Loss Factorization, Weakly Supervised Learning and Label Noise Robustness

Giorgio Patrini[1,2] GIORGIO.PATRINI@ANU.EDU.AU
Frank Nielsen[3,4] NIELSEN@LIX.POLYTECHNIQUE.FR
Richard Nock[2,1] RICHARD.NOCK@DATA61.CSIRO.AU
Marcello Carioni[5] MARCELLO.CARIONI@MIS.MPG.DE
Australian National University[1], Data61[2], École Polytechnique[3],
Sony Computer Science Laboratories Inc[4], Max Planck Institute for Mathematics in the Sciences[5]

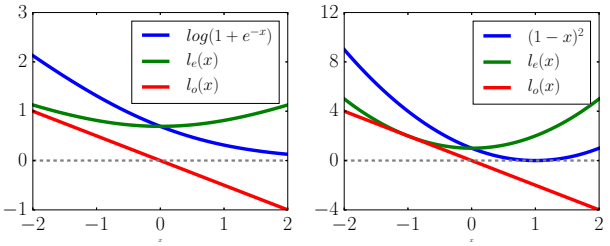
Abstract

We prove that the empirical risk of most well-known loss functions factors into a linear term aggregating all labels with a term that is label free, and can further be expressed by sums of the same loss. This holds true even for non-smooth, non-convex losses and in any RKHS. The first term is a (kernel) mean operator — the focal quantity of this work — which we characterize as the sufficient statistic for the labels. The result tightens known generalization bounds and sheds new light on their interpretation.

Factorization has a direct application on weakly supervised learning. In particular, we demonstrate that algorithms like SGD and proximal methods can be adapted with minimal effort to handle weak supervision, once the mean operator has been estimated. We apply this idea to learning with asymmetric noisy labels, connecting and extending prior work. Furthermore, we show that most losses enjoy a data-dependent (by the mean operator) form of noise robustness, in contrast with known negative results.

1. Introduction

Supervised learning is by far the most effective application of the machine learning paradigm. However, there is a growing need of decoupling the success of the field from its topmost framework, often unrealistic in practice. In fact while the amount of available data grows continuously, its relative training labels — often derived by human effort — become rare, and hence learning is performed with partially missing, aggregate-level and/or noisy labels. For this reason, *weakly supervised learning* (WSL) has attracted much research. In this work, we focus on binary classification under weak supervision. Traditionally, WSL problems are attacked by designing *ad-hoc* loss functions and optimization algorithms tied to the particular learning setting. Instead, we advocate to "do not reinvent the wheel" and present an unifying treatment. In summary, we show that, under a mild decomposability assumption,



(a) logistic loss (b) square loss

Figure 1. Loss factorization: $\ell(x) = \ell_e(x) + \ell_o(x)$.

Any loss admitting a minimizing algorithm over fully labelled data, can also be minimized in WSL *setting with provable generalization and noise robustness guarantees. Our proof is constructive: we show that a simple change in the input and of one line of code is sufficient.*

We introduce *linear-odd losses* (LOLs), a definition not demanding smoothness or convexity and embracing many losses of practical interest, *e.g.* logistic and square. They decompose into an even and an odd function (Figure 1) as

$$\ell(x) = \ell_e(x) + \ell_o(x) \ ,$$

where $\ell_e(x) \doteq (\ell(x) + \ell(-x))/2$ is label-independent. We term this result *Factorization*. When we consider linear or kernel models, the ℓ-*risk* factors in a label free term with another incorporating a *sufficient statistic of the labels*, the mean operator $\boldsymbol{\mu} \doteq \mathbb{E}[y\boldsymbol{x}]$. The interplay of the two components is apparent in a new generalization bound, that also improves Kakade et al. (2009). The result is reminiscent of Fisher-Neyman's factorization (Lehmann & Casella, 1998) of the exponential family that can be seen as a special case.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

Isolating labels is advantageous in applications on WSL, where training labels are only partially observable due to a noise process (Garcıa-Garcıa & Williamson, 2011; Hernandez-Gonzalez et al., 2016). For example, labels may missing as with *semi-supervision* (Chapelle et al., 2006) and *positive and unlabelled data* (du Plessis et al., 2015), *noisy* (Natarajan et al., 2013), or aggregated as it happens in *multiple instance learning* (Dietterich et al., 1997) and *learning from label proportions* (LLP) (Quadrianto et al., 2009). As the success of those areas shows, labels are not strictly needed for learning. However, most WSL methods implicitly assumes that labels must be recovered in training, as pointed out by Joulin & Bach (2012). Instead, sufficiency supports a more principled two-step approach: (1) estimate the mean operator $\boldsymbol{\mu}$ from weakly supervised data and (2) plug it into any LOL and resort to known procedures for empirical risk minimization (ERM). Thus, (1) becomes the only technical obstacle in adapting an algorithm, although often easy to surpass. Indeed, this 2-step approach unifies a growing body of literature (Quadrianto et al., 2009; Patrini et al., 2014; van Rooyen et al., 2015; Gao et al., 2016). As a showcase, we implement (2) by adapting stochastic gradient descent (SGD) to WSL. We only require to transform the input to a "double sample" $\mathcal{S}_{2x}$ — each example is given twice, labelled once with each label — and to sum $\boldsymbol{\mu}$ in the model update. For each example $(\boldsymbol{x}_i, y_i)$ in $\mathcal{S}_{2x}$, a constant depending on the choice of ℓ and η the learning rate, we have:

$$\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t - \eta \nabla \ell(y_i \langle \boldsymbol{\theta}^t, \boldsymbol{x}_i \rangle) - \frac{1}{2} \eta a \boldsymbol{\mu} \ .$$

We then focus on learning with asymmetric label noise, with noise rates (p_+, p_-). We extend the work of Natarajan et al. (2013) by designing an unbiased estimator of $\boldsymbol{\mu}$:

$$\hat{\boldsymbol{\mu}} \doteq \mathbb{E}_{(\boldsymbol{x},y)} \left[\frac{y - (p_- - p_+)}{1 - p_- - p_+} \boldsymbol{x} \right] \ ,$$

on which we derive a generalization bound not tied to neither loss or algorithm. Long & Servedio (2010) has shown that the strongest form of robustness — on any possible noisy sample — rules out most losses commonly used, and have drifted research focus on non-convex (Stempfel & Ralaivola, 2009; Masnadi-Shirazi et al., 2010; Ding & Vishwanathan, 2010) or linear losses (van Rooyen et al., 2015). More pragmatically, we show that *any* LOL enjoys an *approximate* form of noise robustness. The mean operator is still central here, being the data-dependent quantity that shapes the bound. The theory is validated by experiments in which we call the adapted SGD as a black box.

Next, Section 2 settles notations and background. Section 3 states the Factorization Theorem. Sections 4 and 5 focus on WSL and noisy labels. Section 6 discusses the paper. Most proofs and additional results appear in the supplementary material (SM).

2. Preliminaries

2.1. Learning setting

We denote vectors in bold as $\boldsymbol{x}$ and $1\{p\}$ the indicator of p being TRUE. We define $[m] \doteq \{1, \ldots, m\}$ and $[x]_+ \doteq \max(0, x)$. In binary classification, a learning sample $\mathcal{S} = \{(\boldsymbol{x}_i, y_i), i \in [m]\}$ is a sequence of (observation, label) pairs, the examples, drawn from an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, with $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$. Expectation (or average) over $(\boldsymbol{x}, y) \sim \mathcal{D}$ ($\mathcal{S}$) is denoted as $\mathbb{E}_{\mathcal{D}}$ ($\mathbb{E}_{\mathcal{S}}$).

Given a hypothesis (or model) $h \in \mathcal{H}, h : \mathcal{X} \to \mathbb{R}$, a loss is a function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$. A loss gives a penalty $\ell(y, h(\boldsymbol{x}))$ when predicting the value $h(\boldsymbol{x})$ and the observed label is y. We consider *margin losses*, *i.e.* $\ell(y, h(\boldsymbol{x})) = \ell(yh(\boldsymbol{x}))$ (Reid & Williamson, 2010), which are implicitly symmetric: $\ell(yh(\boldsymbol{x})) = \ell(-y \cdot (-h(\boldsymbol{x})))$. For notational convenience, we will often use a generic scalar argument $\ell(x)$. Examples are 01 loss $1\{x < 0\}$, logistic loss $\log(1 + e^{-x})$, square loss $(1 - x)^2$ and hinge loss $[1 - x]_+$.

The goal of binary classification is to select a hypothesis $h \in \mathcal{H}$ that generalizes on $\mathcal{D}$. That is, we aim to minimize the *true risk* on the 01 loss $R_{\mathcal{D},01}(h) \doteq \mathbb{E}_{\mathcal{D}}[1\{yh(\boldsymbol{x}) < 0\}]$. In practice, we only learn from a finite learning sample $\mathcal{S}$ and thus minimize the *empirical ℓ-risk* $R_{\mathcal{S},\ell}(h) \doteq \mathbb{E}_{\mathcal{S}}[\ell(yh(\boldsymbol{x}))] = \frac{1}{m} \sum_{i \in [m]} \ell(y_i h(\boldsymbol{x}_i))$, where ℓ is a tractable upperbound of 01 loss.

Finally, we discuss the meaning of WSL — and in particular of weakly supervised *binary classification*. The difference with the above is at training time: we learn on a sample $\tilde{\mathcal{S}}$ drawn from a noisy distribution $\tilde{\mathcal{D}}$ that may flip, aggregate or suppress labels, while observations are the same. Still, the purpose of learning is unchanged: to minimize the true risk. A rigorous definition is not relevant in our study.

2.2. Background: exponential family and logistic loss

Some background on the exponential family is to come. We can learn a binary classifier fitting a model in the conditional exponential family parametrized by $\boldsymbol{\theta}$: $p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) = \exp(\langle \boldsymbol{\theta}, y\boldsymbol{x} \rangle - \log \sum_{y \in \mathcal{Y}} \exp \langle \boldsymbol{\theta}, y\boldsymbol{x} \rangle)$, with y random variable. The two terms in the exponent are the log-partition function and the sufficient statistic $y_i \boldsymbol{x}_i$, which fully summarizes one example $(\boldsymbol{x}, y)$. The Fisher-Neyman theorem (Lehmann & Casella, 1998, Theorem 6.5) gives a sufficient and necessary condition for sufficiency of a statistic $T(y)$: the probability distribution factors in two functions, such that $\boldsymbol{\theta}$ interacts with the y only through T:

$$p_{\boldsymbol{\theta}}(y) = g_{\boldsymbol{\theta}}(T(y)) g'(y) \ .$$

In our case, it holds that $g'(y|\boldsymbol{x}) = 1$, $T(y|\boldsymbol{x}) = y\boldsymbol{x}$ and $g_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x}) = \exp(\langle \boldsymbol{\theta}, \cdot \rangle - \log \sum_{y \in \mathcal{Y}} \exp(\langle \boldsymbol{\theta}, y\boldsymbol{x} \rangle))$, since the value of y is not needed to compute $g_{\boldsymbol{\theta}}$. This shows how $y\boldsymbol{x}$ is indeed sufficient for y. Now, under the *i.i.d.* assumption,

the log-likelihood of $\boldsymbol{\theta}$ is (the negative of)

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

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

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

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

Step (2) is true since $y \in \mathcal{Y}$. At last, by re-parameterizing $\boldsymbol{\theta}$ and normalizing, we obtain logistic loss. Equation (1) shows how the loss splits into a linear term aggregating the labels and another, label free term. Next, we aim to translate this property for classification with ERM, by transferring the ideas of sufficiency and factorization to a wide set of losses including the ones of Patrini et al. (2014).

3. Loss factorization and sufficiency

The linear term just encountered in logistic loss integrates a well-studied statistical object.

Definition 1 *The (empirical)* mean operator *of a learning sample $\mathcal{S}$ is* $\boldsymbol{\mu}_{\mathcal{S}} \doteq \mathbb{E}_{\mathcal{S}} [y\boldsymbol{x}]$.

We drop the $\mathcal{S}$ when clear by the context. The name *mean operator*, or mean map, is borrowed from the theory of Hilbert space embedding (Quadrianto et al., 2009)[1]. Its importance is due to the injectivity of the map — under conditions on the kernel — which is used in applications such as two-sample and independence tests, feature extraction and covariate shift (Smola et al., 2007). Here, $\boldsymbol{\mu}$ plays the role of sufficient statistic for labels *w.r.t.* a set of losses.

Definition 2 *A function $T(\mathcal{S})$ is said to be a sufficient statistic for a variable z* w.r.t. *a set of losses $\mathcal{L}$ and a hypothesis space $\mathcal{H}$ when for any $\ell \in \mathcal{L}$, any $h \in \mathcal{H}$ and any two samples $\mathcal{S}$ and $\mathcal{S}'$ the empirical ℓ-risk is such that*

$$R_{\mathcal{S},\ell}(h) - R_{\mathcal{S}',f}(h) \text{ does not depend on } z \Leftrightarrow T(\mathcal{S}) = T(\mathcal{S}').$$

This is motivated by the one in Statistics, taking log-odd ratios (Patrini et al., 2014). With the next results, we establish sufficiency of mean operators for a large set of losses.

Theorem 3 (Factorization) *Let $\mathcal{H}$ be the space of linear hypotheses. Assume that a loss ℓ is such that $\ell_o(x) \doteq (\ell(x) - \ell(-x))/2$ is linear. Then, for any sample $\mathcal{S}$ and hypothesis $h \in \mathcal{H}$ the empirical ℓ-risk can be written as*

$$R_{\mathcal{S},\ell}(h) = \frac{1}{2} R_{\mathcal{S}_{2x},\ell}(h) + \ell_o(h(\boldsymbol{\mu}_{\mathcal{S}})) \ ,$$

where $\mathcal{S}_{2x} \doteq \{(\boldsymbol{x}_i, \sigma), i \in [m], \forall \sigma \in \mathcal{Y}\}$.

[1] We keep the lighter notation of linear classifiers, but nothing should prevent the extension to non-parametric models, exchanging $\boldsymbol{x}$ with an implicit feature map $h(\boldsymbol{x})$. See also Theorem 13.

	loss ℓ	odd term ℓ_o
LOL	$\ell(x)$	$-ax$
ρ-loss	$\rho\lvert x\rvert - \rho x + 1$	$-\rho x \ \ (\rho \geq 0)$
unhinged	$1 - x$	$-x$
perceptron	$\max(0, -x)$	$-x$
2-hinge	$\max(-x, 1/2 \max(0, 1 - x))$	$-x$
SPL	$a_\ell + \ell^\star(-x)/b_\ell$	$-x/(2b_\ell)$
logistic	$\log(1 + e^{-x})$	$-x/2$
square	$(1 - x)^2$	$-2x$
Matsushita	$\sqrt{1 + x^2} - x$	$-x$

Table 1. Factorization of linear-odd losses: SPL (including logistic, square and Matsushita) (Nock & Nielsen, 2009), double "2"-hinge and perceptron (du Plessis et al., 2015), unhinged (van Rooyen et al., 2015). For ρ-loss see the text.

Proof We write $R_{\mathcal{S},\ell}(h) = \mathbb{E}_{\mathcal{S}}[\ell(yh(\boldsymbol{x}))\ ]$ as

$$\frac{1}{2}\mathbb{E}_{\mathcal{S}}\Big[\ell(yh(\boldsymbol{x})) + \ell(-yh(\boldsymbol{x})) + \ell(yh(\boldsymbol{x})) - \ell(-yh(\boldsymbol{x}))\Big]$$

$$= \frac{1}{2}\mathbb{E}_{\mathcal{S}}\left[\sum\nolimits_{\sigma \in \mathcal{Y}} \ell(\sigma h(\boldsymbol{x}))\right] + \mathbb{E}_{\mathcal{S}}\Big[\ell_o(yh(\boldsymbol{x}))\Big]$$

$$= \frac{1}{2}\mathbb{E}_{\mathcal{S}_{2x}}\Big[\ell(\sigma h(\boldsymbol{x}))\Big] + \mathbb{E}_{\mathcal{S}}\Big[\ell_o(h(y\boldsymbol{x}))\Big] \ . \tag{3}$$

Step 3 is due to the definition of $\mathcal{S}_{2x}$ and linearity of h. The Theorem follows by linearity of ℓ_o and expectation. ■

Factorization splits ℓ-risk in two parts. A first term is the ℓ-risk computed *on the same loss* on the "doubled sample" $\mathcal{S}_{2x}$ that contains each observation twice, labelled with opposite signs, and hence it is label free. A second term is a loss ℓ_o of h applied to the mean operator $\boldsymbol{\mu}_{\mathcal{S}}$, which aggregates all sample labels. Also observe that ℓ_o is by construction an odd function, *i.e.* symmetric *w.r.t.* the origin. We call the losses satisfying the Theorem linear-odd losses.

Definition 4 *A loss ℓ is a-linear-odd (a-*LOL*) when $\ell_o(x) = (\ell(x) - \ell(-x))/2 = ax$, for any $a \in \mathbb{R}$.*

Notice how this does not exclude losses that are not smooth, convex, or proper (Reid & Williamson, 2010). From now on, we also consider $\mathcal{H}$ as the space linear hypotheses $h(\cdot) = \langle \boldsymbol{\theta}, \cdot \rangle$. (Theorem 12 in Section 6 applies beyond LOLs and linear models.) As a consequence of Theorem 3, $\boldsymbol{\mu}$ is sufficient for all labels.

Corollary 5 *The mean operator $\boldsymbol{\mu}$ is a sufficient statistic for the label y with regard to* LOL*s and $\mathcal{H}$.*

The corollary is at the core of the applications in the paper: the single vector $\boldsymbol{\mu} \in \mathbb{R}^d$ summarizes all information concerning the linear relationship between y and $\boldsymbol{x}$ for losses that are LOL (see also Section 6). Many known losses belong to this class; see Table 1. For logistic loss it holds that (Figure 1(a)):

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

This "symmetrization" is known in the literature (Jaakkola & Jordan, 2000; Gao et al., 2016). Another case of LOL is unhinged loss $\ell(x) = 1 - x$ (van Rooyen et al., 2015) — while standard hinge loss does not factor in a linear term.

The Factorization Theorem 3 generalizes Patrini et al. (2014, Lemma 1) that works for *symmetric proper losses* (SPLs), *e.g.* logistic, square and Matsushita losses. Given a permissible generator ℓ (Kearns & Mansour, 1996; Nock & Nielsen, 2009), *i.e.* $\mathrm{dom}(l) \supseteq [0,1]$, ℓ is strongly convex, differentiable and symmetric with respect to $1/2$, SPLs are defined as $\ell(x) = a_\ell + \ell^\star(-x)/b_\ell$, where $\ell^\star$ is the convex conjugate of ℓ. Then, since $\ell^\star(-x) = \ell^\star(x) - x$:

$$\ell_o(x) = \frac{1}{2}\left(a_l + \frac{\ell^\star(-x)}{b_\ell} - a_l - \frac{\ell^\star(x)}{b_\ell}\right) = -\frac{x}{2b_\ell} \ .$$

A similar result appears in Masnadi-Shirazi (2011, Theorem 11). A natural question is whether the classes SPL and LOL are equivalent. We answer in the negative.

Lemma 6 *The exhaustive class of linear-odd losses is in 1-to-1 mapping with a proper subclass of even functions,* i.e. $\ell_e(x) - ax$, *with* ℓ_e *any even function.*

Interestingly, the proposition also let us engineer losses that always factor: choose any even function ℓ_e with desired properties — it need not be convex nor smooth. The loss is then $\ell(x) = \ell_e(x) - ax$, with a to be chosen. For example, let $\ell_e(x) = \rho|x| + 1$, with $\rho > 0$. $\ell(x) = \ell_e(x) - \rho x$ is a LOL; furthermore, ℓ upperbounds 01 loss and intercepts it in $\ell(0) = 1$. Non-convex ℓ can be constructed similarly. Yet, not all non-differentiable losses can be crafted this way since they are not LOLs. We provide in SM B sufficient and necessary conditions to bound other known losses, *e.g.* hinge and Huber, by LOLs.

From the optimization viewpoint, we may be interested in keeping properties of ℓ after factorization. The good news is that we are dealing with the same ℓ plus a linear term. Thus, if the property of interest is closed under summation with linear functions, then it will hold true. An example is convexity: if ℓ is LOL and convex, so is the factored loss.

The next Theorem sheds new light on generalization bounds on Rademacher complexity with linear hypotheses.

Theorem 7 *Assume* ℓ *is* a-LOL *and* L-*Lipschitz. Suppose* $\mathbb{R}^d \supseteq \mathcal{X} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq X < \infty\}$ *and* $\mathcal{H} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 \leq B < \infty\}$. *Let* $c(X,B) \doteq \max_{y\in\mathcal{Y}} \ell(yXB)$ *and* $\hat{\boldsymbol{\theta}} \doteq \mathrm{argmin}_{\boldsymbol{\theta}\in\mathcal{H}} R_{\mathcal{S},\ell}(\boldsymbol{\theta})$. *Then for any* $\delta > 0$, *with probability at least* $1-\delta$:

$$R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - \inf_{\boldsymbol{\theta}\in\mathcal{H}} R_{\mathcal{D},\ell}(\boldsymbol{\theta}) \leq \left(\frac{\sqrt{2}+1}{4}\right)\cdot\frac{XBL}{\sqrt{m}} + \frac{c(X,B)L}{2}\cdot\sqrt{\frac{1}{m}\log\left(\frac{1}{\delta}\right)} + 2|a|B\cdot\|\boldsymbol{\mu}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathcal{S}}\|_2 \ ,$$

or more explicity

$$R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - \inf_{\boldsymbol{\theta}\in\mathcal{H}} R_{\mathcal{D},\ell}(\boldsymbol{\theta}) \leq \left(\frac{\sqrt{2}+1}{4}\right)\cdot\frac{XBL}{\sqrt{m}} + \frac{c(X,B)L}{2}\sqrt{\frac{1}{m}\log\left(\frac{2}{\delta}\right)} + 2|a|XB\sqrt{\frac{d}{m}\log\left(\frac{2d}{\delta}\right)} \ .$$

The term $\frac{\sqrt{2}+1}{4}\cdot\frac{XBL}{\sqrt{m}}$ is derived by an improved upperbound to the Rademacher complexity of $\mathcal{H}$ computed on $\mathcal{S}_{2x}$ (SM A.3, Lemma 1); we call it in short *complexity*. The former expression displays the contribution of the non-linear part of the loss, keeping aside what is missing: a deviation of the empirical mean operator from its population mean. When $\boldsymbol{\mu}_{\mathcal{S}}$ is not known because of partial label knowledge, the choice of any estimator would affect the bound only through that norm discrepancy. The second expression highlights the interplay of the two loss components. $c(X,B)$ is the only non-linear term, which may well be predominant in the bound for fast-growing losses, *e.g.* strongly convex. Moreover, we confirm that the linear-odd part does not change the complexity and only affects the statistical penalty by a linear factor, with a dependency on d. A last important remark comes from comparing the bound with the one due to Kakade et al. (2009, Corollary 4). Our complexity coefficient is $(\sqrt{2}+1)/4 \approx 0.6$ instead of 2, that is three times smaller. A similar statement may be derived for RKHS on top of Bartlett & Mendelson (2002).

4. Weakly supervised learning

In the next two Sections we discuss applications to WSL. Recall that in this scenario we learn on $\tilde{\mathcal{S}}$ with partially observable labels, but aim to generalize to $\mathcal{D}$. Assume to know an algorithm that learns on $\mathcal{S}$. By sufficiency, Corollary 5, a principled approach to use $\tilde{\mathcal{S}}$ is: (1) estimate $\boldsymbol{\mu}$ from $\tilde{\mathcal{S}}$ and (2) run the algorithm with the LOL computed on the estimated $\boldsymbol{\mu}$. This direction was explored by work on LLP by Quadrianto et al. (2009, with logistic loss) and Patrini et al. (2014, SPL), and in the setting of noisy labels by van Rooyen et al. (2015, unhinged loss) and (Gao et al., 2016, logistic loss). The approach contrasts with *ad-hoc*

Algorithm 1 $\boldsymbol{\mu}$SGD

Input: $\mathcal{S}_{2x}, \boldsymbol{\mu}$, ℓ is a-LOL; $\lambda > 0$; $T > 0$
$m' \leftarrow |\mathcal{S}_{2x}|$
$\boldsymbol{\theta}^0 \leftarrow \mathbf{0}$
For any $t = 1, \ldots, T$:
  Pick $i = i^t \in [m']$ uniformly at random
  $\eta^t \leftarrow (\lambda t)^{-1}$
  Pick any $\boldsymbol{v} \in \partial \ell(y_i \langle \boldsymbol{\theta}^t, \boldsymbol{x}_i \rangle)$
  $\boldsymbol{\theta}^{t+1} \leftarrow (1 - \eta^t \lambda)\boldsymbol{\theta}^t - \eta^t (\boldsymbol{v} + a\boldsymbol{\mu}/2)$
  $\boldsymbol{\theta}^{t+1} \leftarrow \min \left\{ \boldsymbol{\theta}^{t+1}, \boldsymbol{\theta}^{t+1} \sqrt{\lambda^{-1}} / \|\boldsymbol{\theta}^{t+1}\|_2 \right\}$
Output: $\boldsymbol{\theta}^{t+1}$

optimization methods often aiming to recover the latent labels by coordinate descent and EM (Joulin & Bach, 2012). Instead, the only difficulty here is to come up with a well-behaved estimator of $\boldsymbol{\mu}$ — a statistic independent from both h and ℓ. Thereom 7 then assures bounded ℓ-risk and, in turn, true risk. Finite-sample bounds hold under stricter conditions on ℓ (Altun & Smola, 2006; Patrini et al., 2014).

Algorithm 1, $\boldsymbol{\mu}$SGD, adapts SGD to weak supervision. For the sake of presentation, we work on a simple version of SGD based on subgradient descent with L_2 regularization inspired by Shalev-Shwartz et al. (2011). Given $\boldsymbol{\mu}$ changes are trivial: (i) construct $\mathcal{S}_{2x}$ from $\tilde{\mathcal{S}}$ and (ii) sum $a\boldsymbol{\mu}/2$ to the subgradients of each example of $\mathcal{S}_{2x}$. In Section 6 we upgrade proximal algorithms with the same minimal-effort strategy. The next Section shows an estimator of $\boldsymbol{\mu}$ in the case of noisy labels and specializes $\boldsymbol{\mu}$SGD. We also analyze the effect of noise through the lenses of Theorem 7 and discuss a non-standard notion of robustness.

5. Asymmetric label noise

In learning with noisy labels, $\tilde{\mathcal{S}}$ is a sequence of examples drawn from a distribution $\tilde{\mathcal{D}}$, which samples from $\mathcal{D}$ and flips labels at random. Each example $(\boldsymbol{x}_i, \tilde{y}_i)$ is $(\boldsymbol{x}_i, -y_i)$ with probability at most $1/2$ or it is $(\boldsymbol{x}_i, y_i)$ otherwise. The *noise rates* are label dependent[2] by $(p_+, p_-) \in [0, 1/2)^2$ respectively for positive and negative examples, that is, *asymmetric* label noise (ALN) (Natarajan et al., 2013).

5.1. Unbiased estimation of μ

Our first result builds on Natarajan et al. (2013, Lemma 1) that provides a recipe for unbiased estimators of losses. Thanks to the Factorization Theorem 3, instead of estimating the whole ℓ we act on the sufficient statistic:

$$\hat{\boldsymbol{\mu}}_{\mathcal{S}} \doteq \mathbb{E}_{\mathcal{S}} \left[\frac{y - (p_- - p_+)}{1 - p_- - p_+} \boldsymbol{x} \right] . \tag{4}$$

[2]While being independent from the observation.

The estimator is unbiased, that is, its expectation over the noise distribution $\tilde{\mathcal{D}}$ is the population mean operator: $\hat{\boldsymbol{\mu}}_{\tilde{\mathcal{D}}} = \boldsymbol{\mu}_{\mathcal{D}}$. Denote then the risk computed on the estimator as $\hat{R}_{\mathcal{S},\ell}(\boldsymbol{\theta}) \doteq \frac{1}{2} R_{\mathcal{S}_{2x},\ell}(\boldsymbol{\theta}) + a\langle \boldsymbol{\theta}, \hat{\boldsymbol{\mu}}_{\mathcal{S}} \rangle$. Unbiasedness transfers to ℓ-risk: $\hat{R}_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}) = R_{\mathcal{D},\ell}(\boldsymbol{\theta})$, $\forall \boldsymbol{\theta}$ (Proofs in SM A.4). We have thus obtained a good candidate as input for any algorithm implementing our 2-step approach, like $\boldsymbol{\mu}$SGD. But there is more. On one hand, the estimators of Natarajan et al. (2013) may not be convex even when ℓ is so, but this is never the case with LOLs; in fact, $\ell(x) - \ell(-x) = 2ax$ may be seen as alternative sufficient condition to Natarajan et al. (2013, Lemma 4) for convexity. On the other hand, we generalize the approach of van Rooyen et al. (2015) to losses beyond unhinged and to asymmetric noise. We now prove that *any* algorithm minimizing LOLs that uses the estimator in Equation 4 has a non-trivial generalization bound. We further assume that ℓ is Lipschitz.

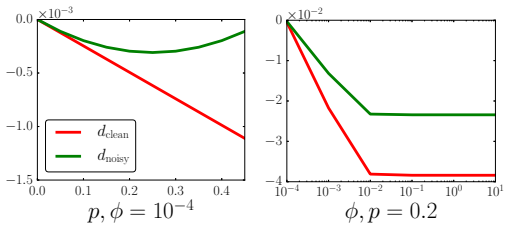
Theorem 8 *Consider the setting of Theorem 7, except that here* $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathcal{H}} \hat{R}_{\tilde{\mathcal{S}},\ell}(\boldsymbol{\theta})$. *Then for any* $\delta > 0$, *with probability at least* $1 - \delta$:

$$R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - \inf_{\boldsymbol{\theta} \in \mathcal{H}} R_{\mathcal{D},\ell}(\boldsymbol{\theta}) \leq \left(\frac{\sqrt{2} + 1}{4} \right) \cdot \frac{XBL}{\sqrt{m}} + \frac{c(X,B)L}{2} \sqrt{\frac{1}{m} \log \left(\frac{2}{\delta} \right)} + \frac{2|a|XB}{1 - p_- - p_+} \sqrt{\frac{d}{m} \log \left(\frac{2d}{\delta} \right)} .$$

Again, the complexity term is tighter than prior work. Natarajan et al. (2013, Theorem 3) proves a factor of $2L/(1 - p_- - p_+)$ that may even be unbounded due to noise, while our estimate shows a constant of about $0.6 < 2$ and it is noise free. In fact, LOLs are such that noise affects only the linear component of the bound, as a direct effect of factorization. Although we are not aware of any other such results, this is intuitive: Rademacher complexity is computed regardless of sample labels and therefore is unchanged by label noise. Furthermore, depending on the loss, the effect of (limited) noise on generalization may be also be negligible since $c(X, B)$ could be very large for losses like strongly convex. This last remark fits well with the property of robustness that we are about to investigate.

5.2. Every LOL is approximately noise-robust

The next result comes in pair with Theorem 8: it holds regardless of algorithm and (linear-odd) loss of choice. In particular, we demonstrate that every learner enjoys a distribution-dependent property of robustness against asymmetric label noise. No estimate of $\boldsymbol{\mu}$ is involved and hence the theorem applies to any naïve supervised learner on $\tilde{\mathcal{S}}$. We first refine the notion of robustness of Ghosh et al. (2015) and van Rooyen et al. (2015) in a weaker sense.



$p, \phi = 10^{-4}$ $\phi, p = 0.2$

Figure 2. Behavior of Theorem 10 on synthetic data. Definition of the axes within the text.

Definition 9 *Let $(\boldsymbol{\theta}^\star, \tilde{\boldsymbol{\theta}}^\star)$ respectively be the minimizers of $(R_{\mathcal{D},\ell}(\boldsymbol{\theta}), R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}))$ in $\mathcal{H}$. ℓ is said ϵ-*ALN *robust if for any $\mathcal{D}, \tilde{\mathcal{D}}$, $R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}^\star) - R_{\tilde{\mathcal{D}},\ell}(\tilde{\boldsymbol{\theta}}^\star) \leq \epsilon$.*

The distance of the two minimizers is measured by empirical ℓ-risk under expected label noise. 0-ALN robust losses are also ALN robust: in fact if $R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}^\star) = R_{\tilde{\mathcal{D}},\ell}(\tilde{\boldsymbol{\theta}}^\star)$ then $\boldsymbol{\theta}^\star \in \operatorname{argmin}_{\boldsymbol{\theta}} R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta})$. And hence if $R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta})$ has a unique global minimum, that will be $\boldsymbol{\theta}^\star$. More generally

Theorem 10 *Assume $\{\boldsymbol{\theta} \in \mathcal{H} : ||\boldsymbol{\theta}||_2 \leq B\}$. Then every a-*LOL *is ϵ-*ALN*. That is*

$$R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}^\star) - R_{\tilde{\mathcal{D}},\ell}(\tilde{\boldsymbol{\theta}}^\star) \leq 4|a|B \max\{p_+, p_-\}\|\boldsymbol{\mu}_{\mathcal{D}}\|_2$$

Moreover: (1) If $\|\boldsymbol{\mu}_{\mathcal{D}}\|_2 = 0$ for $\mathcal{D}$ then every LOL *is* ALN *for any $\tilde{\mathcal{D}}$. (2) Suppose that ℓ is also once differentiable and γ-strongly convex. Then $\|\boldsymbol{\theta}^\star - \tilde{\boldsymbol{\theta}}^\star\|_2^2 \leq 2\epsilon/\gamma$.*

Unlike Theorem 8, this bound holds in expectation over the noisy risk $R_{\tilde{\mathcal{D}},\ell}$. Its shape depends on the population mean operator, a *distribution-dependent* quantity. There are two immediate corollaries. When $\|\boldsymbol{\mu}_{\mathcal{D}}\|_2 = 0$, we obtain optimality for all LOLs. The second corollary goes further, limiting the minimizers' distance when losses are differentiable and strongly convex. But even more generally, under proper compactness assumptions on the domain of ℓ, Theorem 10 tells us much more: in the case $R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta})$ has a unique global minimizer, the smaller $\|\boldsymbol{\mu}_{\mathcal{D}}\|_2$, the closer the minimizer *on noisy data* $\tilde{\boldsymbol{\theta}}^\star$ will be to the minimizer *on clean data* $\boldsymbol{\theta}^\star$. Therefore, assuming an efficient algorithm that computes a model not far from the global optimum $\tilde{\boldsymbol{\theta}}^\star$, that will be not far from $\boldsymbol{\theta}^\star$ either. This is true in spite of the presence of local minima and/or saddle points.

Long & Servedio (2010) proves that no convex potential[3] is noise tolerant, that is, 0-ALN robust. This is not a contradiction. To show the negative statement, the authors craft a case of $\mathcal{D}$ breaking all such losses. And in fact that choice of $\mathcal{D}$ does not meet optimality in our bound, because $\|\boldsymbol{\mu}_{\mathcal{D}}\|_2 = \frac{1}{4}(18\gamma^2 + 6\gamma + 1) > 0$, with $\gamma \in (0, 1/6)$. In contrast, we show that every element of the broad class of LOLs is approximately robust, as opposed to a *worst-case* statement. Finally, compare our ϵ-robustness to the one of Ghosh et al. (2015): $R_{\mathcal{D},\ell}(\tilde{\boldsymbol{\theta}}^*) \leq (1-2\max(p_-, p_+))^{-1} R_{\mathcal{D},\ell}(\boldsymbol{\theta}^*)$. Such bound, while relating the (non-noisy) ℓ-risks, is not data-dependent and may be not much informative for high noise rates.

[3] A convex potential is a loss $l \in C^1$, convex, such that $\ell(0) < 0$ and $\ell(x) \to 0$ for $x \to \infty$. Many convex potentials are LOLs but not all. An example is e^{-x}.

Algorithm 2 $\boldsymbol{\mu}$SGD applied on noisy labels

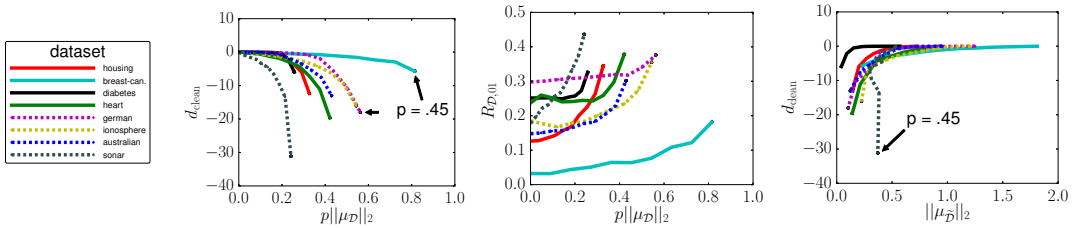
Input: $\tilde{\mathcal{S}}, \ell \in$ LOL; $\lambda > 0$; $T > 0$
$\mathcal{S}_{2x} \doteq \{(\boldsymbol{x}_i, \sigma), i \in [m], \forall \sigma \in \mathcal{Y}\}$
$\hat{\boldsymbol{\mu}}_{\tilde{\mathcal{S}}} \leftarrow$ Equation 4
$\boldsymbol{\theta} \leftarrow \boldsymbol{\mu}$-SGD$(\mathcal{S}_{2x}, \hat{\boldsymbol{\mu}}_{\tilde{\mathcal{S}}}, \lambda, T)$
Output: $\boldsymbol{\theta}$

5.3. Experiments

We analyze experimentally the theory so far developed. From now on, we assume to know p_+ and p_- at learning time. In principle they may be tuned as hyper-parameters (Natarajan et al., 2013), at least for small $|\mathcal{Y}|$ (Sukhbaatar & Fergus, 2014). While being out of scope, practical noise estimators are studied (Menon et al., 2015; Scott, 2015).

We begin by building a toy planar dataset to probe the behavior of Theorem 10. It is made of four observations: $(0, 1)$ and $(\phi/3, 1/3)$ three times, with the first example the only negative, repeated 5 times. We consider this the distribution $\mathcal{D}$ so as to calculate $\|\boldsymbol{\mu}_{\mathcal{D}}\|_2 = \phi^2/4$. We fix $p_+, p_- = 0.2 = p$ and control ϕ to measure the discrepancy $d_{\text{noisy}} \doteq R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}^\star) - R_{\tilde{\mathcal{D}},\ell}(\tilde{\boldsymbol{\theta}}^\star)$, its counterpart d_{clean} computed on $\mathcal{D}$, and how the two minimizers "differ in sign" by $d_{\text{models}} \doteq \langle \boldsymbol{\theta}^\star, \tilde{\boldsymbol{\theta}}^\star \rangle / \|\boldsymbol{\theta}^\star\|_2 \|\tilde{\boldsymbol{\theta}}^\star\|_2$. The same simulation is run varying the noise rates with constant $\phi = 10^{-4}$. We learn with $\lambda = 10^{-6}$ by standard square loss. Results are in Figure 2. The closer the parameters to 0, the smaller $d_{\text{clean}} - d_{\text{noisy}}$, while they are equal when each parameter is individually 0. d_{models} is negligible, which is good news for the 01-risk on sight.

Algorithm 2 learns with noisy labels on the estimator of Equation 4 and by calling the black box of $\boldsymbol{\mu}$SGD. The next results are based on UCI data. We learn with logistic loss, without model's intercept and set $\lambda = 10^{-6}$ and $T = 4 \cdot 2m$ (4 epochs). We measure d_{clean} and $R_{\mathcal{D},01}$, injecting symmetric label noise $p \in [0, 0.45)$ and averaging over 25 runs. Again, we consider *the whole distribution* so as to play with the ingredients of Theorem 10. Figure 3(a) confirms how the combined effect of $p\|\boldsymbol{\mu}_{\mathcal{D}}\|_2$ can explain most variation of d_{clean}. While this is not strictly implied by Theorem 10 that only involves d_{noisy}, the observed behavior is expected. A similar picture is given in Figure 3(b) which displays the true risk $R_{\mathcal{D},01}$ computed on the minimizer $\tilde{\boldsymbol{\theta}}^\star$ of $\tilde{\mathcal{S}}$. From 3(a) and 3(b) we also deduce that



(a) $p\|\mu_{\mathcal{D}}\|_2$ *vs.* d_{clean} (b) $p\|\mu_{\mathcal{D}}\|_2$ *vs.* $R_{\mathcal{D},01}$ (c) $\|\mu_{\tilde{\mathcal{D}}}\|_2$ *vs.* d_{clean}

Figure 3. How mean operator and noise rate condition risks. $d_{\text{clean}} \doteq R_{\mathcal{D},\ell}(\boldsymbol{\theta}^\star) - R_{\mathcal{D},\ell}(\tilde{\boldsymbol{\theta}}^\star)$.

large $\|\boldsymbol{\mu}_{\mathcal{D}}\|_2$ is a good proxy for generalization with linear classifiers; see the relative difference between points at the same level of noise. Finally, we have also monitored $\boldsymbol{\mu}_{\tilde{\mathcal{D}}}$. Figure 3(c) shows that large $\|\boldsymbol{\mu}_{\tilde{\mathcal{D}}}\|_2$ indicates small d_{clean} as well. This remark can be useful in practice, when the norm can be estimated from $\tilde{\mathcal{S}}$, as opposite to p and $\boldsymbol{\mu}_{\mathcal{D}}$, and used to anticipate the effect of noise on the task at hand.

We conclude with a systematic study of hold-out error of $\boldsymbol{\mu}$SGD. The same datasets are now split in 1/5 test and 4/5 training sets once at random. In contrast with the previous experimental setting we perform cross-validation of $\lambda \in 10^{\{-3,\dots,+3\}}$ on 5-folds in the training set. We compare with vanilla SGD run on corrupted sample $\tilde{\mathcal{S}}$ and measure the gain from estimating $\hat{\boldsymbol{\mu}}_{\tilde{\mathcal{S}}}$. The other parameters l, T, λ are the same for both algorithms; the learning rate η is untouched from Shalev-Shwartz et al. (2011) and not tuned for $\boldsymbol{\mu}$SGD. The only differences are in input and gradient update. Table 2 reports test error for SGD and its difference with $\boldsymbol{\mu}$SGD, for a range of values of (p_-, p_+). $\boldsymbol{\mu}$SGD beats systematically SGD with large noise rates, and yet performs in pair under low or null noise. Interestingly, in the presence of very intense noise $p_+ \approx .5$, $\boldsymbol{\mu}$SGD still learns sensible models and improves up to 55% relatively to the error of SGD, which is often doomed to random guess.

6. Discussion and conclusion

Mean and covariance operators The intuition behind the relevance of the mean operator becomes clear once we rewrite it as follows.

Lemma 11 *Let $\pi_+ \doteq \mathbb{E}_{\mathcal{S}} 1\{y > 0\}$ be the positive label proportion of $\mathcal{S}$. Then $\boldsymbol{\mu}_{\mathcal{S}} = \mathbb{C}ov_{\mathcal{S}}[\boldsymbol{x}, y] + (2\pi_+ - 1)\mathbb{E}_{\mathcal{S}}[\boldsymbol{x}]$.*

We have come to the unsurprising fact that — when observations are centered — the covariance $\mathbb{C}\text{ov}_{\mathcal{S}}[\boldsymbol{x}, y]$ is what we need to know about the labels for learning linear models. The rest of the loss may be seen as a data-dependent regularizer. However, notice how the condition $\|\boldsymbol{\mu}_{\mathcal{D}}\|_2 = 0$ does not implies $\mathbb{C}\text{ov}_{\mathcal{D}}[\boldsymbol{x}, y] = 0$, which would make linear classification hard and limit Theorem 10's validity to degenerate cases. A kernelized version of this Lemma is given in Song et al. (2009).

The generality of factorization Factorization is ubiquitous for any (margin) loss, beyond the theory seen so far. A basic fact of real analysis supports it: a function ℓ is (uniquely) the sum of an even function ℓ_e and an odd ℓ_o:

$$\ell(x) = \frac{1}{2}\left(\ell(x) + \ell(-x) + \ell(x) - \ell(-x)\right) = \ell_e(x) + \ell_o(x)$$

One can check that ℓ_e and ℓ_o are indeed even and odd (Figure 1). This is actually all we need to factor ℓ.

Theorem 12 (Factorization) *For any sample $\mathcal{S}$ and hypothesis h the empirical ℓ-risk can be written as*

$$R_{\mathcal{S},\ell}(h) = \frac{1}{2}\mathbb{E}_{\mathcal{S}}\left[\sum\nolimits_{\sigma\in\mathcal{Y}} \ell(\sigma h(\boldsymbol{x}))\right] + \mathbb{E}_{\mathcal{S}}\left[\ell_o(yh(\boldsymbol{x}))\right]$$

where $\ell_o(\cdot)$ is odd and $\ell_e(\cdot) \doteq \sum_{\sigma\in\mathcal{Y}} \ell(\sigma h(\cdot))$ is even and both uniquely defined.

Its range of validity is exemplified by 01 loss, a non-convex discontinuous piece-wise linear function, which factors as

$$\ell_e(x) = \begin{cases} \frac{1}{2} & x \neq 0 \\ 1 & \text{otherwise} \end{cases}, \quad \ell_o(x) = -\frac{1}{2}\,\text{sign}(x) \ .$$

It follows immediately that $\mathbb{E}_{\mathcal{S}}[\ell_o(\cdot)]$ is sufficient for y. However, ℓ_o is a function of model $\boldsymbol{\theta}$. This defeats the purpose of a sufficient statistic, which we aim to be computable from data only and it is the main reason to define LOLs. The Factorization Theorem 12 can also be stated for RKHS. To show that, notice that we satisfy all hypotheses of the Representer Theorem (Schölkopf & Smola, 2002).

Theorem 13 *Let $h(\boldsymbol{x}) : \mathcal{X} \to \mathcal{H}$ be a feature map into a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ with symmetric positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, such that $h : \boldsymbol{x} \to k(\cdot, \boldsymbol{x})$. For any learning sample $\mathcal{S}$, the empirical ℓ-risk $R_{\mathcal{S},\ell}(h)$ with $\Omega : ||h||_{\mathcal{H}} \to \mathbb{R}^+$ regularization can be written as*

$$\frac{1}{2}\mathbb{E}_{\mathcal{S}}\left[\sum\nolimits_{\sigma\in\mathcal{Y}} \ell(\sigma h(\boldsymbol{x}))\right] + \mathbb{E}_{\mathcal{S}}\left[\ell_o(yh(\boldsymbol{x}))\right] + \Omega(||h||_{\mathcal{H}})$$

and the optimal hypothesis admits a representation of the form $h(\boldsymbol{x}) = \sum_{i\in[m]} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i)$.

$(p_-, p_+) \rightarrow$	$(.00, .00)$		$(.20, .00)$		$(.20, .10)$		$(.20, .20)$		$(.20, .30)$		$(.20, .40)$		$(.20, .49)$	
dataset	SGD	$\boldsymbol{\mu}$SGD	SGD	$\boldsymbol{\mu}$SGD	SGD	$\boldsymbol{\mu}$SGD	SGD	$\boldsymbol{\mu}$SGD	SGD	$\boldsymbol{\mu}$SGD	SGD	$\boldsymbol{\mu}$SGD	SGD	$\boldsymbol{\mu}$SGD
australian	0.13	$+.01$	0.15	$\mathbf{-.01}$	0.14	$\pm.00$	0.14	$+.01$	0.16	$\mathbf{-.01}$	0.26	$\mathbf{-.09}$	0.45	$\mathbf{-.25}$
breast-can.	0.02	$+.01$	0.03	$\pm.00$	0.03	$\pm.00$	0.03	$\pm.00$	0.05	$\mathbf{-.01}$	0.11	$\mathbf{-.06}$	0.17	$\mathbf{-.08}$
diabetes	0.28	$\mathbf{-.03}$	0.29	$\mathbf{-.03}$	0.29	$\mathbf{-.03}$	0.27	$\mathbf{-.02}$	0.28	$\mathbf{-.02}$	0.39	$\mathbf{-.13}$	0.59	$\mathbf{-.22}$
german	0.27	$\mathbf{-.02}$	0.26	$\pm.00$	0.27	$\mathbf{-.02}$	0.29	$\mathbf{-.02}$	0.31	$\mathbf{-.01}$	0.31	$\pm.00$	0.31	$\pm.00$
heart	0.15	$+.01$	0.17	$\mathbf{-.01}$	0.16	$\pm.00$	0.17	$\pm.00$	0.18	$\mathbf{-.01}$	0.26	$\mathbf{-.08}$	0.35	$\mathbf{-.15}$
housing	0.17	$\mathbf{-.03}$	0.23	$\mathbf{-.05}$	0.22	$\mathbf{-.04}$	0.20	$\mathbf{-.02}$	0.22	$\mathbf{-.03}$	0.34	$\mathbf{-.12}$	0.41	$\mathbf{-.13}$
ionosphere	0.14	$+05$	0.19	$\mathbf{-.05}$	0.20	$\mathbf{-.05}$	0.20	$\mathbf{-.03}$	0.21	$\mathbf{-.03}$	0.35	$\mathbf{-.13}$	0.54	$\mathbf{-.29}$
sonar	0.27	$\pm.00$	0.29	$+.02$	0.29	$+.01$	0.34	$\mathbf{-.04}$	0.36	$\mathbf{-.03}$	0.43	$\mathbf{-.10}$	0.45	$\mathbf{-.05}$

Table 2. Test error for SGD and $\boldsymbol{\mu}$SGD over 25 trials of artificially corrupted datasets.

All paper may be read in the context of non-parametric models, with the *kernel* mean operator as sufficient statistic. Finally, we can show factorization for regression with square loss (SM C), a result that opens further applications.

The linear-odd losses of du Plessis et al. (2015) This work shows that a linear-odd condition on a *convex* ℓ allows one to derive a tractable, *i.e.* still convex, loss for learning with *positive and unlabelled data*. The approach is similar to ours as it isolates a label-free term in the loss, with the goal of leveraging on the unlabelled examples too. Interestingly, the linear term of their Equation 4 can be seen as a mean operator estimated as $\hat{\boldsymbol{\mu}} \doteq \mathbb{P}(y=1) \cdot \mathbb{E}_{\mathcal{S}_+}[\boldsymbol{x}]$, where $\mathcal{S}_+$ is the set of positive examples. Their manipulation of the loss *is not* equivalent to Theorem 3 though, as explained with details in (SM D). Beside that, since we reason at the higher level of WSL, we can frame a solution for this setting by calling $\boldsymbol{\mu}$SGD on $\hat{\boldsymbol{\mu}}$ defined above or by building on estimators derived from Patrini et al. (2014).

Learning reductions Solving a machine learning problem by solutions to other learning problems is a *learning reduction* (Beygelzimer et al., 2015). Our work *does* fit into this framework. Following Beygelzimer et al. (2005), we define a WSL task as a triple $(\mathcal{K}, \mathcal{Y}, \ell)$, with weakly supervised advice $\mathcal{K}$, predictions space $\mathcal{Y}$ and loss ℓ, and we reduce to binary classification $(\mathcal{Y}, \mathcal{Y}, \ell)$. Our reduction is somehow simple, in the sense that $\mathcal{Y}$ does not change *and neither does* ℓ. Although, Algorithm 1 modifies the internal code of the "oracle learner" which contrasts with the concept of reduction. Anyway, we could as well write subgradients as

$$\frac{1}{2}\left(\partial\ell(\langle\boldsymbol{\theta}^t, \boldsymbol{x}_i\rangle) + \partial\ell(-\langle\boldsymbol{\theta}^t, \boldsymbol{x}_i\rangle) + a\boldsymbol{\mu}\right) \enspace ,$$

which equals $\partial\ell$, and thus the oracle would be untouched.

Beyond $\boldsymbol{\mu}$SGD META-$\boldsymbol{\mu}$SGD is intimately similar to stochastic average gradient (SAG) (Schmidt et al., 2013). Let $g_e^{i,t}(\boldsymbol{\theta}) \in \partial\ell_e(y_i\langle\boldsymbol{\theta}, \boldsymbol{x}_i\rangle)$ if $i = i^t$ (example i picked at time t), otherwise $= g_e^{i,t-1}(\boldsymbol{\theta})$. Define the same for ℓ_o accordingly. Then, SAG's model update is:

$$\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t - \frac{\eta}{m}\sum\nolimits_{i\in[m]} g_e^{i,t}(\boldsymbol{\theta}^t) - \frac{\eta}{m}\sum\nolimits_{i\in[m]} g_o^{i,t}(\boldsymbol{\theta}^t) \enspace ,$$

and recalling that $a\boldsymbol{\mu}_{\mathcal{S}} = \mathbb{E}_{\mathcal{S}}[\partial\ell_o(\boldsymbol{\theta})]$, $\boldsymbol{\mu}$SGD's update is

$$\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t - \eta\,\partial\ell_e^i(\boldsymbol{\theta}^t) - \frac{\eta}{m}\sum\nolimits_{i\in[m]} \partial\ell_o^i(\boldsymbol{\theta}^t) \enspace .$$

From this parallel, the two algorithms appear to be variants of a more general sampling mechanism of examples *and* gradient components, at each step. More generally, stochastic gradient is just *one* suit of algorithms that fits into our 2-step learning framework. Proximal methods (Bach et al., 2012) are another noticeable example. The same *modus operandi* leads to a proximal step of the form:

$$\boldsymbol{\theta}^{t+1} \leftarrow \mathrm{prox}_{\Theta}\left(\boldsymbol{\theta}^t + \eta\left(\partial R_{\mathcal{S}_{2x},\ell}(\boldsymbol{\theta}^t) + \frac{a}{2}\boldsymbol{\mu}\right)\right)$$

with $\mathrm{prox}_g(x) = \mathrm{argmin}_{x'}\, g(x') + \frac{1}{2}\|x - x'\|_2^2$ and $\Theta(\cdot)$ the regularizer. Once again, the adaptation works by summing $\boldsymbol{\mu}$ in the gradient step and changing the input to $\mathcal{S}_{2x}$.

A better (?) picture of robustness The worst-case result of Long & Servedio (2010), like any extreme-case argument, should be handled with care. It does not give the big picture for all data we may encounter in a real world, but only the most pessimistic. We present such a global view which appears better than expected: learning from noisy data does not necessarily reduce convex losses to a singleton (van Rooyen et al., 2015) but depends on the mean operator for a broad set of them. Quite surprisingly, factorization also marries two opposite views in one formula[4]:

$$\ell(x) = \frac{1}{2}\Big(\underbrace{\ell(x) + \ell(-x)}_{=\text{const} \Rightarrow \text{0-ALN}} + \underbrace{\ell(x) - \ell(-x)}_{=ax \Rightarrow \epsilon\text{-ALN}}\Big)$$

To conclude, we have seen how losses factor in a way that we can isolate the *contribution of supervision*. This has several implications both on theoretical and practical grounds: learning theory, formal analysis of label noise robustness, and adaptation of algorithms to handle poorly labelled data. An interesting question is whether factorization would let one identify what really matters in learning that is instead *completely unsupervised*, and to do so with more complex models than the ones considered here, as for example deep architectures.

[4] See (Ghosh et al., 2015, Theorem 1).

Acknowledgements

This research originated during a visit of the first author to École Polytechnique. We thank Aditya Menon for insightful feedback on an earlier version. Work carried out in NICTA which was supported by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Center of Excellence Program.

References

Altun, Y. and Smola, A. J. Unifying divergence minimization and statistical inference via convex duality. In *19 th COLT*, pp. 139–153, 2006.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

Bartlett, P.-L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3, 2002.

Beygelzimer, A., Dani, V., Hayes, T., Langford, J., and Zadrozny, B. Error limiting reductions between classification tasks. In *22 th ICML*, pp. 49–56, 2005.

Beygelzimer, A., III, H. Daumé, Langford, J., and Mineiro, P. Learning reductions that really work. *arXiv:1502.02704*, 2015.

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

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

Ding, N. and Vishwanathan, S. V. N. t-logistic regression. In *NIPS*24*, pp. 514–522, 2010.

du Plessis, M C., Niu, G., and Sugiyama, M. Convex formulation for learning from positive and unlabeled data. In *32 th ICML*, pp. 1386–1394, 2015.

Gao, W., Wang, L., Li, Y.-F., and Zhou, Z.-H. Risk minimization in the presence of label noise. In *Proc. of the 30th AAAI Conference on Artificial Intelligence*, 2016.

Garcıa-Garcıa, D. and Williamson, R. C. Degrees of supervision. In *NIPS*25 workshops*, 2011.

Ghosh, A., Manwani, N., and Sastry, P. S. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015.

Hernandez-Gonzalez, J., Inza, I., and Lozano, J.A. Weak supervision and other non-standard classification problems: a taxonomy. In *Pattern Recognition Letters*. Elsevier, 2016.

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

Joulin, A. and Bach, F. R. A convex relaxation for weakly supervised classifiers. In *29 th ICML*, pp. 1279–1286, 2012.

Kakade, S. M., Sridharan, K., and Tewari, A. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*23*, pp. 793–800, 2009.

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

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

Long, P. M. and Servedio, R. A. Random classification noise defeats all convex potential boosters. *Machine learning*, 78(3):287–304, 2010.

Masnadi-Shirazi, H. *The design of bayes consistent loss functions for classification*. PhD thesis, University of California at San Diego, 2011.

Masnadi-Shirazi, H., Mahadevan, V., and Vasconcelos, N. On the design of robust classifiers for computer vision. In *Proc. of the 23rd IEEE CVPR*, pp. 779–786, 2010.

Menon, A., Rooyen, B. Van, Ong, C. S., and Williamson, B. Learning from corrupted binary labels via class-probability estimation. In *32 th ICML*, 2015.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In *NIPS*27*, pp. 1196–1204, 2013.

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

Patrini, G., Nock, R., Rivera, P., and Caetano, T. (Almost) no label no cry. In *NIPS*28*, pp. 190–198, 2014.

Quadrianto, N., Smola, A. J., Caetano, T. S., and Le, Q. V. Estimating labels from label proportions. *JMLR*, 10, 2009.

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

Schmidt, M., Roux, N. L., and Bach, F. Minimizing finite sums with the stochastic average gradient. *arXiv:1309.2388*, 2013.

Schölkopf, B. and Smola, A. J. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Scott, C. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *18^{th} AISTATS*, 2015.

Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

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

Song, L., Huang, J., Smola, A. J., and Fukumizu, K. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *26 th ICML*. ACM, 2009.

Stempfel, G. and Ralaivola, L. Learning SVMs from sloppily labeled data. In *Artificial Neural Networks (ICANN)*, pp. 884–893. Springer, 2009.

Sukhbaatar, S. and Fergus, R. Learning from noisy labels with deep neural networks. *arXiv:1406.2080*, 2014.

van Rooyen, B., Menon, A. K., and Williamson, R. C. Learning with symmetric label noise: The importance of being unhinged. In *NIPS*29*, 2015.