risk built on standard logistic loss. Equation (1) is the first instance of a fundamental property that we investigate in this work. The logistic loss factorizes in two components: a linear part that corresponds to $\langle \boldsymbol{\theta}, \sum_i y_i \boldsymbol{x}_i \rangle$, and a non-linear part that maps back to the cumulant of the exponential family. The key fact is that the non-linear part is independent from y: we only need the label to compute the component which involves the mean operator. In other words, knowing the mean operator (or an estimate of it) is sufficient for evaluating and minimizing the loss. The following theorem states the property in its generality.

name	loss	even term	odd term
generic (linear-odd)	$f(x)$	$\frac{f(x)+f(-x)}{2}$	$-\frac{a}{2}x$
-	$\rho\lvert x\rvert - \rho x + 1$	$\rho\lvert x\rvert + 1$	$-\rho x$
linear	$-x$	0	$-x$
SPL [17]	$a_\phi + \frac{\phi^*(-x)}{b_\phi}$	$a_\phi + \frac{\phi^*(x)+\phi^*(-x)}{2b_\phi}$	$-\frac{x}{2b_\phi}$
logistic	$\log(1+e^{-x})$	$\frac{1}{2}\log(2+e^x+e^{-x})$	$-\frac{x}{2}$
square	$(1-x)^2$	$1+x^2$	$-2x$
Matsushita	$\sqrt{1+x^2}-x$	$\sqrt{1+x^2}$	$-x$

Table 1: Factorization of losses. Permissible convex surrogate enjoy the property that $\phi^*(-x) = \phi^*(x) - x$, and include logistic, square and Matsushita losses.

Theorem 3 (Linear-odd loss factorization) *Let $f(x)$ be a loss function such that $f(x) - f(-x) = -ax$ with $a \in \mathbb{R}$. For any learning sample $\mathcal{S}$, the empirical risk $L(\mathcal{S}, \boldsymbol{\theta})$ computed on f can be written as*

$$L(\mathcal{S}, \boldsymbol{\theta}) = \underbrace{\frac{1}{2m}\sum_i \sum_{u\in\{-1,1\}} f(u\langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle)}_{\textit{same } f\textit{, label independent}} - \frac{a}{2}\langle \boldsymbol{\theta}, \boldsymbol{\mu}\rangle \tag{4}$$

Furthermore, the factorization is unique.

Proof The proof relies on the (unique) odd/even factorization of any arbitrary function with the further assumption that the odd part is linear:

$$\begin{aligned} f(x) &= \frac{1}{2}\big(f(x) + f(-x) + f(x) - f(-x)\big) \\ &= \frac{1}{2}\sum_{u\in\{-1,1\}} f(ux) - \frac{a}{2}x \ . \end{aligned}$$

The statement follows by computing the empirical risk on $f(y\langle \boldsymbol{\theta}, \boldsymbol{x}\rangle)$. This yields the mean operator by linearity of the model. ■

From now on we can denote losses by $L(\boldsymbol{\theta}, \boldsymbol{\mu})$ –with implicit dependence on the learning sample $\mathcal{S}$– and their regularized version accordingly. Label-independent and -dependent terms in the risk are respectively even and odd. By construction, the even function takes the same functional form of the original loss f. Moreover, we did not base the argument on regularity conditions on the loss function; thus the result holds regardless of smoothness and differentiability and even convexity and properness [20]. A consequence is the next Corollary.

Corollary 4 *Under the hypotheses of Theorem 3, $\boldsymbol{\mu}(\mathcal{S})$ is a sufficient statistic for the y variable.*

The *linear-odd* condition $f(x) - f(-x) = -ax$ may seem very strong at first sight. However, it turns out that some commonly used losses satisfy the condition and that Theorem 3 strictly generalizes the decomposition of Lemma 1 of [18]. The known result is valid for *symmetric proper losses* (SPL), the set of bounded proper losses that are twice differentiable and symmetric [17]; they include logistic, square and Matsushita losses. Theorem 3 is less restrictive and has an extremely simple proof based on the unique odd/even decomposition of real functions.

Table 1 provides some instances of the factorization, where the linear-odd condition is straightforward to check. They include the *linear loss*, proven to be robust against label noisy only recently [22]. We consider here some interesting examples. For logistic loss it holds that

$$\frac{1}{2}(f(x) - f(-x)) = \frac{1}{2}\log\frac{1+e^{-x}}{1+e^{x}} = \frac{1}{2}\log\frac{e^{-\frac{x}{2}}(e^{\frac{x}{2}} + e^{-\frac{x}{2}})}{e^{\frac{x}{2}}(e^{-\frac{x}{2}} + e^{\frac{x}{2}})} = -\frac{x}{2}$$

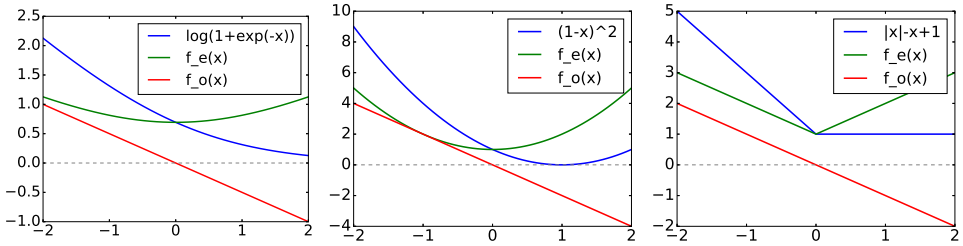

Figure 1: Examples of loss factorization: even and odd parts sum up to make the loss.

and hence the whole function factorizes in $\frac{1}{2}\sum_{u\in\mathcal{Y}}\log(1+\exp(-ux))-\frac{x}{2}$. This "symmetrization" was known in the literature [12]. For the case of SPL, given a permissible generator ϕ [13, 17], *i.e.* $\mathrm{dom}(\phi)\supseteq[0,1]$, ϕ is strictly convex, differentiable and symmetric with respect to $1/2$, any loss can be written as

$$f_\phi(x) = a_\phi + \frac{\phi^*(-x)}{b_\phi}$$

where $\phi^\star$ is the convex conjugate of ϕ. Since $\phi^*(-x)=\phi^*(x)-x$, we have

$$\frac{1}{2}(f_\phi(x)-f_\phi(-x)) = \frac{1}{2}\Big(a_\phi+\frac{\phi^*(-x)}{b_\phi}-a_\phi-\frac{\phi^*(x)}{b_\phi}\Big) = \frac{\phi^*(-x)-\phi^*(x)}{2b_\phi} = -\frac{x}{2b_\phi}\ .$$

A natural question is whether the class SPL is equivalent to the set of losses for which Theorem 3 applies. We answer by negative since there is at least a family of surrogate losses, in particular non differentiable, that yields the mean operator. Let $f(x)=\rho|x|-\rho x+1$, with $\rho>0$. $f(x)$ upper bounds the 0-1 loss and intercept it in $f(0)=1$. We have the following

$$\frac{1}{2}(f(x)-f(-x)) = \frac{1}{2}((\rho|x|-\rho x+1)-(\rho|x|+\rho x+1)) = -\rho x\ .$$

However, other non-differentiable functions, such as the *hinge loss* $f(x)=\max(0,1-x)$, do not satisfy the linear-odd requirement. A full characterization of the class of losses of interest, although somehow not constructive, comes by mapping it to a proper subset of even functions.

Lemma 5 *The exhaustive class of loss functions that satisfies the linear-odd condition is in 1-to-1 mapping with a proper subclass of even functions.*

Proof Consider the class of functions satisfying $f(x)-f(-x)=-ax$, with $a\in\mathbb{R}^+$. Define another function $g(x)=f(x)+\frac{a}{2}x$, which is even. In fact we have $g(-x)=f(-x)-\frac{a}{2}x=f(x)+ax-\frac{a}{2}x=f(x)+\frac{a}{2}x=g(x)$. ■

3 Weakly supervised learning

Armed with the factorization theorem, we can design Meta-Algorithm 1, a generic approach to tackle weakly supervised classification problems. Every setting is characterized by a peculiar kind of supervision granted to the learner and a set of domain-specific assumptions. We represent the weak supervision W as a function of all labels (and potentially the feature vectors as well) in the "original" unknown learning sample, which in practice may have never existed: $W:\mathcal{S}\to\mathcal{W}$, with space $\mathcal{W}$ defined by the learning setting. Moreover, a set of hypotheses $\mathcal{H}$ is usually necessary for the formulation of a solution to those problems, which are ill-posed by definition. Hypotheses may concern the process the supervision itself, or the whole data distribution more generally. For examples, three *a priori* criteria have led the development of the majority of semi-supervised algorithms: smoothness, clustering and manifold assumptions [3]. From $(W(\mathcal{S}),\mathcal{H})$ and the unlabeled $\boldsymbol{x}$s, all we need is to come up with an estimator g of the mean operator. A first instantiation of Meta-Algorithm 1 is the *Mean Map* algorithm for LLP in [19]. In this setting, supervision is only given by label

Algorithm 1 Weakly supervised classification via sufficiency

Input $x_i, \forall i \in [m]$; $W(\mathcal{S})$; $\mathcal{H}$; $\lambda > 0$
Estimate the mean operator: $\tilde{\boldsymbol{\mu}} \leftarrow g(\{x_i, \forall i \in [m]\}, W(\mathcal{S}), \mathcal{H})$
Solve

$$\boldsymbol{\theta}^* \leftarrow \arg\min_{\boldsymbol{\theta}} \frac{1}{2m} \sum_{i,u \in \mathcal{Y}} f(u\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle) - \frac{a}{2} \langle \boldsymbol{\theta}, \tilde{\boldsymbol{\mu}} \rangle + \lambda \Omega(\boldsymbol{\theta})$$

Output $\boldsymbol{\theta}^*$

proportions over set of examples called *bags*. The algorithm fits data to the conditional exponential family and assumes *homogeneity*, namely that $\mathbb{E}[x|y,j] = \mathbb{E}[x|j]$ for each of the $j \in [n]$ bags. [18] relaxes the assumption calling instead for a manifold hypothesis via Laplacian regularization for the estimation of the mean operator.

Learning guarantees for Meta-Algorithm 1 are known. We mention here a result due to [1] that applies to $\boldsymbol{\theta}$'s discrepancy. The theorem requires the loss function to be convex and differentiable.

Theorem 6 *[1] Let f be convex and differentiable and let $L_\lambda(\boldsymbol{\mu}, \boldsymbol{\theta}) = L(\boldsymbol{\mu}, \boldsymbol{\theta}) + \lambda \Omega(\boldsymbol{\theta})^2$, with $\lambda > 0$. Let $\boldsymbol{\theta}^*$ and $\tilde{\boldsymbol{\theta}}^*$ be the minimizers of $L_\lambda(\boldsymbol{\mu}, \cdot)$ and $L_\lambda(\tilde{\boldsymbol{\mu}}, \cdot)$ respectively. Then it holds that $\|\boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}}^*\| \leq \frac{1}{\lambda} \|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|$.*

Those bounds directly relate the quality of the estimator with the model that can be obtained from it. An even tighter bound is provided for losses in SPL class in [18]. Next, we see how the very same algorithm can be used to obtain α-label privacy.

4 Differential privacy

We recall here the framework of α-label differential privacy. This is a weaker notion than the more traditional one, since we are only interested in preserving the privacy of the labels in the data. α-label differential privacy was studied systematically in [4].

Definition 7 *An algorithm A is α-label private if, for any two learning samples $\mathcal{S}, \mathcal{S}'$ differing in at most one label and any output θ of A*

$$\mathbb{P}[A(S) = \theta] \leq e^\alpha \cdot \mathbb{P}[A(S') = \theta] \ .$$

Guarantees on differential privacy are usually tied to the sensitivity of the algorithm to be sanitized.

Definition 8 *For any two learning samples $\mathcal{S}, \mathcal{S}'$ differing in at most one label y', the sensitivity of an algorithm A is*

$$\Delta(A) \doteq \max_{\mathcal{S}, \mathcal{S}'} \|A(\mathcal{S}) - A(\mathcal{S}')\|_1 \ .$$

A straightforward way for enforcing (label) differential privacy is the so-called Laplace mechanism. It can be shown that adding Laplacian noise to each component of the output of the algorithm is sufficient for obtaining privacy [9]. However, the obfuscation does not need to apply at the very end of the computation; arguably there are three approaches for assuring differential privacy of an algorithm, namely at the level of input, output or at any internal state of the algorithm. In the framework of *empirical risk minimization*, one can perturbate the final model, *i.e. output obfuscation*, or the objective function of the learning algorithm, *objective perturbation* [5, 6]. In contrast, we advocate for *input perturbation*, halfway between the former techniques and *local privacy* [8].

An α-label private Algorithm is 2. It simply amounts to obfuscate the each component of the mean operator with unbiased Laplacian noise. This is not a learning algorithm itself but it can be used to make private downstream algorithms. From another point of view we are sanitizing the labels, exposing only a noisy summary statistic. As a consequence, any algorithm using this approximated view of the data will be private too. We prove the privacy of Algorithm 2.

Algorithm 2 Obfuscated release of the mean operator

Input $\mathcal{S}$, α
Let $\boldsymbol{b} \leftarrow (b_1, b_2, \ldots, b_d)$, where $b_i \sim \mathrm{Lap}(0, \frac{2}{m\alpha})$ for each $i \in [d]$
Output $\tilde{\boldsymbol{\mu}} = \frac{1}{m}\sum_i y_i \boldsymbol{x}_i + \boldsymbol{b}$

Theorem 9 *Assume that $\|\boldsymbol{x}_i\| \leq 1$ for all i. Algorithm 2 is α-label private.*

Proof The sensitivity $\Delta(\tilde{\boldsymbol{\mu}})$ is computed on two learning samples that only differ by flipping one example's label:

$$\begin{aligned}
\|\tilde{\boldsymbol{\mu}}(\mathcal{S}) - \tilde{\boldsymbol{\mu}}(\mathcal{S}')\| &= \left\| \frac{1}{m} \sum_{(\boldsymbol{x},y)\in\mathcal{S}} y\boldsymbol{x} - \frac{1}{m} \sum_{(\boldsymbol{x},y)\in\mathcal{S}'} y\boldsymbol{x} \right\| \\
&= \left\| \frac{1}{m} y\boldsymbol{x}' - \frac{1}{m} y'\boldsymbol{x}' \right\| \\
&= \frac{1}{m} |y - y'| \, \|\boldsymbol{x}\| \\
&\leq \frac{2}{m}
\end{aligned}$$

where $(\boldsymbol{x}', y')$ is the example which label has been flipped. The perturbed mean operator is differentially private since we are applying the Laplace mechanism with variance $\frac{2}{m\alpha} \geq \frac{\Delta(\tilde{\boldsymbol{\mu}})}{\alpha}$. ■

In light of Theorem 3, the knowledge of the mean operator, along with the unlabeled instances, is sufficient for minimizing the loss function. Meta-Algorithm 1 can be called for learning: here we do not estimate the mean operator by means of weak supervision but we use the noisy release as the best estimator we may hope to get; notice that this estimator is always unbiased.

Our result is close to Theorem 9 in [6], as we obtain an α independent of regularization. In fact, by combining Algorithms 2 and 1 we obtain *exactly* the same algorithm as *objective perturbation*, due to the linear interaction of model and mean operator. Although, there are differences. First, the unlabeled data is common knowledge and therefore our privacy guarantee only applies to the labels. Second, we do not assume any functional property of losses or regularizers considered, except that they yield the mean operator; in particular, we are not restricted by differentiability or (strong) convexity, thanks to Theorem 3. Third, despite we require less assumptions, our proof is considerably simpler. A discussion on privacy related to performing the cross validation of λ is left aside; see for example [6] that considers ways to approach this issue.

4.1 Learning performance

In this Section, we evaluate the impact of a perturbed mean operator on learning performance. The literature has investigated the problem [1, 19, 18]. We present here a generalization result tied to Algorithm 1: once we consider the mean operator to be affected by Laplacian noise, we can prove a generalization bound characterized by a linear rate of convergence *w.r.t.* the number of examples. In the rest of the Section, every norm $\|\cdot\|$ is a 2-norm.

Theorem 10 *Assume that $\|\boldsymbol{x}_i\| \leq 1$ for all $i \in [m]$ and $\|\boldsymbol{\theta}\| \leq R$. Let $L_\lambda(\boldsymbol{\mu}, \boldsymbol{\theta}) = L(\boldsymbol{\mu}, \boldsymbol{\theta}) + \lambda\Omega(\boldsymbol{\theta})$ with $\lambda \geq 0$. Let $\tilde{\boldsymbol{\mu}}$ be the output of Algorithm 2. Then for any $\boldsymbol{\theta}$, with probability $\delta > 0$*

$$\mathbb{P}\left[|L_\lambda(\boldsymbol{\mu}, \boldsymbol{\theta}) - L_\lambda(\tilde{\boldsymbol{\mu}}, \boldsymbol{\theta})| \leq \frac{|a| R \log(\frac{d}{\delta})}{m\alpha} \right] \geq 1 - \delta \ .$$

Proof Let us begin by upper bounding

$$\begin{aligned}
|L_\lambda(\boldsymbol{\mu}, \boldsymbol{\theta}) - L_\lambda(\tilde{\boldsymbol{\mu}}, \boldsymbol{\theta})| &= |a|/2 \cdot |\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - \langle \boldsymbol{\theta}, \tilde{\boldsymbol{\mu}} \rangle| \qquad (5) \\
&= |a|/2 \cdot |\langle \boldsymbol{\theta}, \boldsymbol{\mu} - \tilde{\boldsymbol{\mu}} \rangle| \\
&= |a|/2 \cdot |\langle \boldsymbol{\theta}, \boldsymbol{b} \rangle| \\
&\leq |a|/2 \cdot R\|\boldsymbol{b}\| \qquad (6)
\end{aligned}$$

Algorithm 3 Two-agent private-learning protocol

Agents: labels provider D knows labels $y_i, \forall i \in [m]$; learner L
Common knowledge: $\boldsymbol{x}_i, \forall i \in [m]$
D and L: establish the privacy level α
D: compute $\tilde{\boldsymbol{\mu}}$ with Algorithm 2
D: send $\tilde{\boldsymbol{\mu}}$ to L
L: learn model $\boldsymbol{\theta}^*$ with Algorithm 1

Step 5 follows from Theorem 3. Step 6 is due to the Cauchy-Schwartz inequality and by hypothesis of bounded models. $\boldsymbol{b}$ is the random vector sampled by Algorithm 2. To bound its norm, we use the fact that for Laplacian random variable $l \sim \mathrm{Lap}(0, \sigma)$ holds that $\mathbb{P}(|l| > \epsilon) = \exp(-\epsilon/\sigma)$ [10]. Then applying an union bound

$$\begin{aligned}\mathbb{P}\Big[\|\boldsymbol{b}\| > \epsilon\Big] &\leq \mathbb{P}\Big[\forall i \in [d],\ \ |b_i| > \epsilon\Big] \\ &= d\,\mathbb{P}\Big[|b_i| > \epsilon\Big] \\ &= d \exp\Big(-\frac{\epsilon}{\sigma}\Big)\ .\end{aligned}$$

It follows that with probability at least $1 - \delta$, $\|\boldsymbol{b}\| \leq \sigma \log \frac{d}{\delta}$. Combining this with Equation 6 and setting $\sigma = 2/(m\alpha)$ concludes the proof. ■

The shape of the bound is similar to the one of *objective perturbation* in Lemma 19 of [6], while it is derived by a simpler proof. As already noticed above, this result seems to be independent of requirements on the structural form of the losses consider. However, it is important to remark the presence of a which parallels the role of usual Lipschitz or strong-convexity factors in similar bounds in statistical learning.

4.2 A two-agent scenario

A notable difference between our approach and *objective perturbation* becomes evident in a multi-agent learning setting. *Input perturbation* has the strong advantage that data can be shared and hence there is no need to learn *in loco*, similarly to *local privacy*. Therefore, we can design a simple two-agent learning protocol, based on Algorithms 1-2. A learner L aims to fit data to a linear classifier but does not have the necessary labels; a data provider D has that knowledge. Her objective might be financial gain, under the constraint of privacy of the individual labels she only knows. The two agents negotiate the α parameter: effectively, the number trades off privacy required by D and utility desired by L, in the usual game-theoretic fashion [8, 23]. Algorithm 2 is executed by D, who sends the result to L, which in turn runs Algorithm 1. All the protocol is secure because labels are accessed only through the private mechanism of Algorithm 2. Algorithm 3 summarizes the exchange.

5 Conclusion

We have highlighted how two independent streams of research can be reconnected leveraging the simple but powerful concept of statistical sufficiency for losses, in the special case of binary classification with linear classifiers. The factorization Theorem 3 presented here is a novel result. Its linear-odd requirement needs deeper investigation to clarify to which extent factorization might help to simplify other known results from regularity conditions, beyond privacy-preserving methods.

Acknowledgments

The authors are pleased to acknowledge the contribution of Aaron Defazio and Tiberio Caetano on the seminal ideas that originated this work. 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] Y. Altun and A. J. Smola. Unifying divergence minimization and statistical inference via convex duality. In *19 th COLT*, pages 139–153, 2006.

[2] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *PODS*. ACM, 2007.

[3] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-supervised learning*. MIT press Cambridge, 2006.

[4] K. Chaudhuri and D. Hsu. Sample complexity bounds for differentially private learning. In *JMLR*, volume 2011, page 155, 2011.

[5] K. Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In *NIPS*09*, 2009.

[6] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *JMRL*, 12:1069–1109, 2011.

[7] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89:31–71, 1997.

[8] J. Duchi, M. J. Wainwright, and M. I. Jordan. Local privacy and minimax bounds: Sharp rates for probability estimation. In *NIPS*26*, 2013.

[9] C. Dwork. Differential privacy. In *Encyclopedia of Cryptography and Security*, pages 338–340. Springer, 2011.

[10] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2013.

[11] M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. In *NIPS*12*, 2012.

[12] T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

[13] M. J. Kearns and Y. Mansour. On the boosting ability of top-down decision tree learning algorithms. In *28 th ACM STOC*, pages 459–468, 1996.

[14] E. L. Lehmann and G. Casella. *Theory of point estimation*, volume 31. Springer Science & Business Media, 1998.

[15] A. Montanari. Computational implications of reducing data to sufficient statistics. *arXiv preprint arXiv:1409.3821*, 2014.

[16] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari. Learning with noisy labels. In *NIPS*13*, 2013.

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

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

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

[20] M. D. Reid and R. C. Williamson. Composite binary losses. *JMLR*, 11:2387–2422, 2010.

[21] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[22] B. van Rooyen, A. K. Menon, and R. C. Williamson. Learning with symmetric label noise: The importance of being unhinged. In *NIPS*15*, 2015.

[23] M J Wainwright, M I Jordan, and J C Duchi. Privacy aware learning. In *NIPS*12*, 2012.