Robust Active Label Correction

Jan Kremer
Department of Computer Science
University of Copenhagen

Fei Sha
Department of Computer Science
University of Southern California

Christian Igel
Department of Computer Science
University of Copenhagen

Abstract

Active label correction addresses the problem of learning from input data for which noisy labels are available (e.g., from imprecise measurements or crowd-sourcing) and each true label can be obtained at a significant cost (e.g., through additional measurements or human experts). To minimize these costs, we are interested in identifying training patterns for which knowing the true labels maximally improves the learning performance. We approximate the true label noise by a model that learns the aspects of the noise that are class-conditional (i.e., independent of the input given the observed label). To select labels for correction, we adopt the active learning strategy of maximizing the expected model change. We consider the change in regularized empirical risk functionals that use different pointwise loss functions for patterns with noisy and true labels, respectively. Different loss functions for the noisy data lead to different active label correction algorithms. If loss functions consider the label noise rates, these rates are estimated during learning, where importance weighting compensates for the sampling bias. We show empirically that viewing the true label as a latent variable and computing the maximum likelihood estimate of the model parameters performs well across all considered problems. A maximum a posteriori estimate of the model parameters was beneficial in most test cases. An image classification experiment using convolutional neural networks demonstrates that the class-conditional noise model, which can be learned efficiently, can guide re-labeling in real-world applications.

Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

1 INTRODUCTION

Acquiring data with noisy labels for supervised learning is often cheap and simple, while obtaining reliable labels remains difficult and/or costly. For instance, in astronomy huge amounts of photometric data from sky surveys are available. Noisy labels can be obtained using crowd-sourcing or automated labeling, but getting a reliable label may require an expert or even additional costly spectroscopic measurements. Another example are medical images, which can be labeled either unreliably by medical students or by expensive experts (Urner et al., 2012). If we are willing to invest in getting high quality labels for some of our training data in order to increase the generalization performance of a machine learning model, two fundamental questions arise. First, how should we learn from both noisy and true labels? Second, which training examples should be re-labeled? We address these questions by devising tailored loss functions and corresponding *active label correction* strategies that try to identify examples for which obtaining the true labels would be most helpful. Active label correction has been considered before by Rebbapragada et al. (2012) and as *learning from weak teachers* by Urner et al. (2012). We incorporate a label noise model, based on which we can derive algorithms for learning and re-labeling in a principled way. This noise model must be simple so that its parameters can be estimated efficiently during training.

Therefore, we consider a model for the parts of the noise distribution that are class-conditional (Angluin and Laird, 1988), that is, we learn label noise rates that depend on the true class, but are independent of the instance covariate. In practice, the noise distribution will be more complex. However, as we will see in our experiments on real-world data, the class-conditional model can capture aspects of the label noise that help guiding the re-labeling (see also Patrini et al., 2017). For selecting noisy examples for correction, we adopt the strategy from active learning to select those points that promise to change the model the most in expectation (Settles and Craven, 2008; Settles et al., 2008).

Next, we summarize additional related work and our main contributions. Section 2 introduces the general active label correction framework, different pointwise loss functions for noisy examples, and the resulting general label correction algorithms. Section 3 derives concrete algorithms for logistic regression. Section 4 presents an importance-weighting method for estimating the noise model during learning. We present experimental results on a range of datasets using logistic regression and convolutional neural networks in section 5.

Related work. Label noise can degrade the accuracy of a learning algorithm to a great extent, and there are different ways of dealing with this problem, see the survey by Frenay and Verleysen (2014). One way is to incorporate a label noise model into the loss function of logistic regression (Bootkrajang and Kabán, 2012; Natarajan et al., 2013) or deep neural networks (Reed et al., 2015; Sukhbaatar and Fergus, 2015; Xiao et al., 2015; Patrini et al., 2017). We follow this approach to guide label corrections and to mitigate the effects of label noise on not yet corrected training examples. The work most similar to ours is by Rebbapragada et al. (2012), although they do not model the label noise. They use uncertainty sampling to select the next example to correct and show that this improves over random selection. Two relevant, purely theoretical contributions have been made by Zhang and Chaudhuri (2015) and Urner et al. (2012). Zhang and Chaudhuri are closer to the standard active learning setting. In their scenario it is possible to obtain the labels of newly queried examples using a weak and a strong labeler, while we assume that all noisy labels are readily available right from the beginning. Their algorithm, which has to our knowledge never been empirically tested, comes with theoretical guarantees, but the need to select a target error and a confidence level *a priori* is problematic in practice. Urner et al. (2012) make assumptions on the noisy labeling which are very different from our noise model and which are based on a notion of neighborhood in the input space. Roughly speaking, if the labels are homogeneous in a neighborhood, then the noise rate in that neighborhood has to be low; if the labels are heterogeneous, the noise rate has to be high. Label correction has also been considered in the crowd-sourcing community (Sheng et al., 2008; Zhao et al., 2011). In these works the authors examine the problem of noisy labelers and consider the trade-off between sampling new examples and asking for additional labels for an already sampled example. In our case we assume that labels are corrected in a reliable way, that the dataset is fixed, and that the noise is inherent (i.e., there is not necessarily a distribution of labels from different labelers). This setting is common in practice, where one often has to analyze data without having access to the full acquisition pipeline.

One might ask whether the underlying problem could simply be reduced to the standard active learning problem and vice versa. While possible, this discards the valuable information in the noisy labels. We show empirically that, in general, such reductions do not achieve the same performance as the methods we present.

Main contributions. (i) We introduce noise-aware loss functions for active label correction. These loss functions attenuate the influence of noisy examples and inform the selection for re-labeling. (ii) We adopt the maximum expected model change strategy for the proposed regularized risk functionals and devise novel algorithms for active label correction. (iii) We show how to simultaneously learn noise and classification model parameters using importance-weighting. (vi) We provide an empirical comparison demonstrating that *maximum likelihood weighted uncertainty re-labeling (ML-WURL)*, which views the true label as a latent variable and computes the maximum likelihood estimate of the model parameters, performs well across all considered test problems. The algorithm *robust maximum a posteriori WURL (MAP-WURL)* using a maximum a posteriori estimates of the model parameters was often faster. Both methods perform particularly well when the underlying problem could not be learned with a high accuracy based on just a few noise-free data points and never performed much worse otherwise. (v) A real-world image classification experiment demonstrates that the class-conditional noise model, which can be learned efficiently, can indeed guide re-labeling even if the real noise distribution is more complex.

2 ACTIVE LABEL CORRECTION BY MAXIMIZING EXPECTED MODEL CHANGE

We consider classification problems with input space $\mathcal{X}$, label space $\mathcal{Y}$, and pointwise loss function ℓ. The goal is to minimize the risk $R(h) = \mathbb{E}_{(x,y)\sim p}[\ell(h(x), y)]$ over hypotheses $h \in \mathcal{H}$ for an unknown distribution p. For a training pattern $(x, y) \sim p$, we may initially only know $(x, \tilde{y})$, the input x and a corresponding *noisy label* $\tilde{y} \in \mathcal{Y}$. However, we can obtain the *true label* y at a considerable cost. Although we refer to y as *the* true label, we do not presume zero Bayes risk.[1]

We assume *class-conditional label noise* (Angluin and Laird, 1988). That is, the noisy label is conditionally independent of the input given the true label, $p(\tilde{y} \mid x, y) = p(\tilde{y} \mid y)$. This noise model is well-studied

[1]For all subsequent considerations it is actually not necessary to be able to obtain the true label; a significantly more accurate label is sufficient.

in the label noise literature (Bootkrajang and Kabán, 2012; Natarajan et al., 2013; Menon et al., 2015; Liu and Tao, 2016). The model has the advantage that learning its parameters typically requires only a few observations of noisy labels with corresponding true labels. Let us consider binary classification (the extension to multiple classes is straight-forward) and define the noise rates as $\rho_{+1} := p(\tilde{Y} = -1|Y = +1)$ and $\rho_{-1} := p(\tilde{Y} = +1|Y = -1)$, with $\rho_{+1} + \rho_{-1} < 1$, where $Y, \tilde{Y} \in \mathcal{Y} = \{-1, +1\}$ are random variables for the true and the noisy label, respectively.

Our learning strategy is minimizing the regularized empirical risk with regularizer Ω weighted by λ

$$\ell(h) := \sum_{(x,y)\in\mathcal{S}} \ell(h(x), y) + \sum_{(x,\tilde{y})\in\tilde{\mathcal{S}}} \tilde{\ell}(h(x), \tilde{y}) + \lambda\Omega(h) \ , \tag{1}$$

given a *noisy training set* $\tilde{\mathcal{S}}$ containing inputs with noisy labels $(x, \tilde{y})$ and a *clean training set* $\mathcal{S}$ containing $(x, y) \sim p$. We will investigate different choices for the noise-aware pointwise loss function $\tilde{\ell}$, which may depend on (estimates of) the noise rates.

We assume that we have the possibility to correct labels. That is, we can query the true label y for $(x, \tilde{y}) \in \tilde{\mathcal{S}}$. We query single (or small batches of) labels in an iterative process. For any $(x, \tilde{y})$ for which we obtain the true label y, we remove $(x, \tilde{y})$ from $\tilde{\mathcal{S}}$ and add (x, y) to the clean training set $\mathcal{S}$. As re-labeling is assumed to be costly, we need a method for selecting the potentially most informative examples for correction.

We propose to greedily select the example(s) having the strongest expected influence on the error measure in Eq. (1). This criterion was suggested by Settles et al. (2008) in the context of multiple-instance active learning. It has the computational advantage that re-training of the model is not required for the selection process, as it is, for instance, in *expected error reduction* (Roy and McCallum, 2001). We approximate the expected model change by the difference between the gradient of the error measure before and after correcting the respective label. After we have selected an example $(x_j, \tilde{y}_j)$ and corrected its label to y_j the regularized empirical risk changes to

$$L_j(h) := \sum_{(x,y)\in\mathcal{S}\cup\{(x_j,y_j)\}} \ell(h(x), y) + \sum_{(x,\tilde{y})\in\tilde{\mathcal{S}}\setminus\{(x_j,\tilde{y}_j)\}} \tilde{\ell}(h(x), \tilde{y}) + \lambda\Omega(h) \ .$$

We assume a differentiable error measure L and define

$$g(h) := \frac{\partial L(h)}{\partial w} \quad \text{and} \quad g_j(h) := \frac{\partial L_j(h)}{\partial w}$$

as the gradients of L and L_j with respect to the model parameter w of the hypothesis h. Our approach is to pick $(x^*, \tilde{y}^*) \in \tilde{\mathcal{S}}$ maximizing

$$\begin{aligned} &\mathbb{E}_{y_j|x_j,\tilde{y}_j}\Big[\|g_j(h) - g(h)\|\Big] \\ &= \mathbb{E}_{y_j|x_j,\tilde{y}_j}\left[\left\|\frac{\partial}{\partial w}\Big(\ell(h(x_j), y_j) - \tilde{\ell}(h(x_j), \tilde{y}_j)\Big)\right\|\right] \end{aligned} \tag{2}$$

for re-labeling. Here, the Euclidean norm is a standard choice (e.g., in the work by Settles et al., 2008); other norms could be explored. Different noise-aware loss functions $\tilde{\ell}$ lead to different algorithms. The pseudocode for this algorithmic framework can be found in Figure 2f. We consider the following variants of $\tilde{\ell}$.

Noise-agnostic estimator. The simplest way to deal with label noise is to neglect it and just choose the noise-aware loss $\tilde{\ell}$ to coincide with the standard loss ℓ.

Unbiased estimator. If we know the noise rates ρ_{-1} and ρ_{+1}, we can define a loss ℓ_{u} on the noisy data as an unbiased estimator of the standard loss ℓ on the clean data, such that $\mathbb{E}_{\tilde{y}}\Big[\ell_{\mathrm{u}}(h(x), \tilde{y})\Big] = \ell(h(x), y)$ for all x, y, and h, as discussed by Natarajan et al. (2013).

However, following this approach within our framework did not lead to competitive results in our empirical evaluation. Therefore, we do not discuss the resulting algorithm in the remainder of the main paper, however, the method (referred to as *U-WURL*) and corresponding empirical results are presented in the supplementary material.

Maximum likelihood estimator. Following Bootkrajang and Kabán (2012), we can consider the true label y as a latent variable and write the posterior probability of a noisy label with class-conditional noise as

$$\begin{aligned} p(\tilde{y}|x) &= \sum_y p(\tilde{y}, y|x) = \sum_y p(\tilde{y}|y)p(y|x) \\ &= \big(1 - \rho_{\tilde{y}}\big)p(y = \tilde{y}|x) + \rho_{-\tilde{y}}\, p(y = -\tilde{y}|x) \ . \end{aligned}$$

If we assume that we can model $p(y|x)$ with the current hypothesis $h \in \mathcal{H}$, we can write the likelihood of the model parameters w given a single training example as

$$\mathcal{L}(w|x, \tilde{y}) \propto p(\tilde{y}|x, w) = \big(1-\rho_{\tilde{y}}\big)h(x)+\rho_{-\tilde{y}}\big(1-h(x)\big) \ .$$

Taking the negative log-likelihood, the noise-aware maximum likelihood loss can be defined as

$$\ell_{\mathrm{ML}}(h(x), \tilde{y}) := -\log\Big(\big(1-\rho_{\tilde{y}}\big)h(x)+\rho_{-\tilde{y}}\big(1-h(x)\big)\Big) \ . \tag{3}$$

This maximum likelihood approach coincides with the unbiased estimator only in the noiseless case.

3 ACTIVE LABEL CORRECTION WITH LOGISTIC REGRESSION

In the following, the active label correction strategies are applied to logistic regression. Logistic regression has the advantages that its output can be interpreted as a probability (allowing its use for the noise-aware maximum likelihood estimator), that the unbiased estimator is convex (Natarajan et al., 2013), and that it can be easily extended to multi-class classification and to deep neural network architectures (e.g., convolutional neural networks, see experiments in section 5). For logistic regression we have $h(x) := \sigma(x) = \frac{1}{1+\exp(-w^\top x)}$. A natural choice for the loss is the cross-entropy $\ell(h(x), y) := [y = +1] \log \frac{1}{h(x)} + [y = -1] \log \frac{1}{1-h(x)}$, where $[\cdot]$ is the Iverson bracket. Combining both components gives $\ell(h(x), y) = -\log \sigma(yx)$. To compute the expectation in the criterion Eq. (2), we re-write the probabilities using Bayes' rule and the noise model $\tilde{Y} \perp\!\!\!\perp X|Y$. Assuming that our current output $\sigma(yx)$ models the posterior probability $p(Y|X)$, we get:

$$
\begin{aligned}
&p(Y = -\tilde{y}|\tilde{Y} = \tilde{y}, X = x) \\
&\qquad = \frac{p(\tilde{Y} = \tilde{y}|Y = -\tilde{y})p(Y = -\tilde{y}|X = x)}{p(\tilde{Y} = \tilde{y}|X = x)} \\
&\qquad = \frac{\rho_{-\tilde{y}}\sigma(-\tilde{y}x)}{\rho_{-\tilde{y}}\sigma(-\tilde{y}x) + (1 - \rho_{\tilde{y}})\sigma(\tilde{y}x)}
\end{aligned}
$$

$$
\begin{aligned}
&p(Y = +\tilde{y}|\tilde{Y} = \tilde{y}, X = x) \\
&\qquad = \frac{(1 - \rho_{\tilde{y}})\sigma(\tilde{y}x)}{\rho_{-\tilde{y}}\sigma(-\tilde{y}x) + (1 - \rho_{\tilde{y}})\sigma(\tilde{y}x)}
\end{aligned}
$$

Now, we derive three novel active label correction algorithms. These pick the next example $(x^*, \tilde{y}^*)$ to be corrected based on the aforementioned loss functions.

3.1 Weighted uncertainty re-labeling (WURL)

Using standard regularized logistic regression, the gradient g becomes

$$
g(\sigma) = -\sum_{(x,y)\in\mathcal{S}} yx\sigma(-yx) - \sum_{(x,\tilde{y})\in\tilde{\mathcal{S}}} \tilde{y}x\sigma(-\tilde{y}x) + \frac{\partial}{\partial w}\Omega(w) \ . \tag{4}
$$

The gradient g_j, which measures the change rate after replacing $\tilde{y}_j$ with y_j is then

$$
\begin{aligned}
g_j(\sigma) = &- \sum_{(x,y)\in\mathcal{S}\cup\{(x_j,y_j)\}} yx\sigma(-yx) \\
&- \sum_{(x,\tilde{y})\in\tilde{\mathcal{S}}\setminus\{(x_j,\tilde{y}_j)\}} \tilde{y}x\sigma(-\tilde{y}x) + \frac{\partial}{\partial w}\Omega(w) \ .
\end{aligned} \tag{5}
$$

Inserting Eq. (4) and Eq. (5) into Eq. (2) gives

$$
\begin{aligned}
(x^*, \tilde{y}^*) &= \arg\max_{(x_j,\tilde{y}_j)\in\tilde{\mathcal{S}}} \ \mathbb{E}_{y_j|x_j,\tilde{y}_j}\Big[\|g_j(\sigma) - g(\sigma)\|\Big] \\
&= \arg\max_{(x_j,\tilde{y}_j)\in\tilde{\mathcal{S}}} \ \|x_j\| \, p(Y = -\tilde{y}_j|\tilde{Y} = \tilde{y}_j, X = x_j) \\
&= \arg\max_{(x_j,\tilde{y}_j)\in\tilde{\mathcal{S}}} \ \|x_j\| \frac{\rho_{-\tilde{y}_j}\sigma(-\tilde{y}_j x_j)}{\rho_{-\tilde{y}_j}\sigma(-\tilde{y}_j x_j) + (1 - \rho_{\tilde{y}_j})\sigma(\tilde{y}_j x_j)} \\
&:= \arg\max_{(x_j,\tilde{y}_j)\in\tilde{\mathcal{S}}} \ s_{\mathrm{W}}(x_j, \tilde{y}_j) \ ,
\end{aligned} \tag{6}
$$

where we assume that $p(y \,|\, x)$ can be replaced by $\sigma(yx)$. The criterion s_{W} suggests to pick the example which has the least confidence predicting its given label, weighted by the length of the sample vector $\|x\|$ (in the case of logistic regression; for neural networks the norm of the gradient differences involves all layers). Empirical results suggest that the bias towards input patterns with larger norm does not affect performance in practice (Settles et al., 2008; Settles and Craven, 2008). Note that we do not assume that the given model classifies the clean points perfectly. We only assume that our current model is a good predictor for $p(Y|X)$.

3.2 Robust maximum likelihood weighted uncertainty re-labeling (ML-WURL)

For logistic regression, the maximum likelihood loss in Eq. (3) takes the form as derived by Bootkrajang and Kabán (2012):

$$
\begin{aligned}
&\ell_{\mathrm{ML}}(\sigma(x), \tilde{y}) = -\log\Big((1 - \rho_{\tilde{y}})\sigma(\tilde{y}x) + \rho_{-\tilde{y}}\big(1 - \sigma(\tilde{y}x)\big)\Big) \\
&= \ell(\sigma(x), \tilde{y}) - \log\Big(1 + \rho_{-\tilde{y}}\exp(-\tilde{y}w^\top x) - \rho_{\tilde{y}}\Big)
\end{aligned} \tag{7}
$$

Unfortunately, minimizing ℓ_{ML} is not a convex problem. The form of Eq. (7), however, suggests that future work might employ DC programming by interpreting it as a difference of two convex functions (Tao, 1997). If we use ℓ_{ML} as the noise-aware loss, we get

$$
\begin{aligned}
L(\sigma) = &\sum_{(x,y)\in\mathcal{S}} \ell(\sigma(x), y) + \Omega(w) \\
&+ \sum_{(x,\tilde{y})\in\tilde{\mathcal{S}}} \Big(\ell(\sigma(x), \tilde{y}) - \log\big(1 + \rho_{-\tilde{y}}\exp(-\tilde{y}w^\top x) - \rho_{\tilde{y}}\big)\Big)
\end{aligned}
$$

with gradient

$$
\begin{aligned}
g(\sigma) = &-\sum_{(x,y)\in\mathcal{S}} yx\sigma(-yx) + \frac{\partial}{\partial w}\Omega(w) \\
&+ \sum_{(x,\tilde{y})\in\tilde{\mathcal{S}}} \tilde{y}x\Big(\frac{\rho_{-\tilde{y}}}{\rho_{-\tilde{y}} + (1 - \rho_{\tilde{y}})\exp(\tilde{y}w^\top x)} - \sigma(-\tilde{y}x)\Big)
\end{aligned}
$$

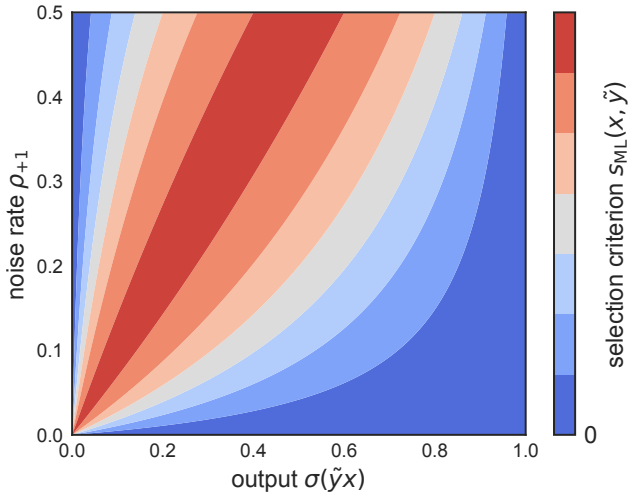


Figure 1: Behavior of the selection criterion s_{ML} in the symmetric case $\rho_{-1} = \rho_{+1}$ as a function of noise rate and classifier output for the case $\tilde{y} = +1$.

leading to:

$$\|g_j(\sigma) - g(\sigma)\| = \|x\| \cdot \left| -y\sigma(-yx) - \tilde{y}\left(\frac{\rho_{-\tilde{y}}}{\rho_{-\tilde{y}} + (1-\rho_{\tilde{y}})\exp(\tilde{y}w^\top x))} - \sigma(-\tilde{y}x) \right) \right|$$

Thus, we select the example $(x^*, \tilde{y}^*)$ for correction that maximizes

$$\begin{aligned} &\mathbb{E}_{y|x,\tilde{y}}\Big[\|g_j(\sigma) - g(\sigma)\|\Big] \\ &= 2\|x\| p(Y = -\tilde{y}|\tilde{Y} = \tilde{y}, X = x)\, p(Y = \tilde{y}|\tilde{Y} = \tilde{y}, X = x) \\ &= 2\|x\| \frac{\rho_{-\tilde{y}}\sigma(-\tilde{y}x)(1-\rho_{\tilde{y}})\sigma(\tilde{y}x)}{\Big(\rho_{-\tilde{y}}\sigma(-\tilde{y}x) + (1-\rho_{\tilde{y}})\sigma(\tilde{y}x)\Big)^2} := s_{\mathrm{ML}}(x,\tilde{y}) \ . \end{aligned}$$

Figure 1 shows how the selection criterion s_{ML} changes for the case of symmetric label noise $\rho = \rho_{-1} = \rho_{+1}$ as a function of the noise rate ρ and the output of the classifier $\sigma(\tilde{y}x)$ for $\tilde{y} = +1$ and $\|x\| = 1$. If $\rho = 0.5$, the observed label does not provide any information and s_{ML} reduces to uncertainty sampling. That is, we are in the standard active learning scenario and sample close to the decision boundary ($\sigma(x) = 0.5$). If the noise rate is zero, there is no preference over the examples as re-labeling does not change anything (note that this is different from the standard active learning scenario). Between these extremes, the criterion depends on how strongly noisy observation and current model prediction deviate. It is high if the observed label $\tilde{y} = +1$ and the prediction by the model disagree (i.e., $\sigma(\tilde{y}x) < 0.5$) and low if they observation and prediction agree. The higher the noise rate, the less pronounced this effect until it vanishes for $\rho = 0.5$.

4 ESTIMATING NOISE RATES

The noise-aware selection criteria assume that the noise rates are known. In practice, it is unlikely that they are. In this case we can draw an initial sample uniformly at random and estimate the noise rates by counting the number of corrected labels:

$$\hat{\rho}_k = \frac{\sum_{(x,y,\tilde{y})\in\mathcal{S}_C} [\tilde{y} = -k, y = k]}{\sum_{(x,y,\tilde{y})\in\mathcal{S}_C} [y = k]} \ , \tag{8}$$

where $[\cdot]$ is the Iverson bracket, and $\mathcal{S}_C$ is the set of all corrected examples $(x, y, \tilde{y})$ with their noisy and true labels. In case $\mathcal{S}_C$ is drawn uniformly at random, $\hat{\rho}_k$ is an unbiased estimator of the corresponding true noise rate. This approach has the drawback that in this initial phase the active learning algorithm does not yet exploit the gathered information about the noise rates.

Importance-weighting. Thus, we want to simultaneously estimate the noise model parameters while using the current noise model for active learning. Drawing examples actively (i.e., non-uniformly) has the drawback that Eq. (8) becomes biased. One way of dealing with this bias is importance-weighting. Instead of deterministically picking examples that maximize our criterion, in each iteration t, we define a sampling probability distribution $p_s(x, \tilde{y}, t)$ over the noisy sample $\tilde{\mathcal{S}}$. This distribution assigns examples with a higher score in the selection criterion a higher probability of being picked. When an example is chosen, it is given an importance weight, defined as the inverse of its sampling probability p_s. To avoid infinite importance weights, we also have to make sure that the sampling probability is bounded away from zero by adding a minimum probability $p_{\min} \leq \frac{1}{n}$. Thus, similar to Ganti and Gray (2012), we can define our sampling probability distribution as:

$$p_s(x,\tilde{y},t) := p_{\min}(t) + (1 - n \cdot p_{\min}(t)) \frac{s(x,\tilde{y})}{\sum_{(x,\tilde{y})\in\mathcal{S}\cup\tilde{\mathcal{S}}} s(x,\tilde{y})} \tag{9}$$

Here, $n = |\mathcal{S} \cup \tilde{\mathcal{S}}|$ and $s(x, \tilde{y})$ is a non-negative selection criterion, for example one of the criteria s_{W} and s_{ML} we introduced above. If $s(x, \tilde{y}) = 0$ for all $(x, \tilde{y})$, we just sample uniformly at random by setting $p_s(x, \tilde{y}, t) = \frac{1}{n}$. Following Ganti and Gray (2012), we define the minimum probability $p_{\min}(t) = \frac{1}{nt^\kappa}$, where κ is a hyperparameter that tunes the trade-off between exploiting the criterion and exploring new examples for noise rate estimation. Note that in order to draw each example independently, we draw with replacement. Therefore, it is possible that an example is selected multiple times. In this case, we just re-use its previously corrected example at no cost.

To account for the non-uniform selection, we can estimate the noise rates in iteration t by

$$\hat{\rho}_k(t) = \frac{\sum_{\tau=1}^{t} w(x_\tau, \tilde{y}_\tau, \tau)[\tilde{y}_\tau = -k, y_\tau = k]}{\sum_{\tau=1}^{t} w(x_\tau, \tilde{y}_\tau, \tau)[y_\tau = k]} \ .$$

Here, $k \in \{-1, 1\}$ is the label of interest. The corrected example in round τ and its labels are denoted by $(x_\tau, \tilde{y}_\tau, y_\tau)$, and we define $w(x, \tilde{y}, \tau) := \frac{1}{p_s(x,\tilde{y},\tau)}$. Although $\hat{\rho}_k(t)$ is not an unbiased estimator of the true unknown noise rate either, its bias vanishes for $t \to \infty$.

By employing importance weights we are able to simultaneously estimate the noise rates while maximizing the accuracy through active learning. In order to avoid an unstable start-up phase, it is possible to integrate a prior probability. This prior can be informed by methods that estimate noise rates from noisy samples only (e.g., Liu and Tao, 2016; Menon et al., 2015).

Maximum likelihood estimation. Instead of estimating the noise rates through importance sampling, we can also maximize the likelihood with respect to the noise rate parameters of the model, analogous to the model weights. Bootkrajang and Kabán (2012) showed that in this case the noise rates can be updated using

$$\hat{\rho}_k^{\mathrm{ML}}(t) \leftarrow \frac{\mu_k^t}{\mu_k^t + \delta_k^t} \ , \tag{10}$$

where

$$\mu_k^t = \sum_{(x,\tilde{y})\in\tilde{\mathcal{S}}} \frac{[\tilde{y} = -k]\, \hat{\rho}_k^{\mathrm{ML}}(t)\, \sigma(kx)}{\hat{\rho}_k^{\mathrm{ML}}(t)\, \sigma(kx) + \left(1 - \hat{\rho}_{-k}^{\mathrm{ML}}(t)\right) \sigma(-kx)}$$

$$\delta_k^t = \sum_{(x,\tilde{y})\in\tilde{\mathcal{S}}} \frac{[\tilde{y} = k]\, \left(1 - \hat{\rho}_k^{\mathrm{ML}}(t)\right) \sigma(kx)}{\left(1 - \hat{\rho}_k^{\mathrm{ML}}(t)\right) \sigma(kx) + \hat{\rho}_{-k}^{\mathrm{ML}}(t)\, \sigma(-kx)}$$

Then we can alternate the optimization of the model weights and noise parameters until convergence similar to expectation maximization. Therefore, we call this variant *EM-WURL*. The advantage of this algorithm is that we do not need any importance-weighting. The optimization problem is non-convex, and we initialize the noise rate estimates randomly from a uniform distribution between 0 and 0.5.

Maximum a posteriori estimation. However, EM-WURL does not make any use of the information we gained about the noise rates by re-labeling a sub-sample of the dataset. Can we combine the sample estimates we use in ML-WURL and the maximum likelihood estimates in EM-WURL? A natural way to combine both sources of information is by treating the sample estimates as parameters of a prior distribution. By assuming that this prior follows a beta distribution, whose parameters can be interpreted as virtual draws from the underlying distribution, we can obtain the maximum a posteriori estimate. Formally, in each round of the re-labeling, we assume the following prior distributions

$$\rho_k \sim \mathrm{Beta}\big(\rho_k | \alpha_k^t, \beta_k^t\big) \ ,$$

where we set the parameters of the distribution to be

$$\alpha_k^t = |\mathcal{S}|\, \hat{\rho}_k(t)$$
$$\beta_k^t = |\mathcal{S}|\, \left(1 - \hat{\rho}_k(t)\right) \ ,$$

which are the estimated numbers of corrected (α_k^t) and confirmed (β_k^t) examples if a random sample had been drawn. If we use this prior distribution, then the maximum a posteriori estimates for the noise rates can be obtained by

$$\hat{\rho}_k^{\mathrm{MAP}}(t) \leftarrow \frac{\alpha_k^t + \mu_k^t - 1}{\alpha_k^t + \beta_k^t + \mu_k^t + \delta_k^t - 2} \ . \tag{11}$$

If we use Eq. (11), we refer to the algorithm as *maximum a posteriori WURL (MAP-WURL)*. As in the case of EM-WURL, we alternate between updating the estimated noise rates and the model weights until convergence.

5 EXPERIMENTS

We started by evaluating the algorithms for label correction on data sets where we injected class-conditional noise. Then we applied these methods to an image classification task using a convolutional neural network.[2]

We randomly sampled training sets of 2000 patterns from different benchmark datasets and flipped their labels with probabilities $\rho_{-1} \in \{0.2, 0.3\}$ and $\rho_{+1} = 0.1$. We evaluated the accuracies of the classifiers on separate test sets of 5000 samples (in the case of the dataset 'ad' we used 359 samples as more data are not available). We averaged each experimental outcome over 30 trials.

The result achieved by training the predictive model on the full training set without label noise is called the *clean baseline*. It indicates the performance limit achieved by correcting the whole dataset. For the active label correction, we started with 0 corrected examples ($\mathcal{S} = \emptyset$) and stop when half of the training set is corrected ($|\mathcal{S}| = 1000$). For all experiments we set the trade-off parameter $\kappa = 0.5$, as suggested in Theorem 3 by Ganti and Gray (2012) for the squared loss. To start with a stable estimated noise rate for ML-WURL, we employ a burn-in phase of sampling $n_{\text{burn-in}} = 50$ examples uniformly at random.

[2]The code reproducing all results is available at https://github.com/kremerj/relabeling.

The algorithms we devised in this paper are referred to as *weighted uncertainty re-labeling (WURL)* and *robust ML weighted uncertainty re-labeling (ML-WURL)*, see section 3.1 and 3.2, respectively. If we determine the noise rate parameters by maximum likelihood, we refer to the algorithm as *EM-WURL*, if we combine maximum likelihood estimate and the importance-sampled estimates, we refer to the algorithm as *MAP-WURL*, see section 4. For comparison, we consider three baselines. The first is *random sampling* for the standard loss, that is, choosing an example to correct uniformly at random. The second is the algorithm presented by Rebbapragada et al. (2012), referred to as *uncertainty re-labeling*, which amounts to selecting the example with the closest absolute distance to the decision hyperplane and then training using Eq. (1) with $\ell = \tilde{\ell}$. Furthermore, we evaluated another simple *reduction*, in which we perform uncertainty re-labeling, but only train on the clean data. This corresponds to simply applying a standard active learning algorithm to our re-labeling task, namely *uncertainty sampling*. Uncertainty sampling, although simple, gives state-of-the-art performance in the standard active learning scenario (Yang and Loog, 2016).

Logistic regression. The algorithms are evaluated on the binary classification benchmark datasets 'a1a', 'ad', 'covtype', 'w1a', 'mushrooms', 'cod-rna' and 'ijcnn1' from the LIBSVM data repository. We evaluated each classifier after correcting one additional example. We tested ℓ_2-regularization $\Omega(h) = \frac{1}{2}\|x\|^2$ with $\lambda \in \{1, 10\}$. Each loss is optimized using L-BFGS and each iteration is warm-started.

Figure 2 shows selected empirical results. Additional experiments on the remaining datasets and with different parameter settings can be found in the supplementary material. We can see that on all shown datasets the robust algorithms were always among the fastest to reach the clean baseline and never failed catastrophically. Although the simple uncertainty sampling dominated on some datasets that are "easy" in the sense that the clean baseline is above 0.9 accuracy, it failed in comparison to the robust methods if the clean baseline accuracy was lower (see supplementary material).

Experiments with realistic noise: Clothing data. We argue that modeling the class-conditional part of real noise is sufficient to improve selection of the next example(s) for correction in practice. This is demonstrated in the following experiment.

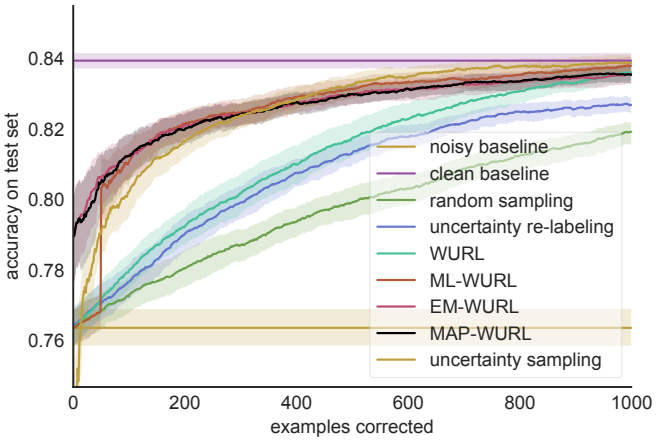
We considered a real-world data set provided by Xiao et al. (2015), who obtained images of clothes from several online shopping websites. A label indicating the type of clothing was automatically obtained by analyzing the text surrounding the images. A subset of these noisy labels was manually corrected afterwards. In our experiments, shown in Figure 2e, we considered dresses and vests because these had the largest difference in noise rates (0.37 and 0.05).

We employed a pre-trained 'AlexNet' (Krizhevsky et al., 2012) implemented using TensorFlow (Abadi et al., 2016). We selected 2048 examples for training and 2000 examples for testing, which we picked randomly in 10 trials, and report the mean performance. We only trained the final fully connected layer and keep the other layers fixed. We trained the network using stochastic gradient descent with weight decay parameter of 0.0005, momentum parameter of 0.9, and learning rate of 0.01 for 100 epochs using a mini-batch size of 64. In contrast to the logistic regression experiments, we greedily selected a full batch of 64 examples for re-labeling in each step. The initial batch was used for burn-in.

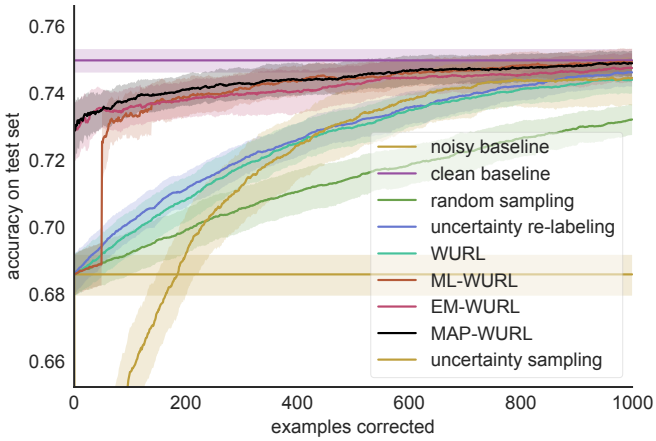
The CNN experiment showed that ML- and MAP-WURL can outperform alternative algorithms even in settings where the class-conditional noise assumption is not a perfect match. Uncertainty sampling performed on par. The EM-WURL algorithm worked less well, most likely because the label noise parameters are more difficult to optimize in the stochastic gradient setting. Adapting the update to stochastic gradient descent is promising future work.

6 CONCLUSION

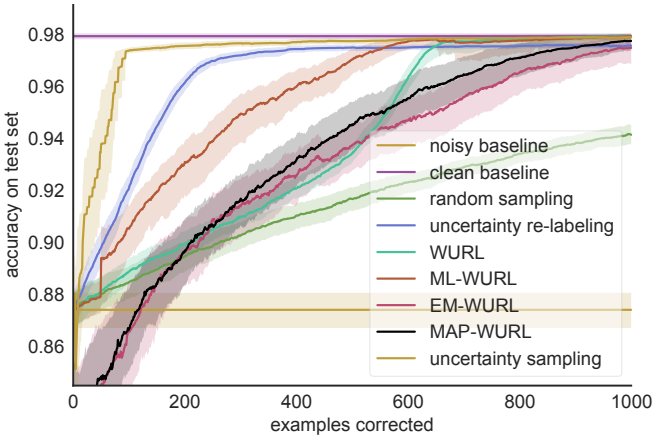
We presented a principled approach to active label correction. We propose to employ loss functions that depend on a noise model and to apply the maximum expected model change criterion to the corresponding regularized risk functionals. Class-conditional noise was assumed as a model for the true noise. We demonstrated how to adapt the parameters of the noise model during learning. Different loss functions were considered and corresponding algorithms were derived. On datasets where training on true labels achieves an accuracy below 0.9 the robust approaches ML-, EM- and MAP-WURL were among the best performing algorithms and in most cases beat uncertainty sampling (i.e., standard active learning). Here, MAP-WURL gave good results regardless of how many re-labelings were provided, while in our deep learning application both ML- and MAP-WURL were among the best. The CNN experiment demonstrated (in accordance with Patrini et al., 2017) that the class-conditional noise model, which can be learned efficiently, can guide re-labeling in real-world applications.



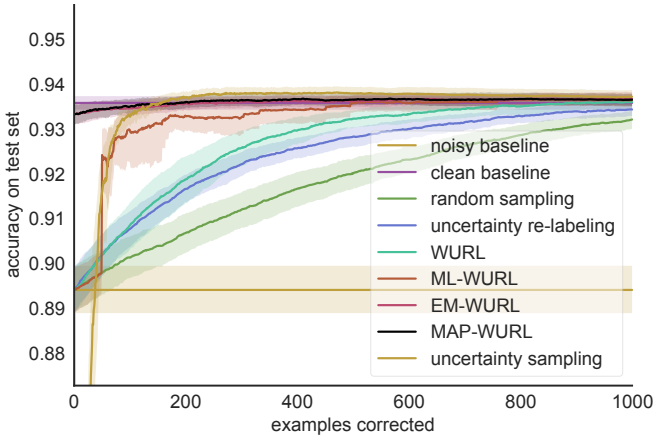
(a) Logistic regression classifying dataset 'a1a'.



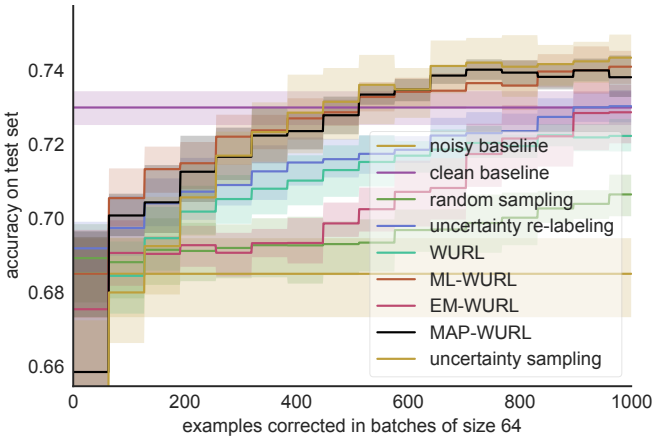
(b) Logistic regression classifying dataset 'covtype'.



(c) Logistic regression classifying dataset 'w1a'.



(d) Logistic regression classifying dataset 'cod-rna'.



(e) CNN classifying 'vest' vs. 'dress'.

Input:
noisy data $\tilde{\mathcal{S}}$, clean data $\mathcal{S} = \emptyset$
rounds t, loss L, selection criterion s
for $\tau = 1$ **to** t **do**
  Estimate noise rates $\hat{\rho}$ from $\mathcal{S}$, $\tilde{\mathcal{S}}$ and optionally, current model parameters $\hat{w}$
  Optimize $\hat{w}, \hat{\rho} = \arg\max_{w,\rho} L(w, \hat{\rho}, \mathcal{S}, \tilde{\mathcal{S}}, \tau)$
  Update $p_s(x, \tilde{y}, \tau)$
  Correct $(x^*, \tilde{y}^*)$ drawn from p_s and add to $\mathcal{S}$
end for
Output: clean data $\mathcal{S}$

(f) Pseudocode of the algorithmic framework.

Figure 2: Empirical results on the benchmark datasets (a-e) evaluating the algorithmic framework (f). Shown are the mean test set errors and bootstrapped 99% confidence intervals over 30 trials for logistic regression ($\lambda = 1.0, \kappa = 0.5, n_{\text{burn-in}} = 50$) and over 10 trials for the CNN ($\kappa = 0.5, n_{\text{burn-in}} = 64$). The noise rates in the logistic regression experiments were $\rho_{-1} = 0.3$, $\rho_{+1} = 0.1$. Experiments on further datasets and for different parameter settings can be found in the supplementary material. That some algorithms performed better than the clean baseline is a random artifact; when training was continued, these finally matched the baseline.

Acknowledgements

The Titan Xp used for this research was donated by the NVIDIA Corporation. JK and CI acknowledge support from the Innovation Fund Denmark through the *Danish Center for Big Data Analytics Driven Innovation* (DABAI).

References

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. TensorFlow: A system for large-scale machine learning. In *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016.

D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2(4), 1988.

J. Bootkrajang and A. Kabán. Label-noise robust logistic regression and its applications. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*. Springer, 2012.

B. Frenay and M. Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5), 2014.

R. Ganti and A. G. Gray. UPAL: Unbiased pool based active learning. *JMLR W&P (AISTATS)*, 22, 2012.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

T. Liu and D. Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3), 2016.

A. K. Menon, B. van Rooyen, C. S. Ong, and B. Williamson. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning (ICML)*, 2015.

N. Natarajan, I. S. Dhillon, P. D. Ravikumar, and A. Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

G. Patrini, A. Rozza, A. Menon, R. Nock, and L. Qu. Making deep neural networks robust to label noise: a loss correction approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

U. Rebbapragada, C. E. Brodley, D. Sulla-Menashe, and M. A. Friedl. Active label correction. In *IEEE International Conference on Data Mining (ICDM)*, 2012.

S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *International Conference on Learning Representations (ICLR) Workshop*, 2015.

N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *International Conference on Machine Learning (ICML)*, 2001.

B. Settles and M. Craven. An analysis of active learning strategies for sequence labeling tasks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL, 2008.

B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Iternational Conference on Knowledge Discovery and Data Mining (SIGKDD)*. ACM, 2008.

S. Sukhbaatar and R. Fergus. Learning from noisy labels with deep neural networks. *International Conference on Learning Representations (ICLR) Workshop*, 2015.

P. D. Tao. Convex analysis approach to d. c. programming: Theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1), 1997.

R. Urner, S. Ben-David, and O. Shamir. Learning from weak teachers. *JMLR W&P (AISTATS)*, 22, 2012.

T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

Y. Yang and M. Loog. A benchmark and comparison of active learning for logistic regression. *arXiv:1611.08618 [stat.ML]*, 2016.

C. Zhang and K. Chaudhuri. Active learning from weak and strong labelers. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

L. Zhao, G. Sukthankar, and R. Sukthankar. Incremental relabeling for active learning with noisy crowdsourced annotations. In *IEEE International Conference on Social Computing (SocialCom)*. IEEE, 2011.