Learning and Evaluation in Presence of Non-i.i.d. Label Noise

Nico Görnitz[1]
Machine Learning Laboratory
Berlin Institute of Technology
Berlin, Germany

Anne K. Porbadnigk[1]
Machine Learning Laboratory
Berlin Institute of Technology
Berlin, Germany

Alexander Binder
Machine Learning Laboratory
Berlin Institute of Technology
Berlin, Germany

Claudia Sannelli
Machine Learning Laboratory
Berlin Institute of Technology
Berlin, Germany

Mikio Braun
Machine Learning Laboratory
Berlin Institute of Technology
Berlin, Germany

Klaus-Robert Müller[2]
Korea University, Seoul, Korea
Machine Learning Laboratory
Berlin Institute of Technology
Berlin, Germany

Marius Kloft[2]
Courant Institute of Mathematical Sciences
Memorial Sloan-Kettering Cancer Center
New York, New York, USA

Abstract

In many real-world applications, the simplified assumption of independent and identically distributed noise breaks down, and labels can have structured, systematic noise. For example, in brain-computer interface applications, training data is often the result of lengthy experimental sessions, where the attention levels of participants can change over the course of the experiment. In such application cases, structured label noise will cause problems because most machine learning methods assume independent and identically distributed label noise. In this paper, we present a novel methodology for learning and evaluation in presence of systematic label noise. The core of which is a novel extension of support vector data description / one-class SVM that can incorporate latent variables. Controlled simulations on synthetic data and a real-world EEG experiment with 20 subjects from the domain of brain-computer-interfacing show that our method achieves accuracies that go beyond the state of the art.

1 Introduction

Most supervised learning algorithms assume that the noise in the labels or output variables (y_i) is independent and identically distributed:

$$y_i = f(x_i) + \epsilon_i\,, \quad \epsilon_1, \ldots, \epsilon_n \overset{\text{i.i.d.}}{\sim} P$$

This assumption is convenient from the viewpoint of theory, as it directly links to the law of large numbers, which explains why learning methods such as support vector machines [1] consistently improve when training set sizes grow, despite label noise. However, in many real-world applications, the assumption of i.i.d. noise is questionable at best because training data can be the result of complex experiments with a time-dependent noise structure and a human in the loop.

For instance, a typical challenge in the context of electroencephalography-based brain-computer interfacing (EEG-BCI; e.g., [2]) is the decoding of mental states [3]. Machine learning has become indispensable in this regard [4], in particular for analyzing the threshold of perception ([5], e.g.). For such applications, EEG signals are recorded in experimental sessions that span several hours, during which the participants are asked to perform complex mental tasks. However, if participants become distracted, bored, or sleepy, this may result in a significant increase of mislabeled training examples in consecutive trials [6]. Thus—from a statistical point of view—the noise level increases, the distribution of the noise changes, and subsequent trials exhibit dependency structures.

What are the implications of such a *systematic* label noise in machine-learning applications? Both the training and the testing phase in machine learning can suffer from non-independent and non-identically distributed noise:

1. common training algorithms fit a noise term to the data, the parameters of which are shared by all training examples, while, in reality, the parameters of the noise may change
2. common procedures for estimating the test error (such as cross validation) rely on accurate test labels—an assertion that renders a fair evaluation challenging.

In this paper, we propose a novel methodology for both learning in presence of systematic label noise and for reliable evaluation of the results. Since we may neither com-

[1] Authors contributed equally.
[2] Corresponding Authors.

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

pletely trust the training nor the test labels, the core of the methodology consists of a new *unsupervised* learning algorithm capable of encoding the *state* of the noise by a latent variable.

The contributions of this paper can be summarized as follows: we propose a new methodology for learning and evaluation in presence of non-i.i.d. label noise, at the core of which lies a novel unsupervised learning method—LATENTSVDD—that is formulated in terms of a DC optimization problem. We give a dual representation of the optimization based on DC Fenchel duality theory and present a DC-programming algorithm for the problem, for which we prove that it locally converges. We show an upper bound on the generalization error of the latter method that converges to zero at the usual convergence rate $O(\sqrt{1/n})$. An empirical analysis of the methodology on synthetic data is presented. Finally, we provide an extensive case study of a real application scenario from the domain of brain-computer interfacing, where LATENTSVDD allows us to re-assess a common test of visual attention. Our analysis shows that even in the difficult scenario of learning in presence of non-i.i.d. label noise, learning and reasonable evaluation can indeed be made possible.

2 Learning Methodology

We are given a data set $\mathcal{D}$ consisting of N data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, lying in some input space $\mathcal{X}$, and labels $y_1, \ldots, y_N \in \mathcal{Y}$. As mentioned in the introduction, we consider a learning scenario where we have varying confidence in the labels (some y_i are more trustworthy than others). To this end, we propose a methodology for learning with non-i.i.d. label noise that consists of the following steps.

Step 1: COMPUTATION OF THE LATENT STATE AND ANOMALY SCORE FOR EACH DATA POINT

Step 2: SANITIZATION: REMOVAL OF THE MOST NOISY DATA POINTS

Step 3: LABEL-DENOISING

Step 4: EVALUATION

As a result of the above steps we obtain a learning methodology that outputs, for a training set $\mathcal{D}$, an inductive rule

$$g_{\mathcal{D}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}\,,$$

that lets us assign to any pair $(\boldsymbol{x}, y)$ a denoised label $\widehat{y} := g_{\mathcal{D}}(y)$, which is our guess for the true underlying label.

The various steps of the above methodology are detailed below.

2.1 Step 1: Latent Support Vector Data Description (LATENTSVDD)

Our approach is based on the paradigms of support vector learning [7] and density level set estimation [8, 9]. Here the data is mapped from the input space into a RKHS feature space $\phi\colon \mathcal{X} \to \mathcal{F}$ that gives rise to a kernel k [10, 11]. The goal is to find a model $f\colon \mathcal{X} \to \mathbb{R}$ and a density level-set $L := \{\boldsymbol{x}\colon f(\boldsymbol{x}) \leq \rho\}$ containing most of the regular data points, while for anomalies and outliers $\boldsymbol{x} \notin L$ holds. In case of the support vector data description (SVDD) method, $f_{\text{SVDD}}(\boldsymbol{x}) = \|\boldsymbol{c} - \phi(\boldsymbol{x})\|^2$ and parameter estimation corresponds to solving a quadratically constrained quadratic program (QCQP):

$$\begin{aligned} &\min_{R,\boldsymbol{c},\boldsymbol{\xi}\geq 0} \; R^2 + C\sum_{i=1}^{n}\xi_i && \text{(SVDD)}\\ &\text{s.t.} \quad \|\boldsymbol{c} - \phi(\boldsymbol{x}_i)\|^2 \leq R^2 + \xi_i \quad \forall\, i \end{aligned}$$

That allows for the following simple geometric interpretation: a ball of radius R is computed that comprises most the regular data points, while all points lying outside of the normality radius are declared being anomalous.

In this paper, we extend the classical mapping f_{SVDD} by the inclusion of a latent variable $\boldsymbol{z} \in \mathcal{Z}$ in a joint feature map $\Psi : \mathcal{X} \times \mathcal{Z} \to \mathcal{F}$. As a consequence, the resulting model

$$f : \mathcal{X} \to \mathbb{R},\; \boldsymbol{x} \mapsto \min_{\boldsymbol{z}\in\mathcal{Z}} \|\boldsymbol{c} - \Psi(\boldsymbol{x}, \boldsymbol{z})\|^2 \qquad (1)$$

becomes more expressive (a similar idea appeared also recently in the context of supervised learning [12, 13]). The latent state variable $\widehat{\boldsymbol{z}}$ of a given data point $\boldsymbol{x}$ can be inferred by $\widehat{\boldsymbol{z}} = \operatorname{argmin}_{\boldsymbol{z}\in\mathcal{Z}} \|\Psi(\boldsymbol{x}, \boldsymbol{z})\|^2 - 2\langle \boldsymbol{c}, \Psi(\boldsymbol{x}, \boldsymbol{z})\rangle$. The idea in the context of non-i.i.d. label noise is: if the structure of the latent space $\mathcal{Z}$ resembles the true underlying label structure, then we are able to *infer* the true labels by taking a purely data driven approach. The extended model, which we call LATENTSVDD, leads to a modified optimization problem:

$$\begin{aligned} &\min_{R,\boldsymbol{c},\boldsymbol{\xi}\geq 0} \; R^2 + C\sum_{i=1}^{n}\xi_i && \text{(LATENTSVDD)}\\ &\text{s.t.} \quad \min_{\boldsymbol{z}\in\mathcal{Z}} \|\boldsymbol{c} - \Psi(\boldsymbol{x}_i, \boldsymbol{z})\|^2 \leq R^2 + \xi_i \quad \forall\, i\,. \end{aligned}$$

Because of the $\min$ operator in the constraints, the resulting optimization problem is no longer convex, but we can derive an optimization strategy by decomposing the problem into convex and concave parts and iteratively linearizing the concave part (*DC Programming* [14, 15]). In order to do so, we re-write the above problem in an equivalent, unconstrained fashion as follows:

$$\min_{\boldsymbol{c},R} \; R^2 + C\sum_{i=1}^{n} \max(0, \min_{\boldsymbol{z}\in\mathcal{Z}} \|\boldsymbol{c} - \Psi(\boldsymbol{x}_i, \boldsymbol{z})\|^2 - R^2).$$

Substituting $\Omega := R^2 - \|\boldsymbol{c}\|^2$, this is equivalent to

$$\min_{\boldsymbol{c},\Omega} \|\boldsymbol{c}\|^2 + \Omega + C\sum_{i=1}^{n} \max\Big(0, -\Omega + \min_{\boldsymbol{z}\in\mathcal{Z}} \|\Psi(\boldsymbol{x}_i, \boldsymbol{z})\|^2 - 2\langle \boldsymbol{c}, \Psi(\boldsymbol{x}_i, \boldsymbol{z})\rangle\Big)$$

subject to the constraint $\|\boldsymbol{c}\|^2 + \Omega \geq 0$, which can be dropped as it is not active in the optimal point. Note that, for any i, the function

$$g_i(\boldsymbol{c}, \Omega) := -\Omega + \min_{\boldsymbol{z} \in \mathcal{Z}} \|\Psi(\boldsymbol{x}_i, \boldsymbol{z})\|^2 - 2\langle \boldsymbol{c}, \Psi(\boldsymbol{x}_i, \boldsymbol{z})\rangle \quad (2)$$

is concave, so $-g_i$ is convex. Furthermore, note that for any $t \in \mathbb{R} : \max(0, t) = \max(0, -t) + t$. Thus we have the decomposition

$$\max(0, g_i(\boldsymbol{c}, \Omega)) \;=\; \underbrace{\max(0, -g_i(\boldsymbol{c}, \Omega))}_{\text{convex}} \;+\; \underbrace{g_i(\boldsymbol{c}, \Omega)}_{\text{concave}},$$

because the maximum of two convex functions is convex. Thus we can equivalently re-write the LATENTSVDD optimization problem as a sum of a convex and a concave function as follows: given the definition of g_i in Eq. (2), solve

(LATENTSVDD-DC)

$$\min_{\boldsymbol{c}, \Omega} \underbrace{\|\boldsymbol{c}\|^2 + \Omega + C \sum_{i=1}^{n} \max(0, -g_i(\boldsymbol{c}, \Omega))}_{\text{convex}} + \underbrace{C \sum_{i=1}^{n} g_i(\boldsymbol{c}, \Omega)}_{\text{concave}}$$

The above problem is an instance of the class of DC optimization problems. We propose to solve the above problem with the simplified DC algorithm [14]. That is, alternatingly, the concave part is linearized and the resulting approximate problem solved. The resulting algorithm is shown in Algorithm Table 1

Algorithm 1 Optimization Algorithm for LATENTSVDD

input data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$
initialize $\boldsymbol{c}^{t=0}$ & $\forall i : \widehat{z}_i^{t=0}$ (e.g., randomly)
repeat
  t:=t+1
  for $i = 1, \ldots, N$ **do**
    $\widehat{\boldsymbol{z}}_i^t := \operatorname{argmin}_{\boldsymbol{z} \in \mathcal{Z}} ||\boldsymbol{c}^{t-1} - \Psi(\boldsymbol{x}_i, \boldsymbol{z})||^2$
    overwriting the notation of g_i in (2), we define
    $g_i(\boldsymbol{c}, \Omega) := -\Omega + ||\Psi(\boldsymbol{x}_i, \widehat{\boldsymbol{z}}_i^t)||^2 - 2\langle \boldsymbol{c}, \Psi(\boldsymbol{x}_i, \widehat{\boldsymbol{z}}^t)\rangle$
  end for
  let $\boldsymbol{c}^t$ and Ω^t the optimal arguments when solving Problem (LATENTSVDD-DC) with the g_i set as above
until $\forall\, i : \; \widehat{\boldsymbol{z}}_i^t := \widehat{\boldsymbol{z}}_i^{t-1}$
return optimal model parameters $\boldsymbol{c} := \boldsymbol{c}^t$, $R := \sqrt{||\boldsymbol{c}^t||^2 + \Omega^t}$, and $\boldsymbol{z}_i := \widehat{\boldsymbol{z}}_i^t \quad \forall\, i = 1, \ldots, N$

The proposed algorithm converges against a local optimum (typically in about 10 iterations, as we found in our experiments). This follows from the following theorem that is taken from [16], which is an extension of the convergence theorem in [14] to non-differentiable objective functions.

Theorem 1 ([16], Theorem 3.3)**.** *Let f, g be convex functions. Let x_0 be any feasible point, and put*

$$\forall t > 0 : \;\; x_t := \operatorname*{argmin}_x f(x) - x^\top \nabla g(x_{t-1}) \,.$$

If the non-smooth parts of f and g are piecewise-linear and the smooth part of f is strictly convex quadratic, then any limit point of the sequence (x_t) is a stationary point.

The proposed algorithm also admits a dual representation via the convex conjugate function $f^*(x) := \sup_y \langle x, y\rangle - f(x)$. The dual of the LATENTSVDD-DC problem is given by

$$\min_{\boldsymbol{c}, \Omega} \left(-C \sum_{i=1}^{n} g_i(\boldsymbol{c}, \Omega) \right)^* - \left(\|\boldsymbol{c}\|^2 + \Omega + C \sum_{i=1}^{n} \max(0, -g_i(\boldsymbol{c}, \Omega)) \right)^* .$$

This completes the presentation of the first step in our proposed methodology. We now turn to step 2.

2.2 Step 2: Outlier Removal

To remove outliers [17], we divide the data set $\mathcal{D}$ into two disjoint sets $L_- := \{\boldsymbol{x} \colon f(\boldsymbol{x}) \leq \rho\}$, containing most of the regular data, and $L_+ := \{\boldsymbol{x} \colon f(\boldsymbol{x}) > \rho\}$, consisting of the anomalies. Here f is defined as in (1). LATENTSVDD provides us with a natural choice of a threshold $\rho = R^2$, but usually we employ a small and thus conservative radius $R << \|\psi(\boldsymbol{x}, \boldsymbol{z})\|_\infty$, so that choosing $\rho = R^2$ would be too aggressive (too many anomalies removed). As a remedy, we apply the following procedure to determine a good threshold ρ. Set $f_i := f(x_i)$ and arrange the f_i in non-decreasing order, $f_{(1)} \leq \ldots \leq f_{(n)}$. Put

$$\rho := \max\left(R^2, \max_{i=1,\ldots,N-1} f_{(i+1)} - f_{(i)}\right).$$

Thus intuitively we determine the threshold where the anomaly score $f(\boldsymbol{x})$ has the steepest slope. The motivation of which is that regular data is quite densely sampled and thus has a rather smooth increase of anomaly scores, so that choosing an area with steep slope of anomaly scores corresponds to an anomalous region in input space. Indeed we have observed that this heuristic often leads to good results in practice. Finally we output $\mathcal{W} := L_-$ as our (sanitized) working training set.

2.3 Step 3: Label Assignment

In this step, we aim at assigning a label $\widehat{\boldsymbol{y}}$ for each data point $\boldsymbol{x}$ using the information from the latent variable $\boldsymbol{z} \in \mathcal{Z}$, as computed by LATENTSVDD. We start by partitioning the working data set $\mathcal{W}$ is into m smaller sets $\mathcal{W}_1, \ldots, \mathcal{W}_m$, where $m := |\mathcal{Z}|$ denotes the cardinality of the latent state space, by grouping all data points that have the same latent state in the LATENTSVDD model.

Then, we wish to flip the labels of data points such that the data within each group $\mathcal{W}_i$ has identical labels. To this end, we could simply perform a majority vote within each group. We follow a different, more sophisticated approach

here: we determine each group's joint label by choosing the labels such that the working set's kernel-target-alignment (KTA) score is maximized after label assignment.

Kernel target alignment (KTA) [18] is a method that measures the fit between the Gram matrix $K = (\langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle)_{1 \leq, i,j \leq n}$ and the label vector $\boldsymbol{y} = (y_1, \ldots, y_n)$ as follows:

$$\mathrm{KTA}(K, \boldsymbol{y}) = \frac{\langle K, \boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} \rangle_F}{\|K\|_F \, \|\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\|_F}$$

Here, $\langle A, B \rangle_F := \sum_{i,j=1}^{n} a_{ij} b_{ij}$ denotes the Frobenius inner product and $\|A\|_F := \langle A, A \rangle_F^{1/2}$ denotes its induced norm. This measure has been utilized for optimizing kernels or feature representations [18, 19]. In this paper, we reverse the perspective: instead of optimizing a kernel to match the labels, we optimize the labels to match the kernel.

Let $\mathcal{W} = \mathcal{W}_1 \cup \ldots \cup \mathcal{W}_m$ be the partition of the working training set $\mathcal{W}$ into disjoint sets $\mathcal{W}_i$ such that examples having the same latent state are grouped within the same $\mathcal{W}_i$. Then we compute the denoised label vector $\widehat{\boldsymbol{y}}$ as

$$\widehat{\boldsymbol{y}} := \operatorname*{argmax}_{\boldsymbol{y} \in \{+1,-1\}^N} \mathrm{KTA}(K, \boldsymbol{y}) \quad \text{s.t.} \quad \forall i, j, k : \boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{W}_k \Rightarrow y_i = y_j.$$

Here, the constraints require that all data points within a group $\mathcal{W}_i$ are assigned with the same label. This ensures that we only have to optimize over a few possible label combinations, e.g., over $2^5 = 32$ instead of 2^N, if we have $m = 5$ groups. This renders the optimization problem feasible.

2.4 Step 4: Evaluation

Fair evaluation of learning algorithms for label denoising is a major challenge: while we cannot trust the observed labels, we usually cannot access the underlying ground truth of an experiment. One approach to circumvent this problem is perform the evaluation solely on controlled synthetic data, where we can access the truly underlying performance of the algorithms. In this paper, we perform a mixed approach of evaluation in controlled synthetic and in real-world setups.

When evaluating our experiments on real-world data, we employ three indicators for the prediction accuracy of an algorithm. First, note that it is our intrinsic interest that the accuracy of a classifier increases after denoising the labels. For this purpose we measure the classification performance in terms of the area under the ROC curve (AUC) [20] before and after denoising, and take the difference as an indicator for a algorithm's performance: a good denoising algorithm should yield a substantial higher classification accuracy after denoising, while not overfitting to the training sample. In the synthetic experiment, we observe a kind of bias-variance tradeoff, which helps us also to guess on real data when an algorithm over- and underfits. Second, we use kernel-target-alignment scores as an indicator for the fit between labels and data before and after denoising. KTA scores are complementary to AUCs in the sense that capture how well the separability of the data correlates with the labels. They are less prone to overfitting than AUCs. Third, we invoke expert opinions to ensure the quality of the delivered solution. This has the advantage that we do not rely on labels in this case, but the disadvantage that the expert opinion is subjective and might be biased. In summary, the combined application of the above described measures lets us obtain a guess for the true performance of a denoising algorithm.

3 Theoretical Analysis

In this section, we wish to present a theoretical analysis of our learning methodology. However, the whole 4-step procedure is hard to access theoretically, which is why—as a first step towards this goal—we focus on the new learning method that forms the core of step 1 of our methodology, that is, the novel LATENTSVDD method. In this section we present a generalization analysis of this unsupervised learning algorithm.

We start by defining, for any $\lambda > 0$, the following hypothesis class

$$\mathcal{F}_{\text{LATENTSVDD}} := \Big\{ f_{\boldsymbol{c},\Omega,\mathcal{Z}} = \big(\boldsymbol{x} \mapsto \Omega + \max_{\boldsymbol{z} \in \mathcal{Z}} 2\langle \boldsymbol{c}, \Psi(\boldsymbol{x}, \boldsymbol{z}) \rangle - \|\Psi(\boldsymbol{x}, \boldsymbol{z})\|^2\big) : 0 \leq \|\boldsymbol{c}\|^2 + \Omega \leq \lambda \Big\},$$

and its corresponding loss class $\mathcal{G}_{\text{LATENTSVDD}} := l \circ \mathcal{F}_{\text{LATENTSVDD}}$, employing the loss function $l(t) := \max(0, -t)$. It is not difficult to verify that (e.g., [21], Proposition 12), by employing the variable substitution $\Omega := R^2 - \|\boldsymbol{c}\|^2$, for any $C > 0$ there is an $\lambda > 0$ such that problem (LATENTSVDD) is equivalent to

$$\min_{f \in \mathcal{F}_{\text{LATENTSVDD}}} \frac{1}{n} \sum_{i=1}^{n} l(f(x_i)) \; = \; \min_{g \in \mathcal{G}_{\text{LATENTSVDD}}} \frac{1}{n} \sum_{i=1}^{n} g(x_i).$$

Hence, we may analyze the proposed LATENTSVDD within the proven framework of empirical risk minimization.

Let us first briefly review the classical setup of empirical risk minimization [1]. Let $x_1, \ldots, x_n$ be an i.i.d. sample drawn from a probability distribution P over $\mathcal{X}$. Let $\mathcal{F}$ be a class of functions mapping from $\mathcal{X}$ to some set $\mathcal{Y}$, and let $l : \mathcal{Y} \to [0, b]$ be a bounded loss function, for some $b > 0$. The goal is to find a function $f \in \mathcal{F}$ that has a low risk $\mathbb{E}[l(f(x))]$. Denoting the loss class by $\mathcal{G} := l \circ \mathcal{F}$, this is equivalent finding a function g with small $\mathbb{E}[g]$. The best function in $\mathcal{G}$ we can hope to learn is $g^* \in \operatorname{argmin}_{g \in \mathcal{G}} \mathbb{E}[g]$. Since g^* is unknown, we instead compute a minimizer $\widehat{g}_n \in \operatorname{argmin}_{g \in \mathcal{G}} \widehat{\mathbb{E}}[g]$, where

$\widehat{\mathbb{E}}[g] := \frac{1}{n}\sum_{i=1}^{n} g(x_i)$. To compare the prediction accuracies of g^* and $\widehat{g}_n$, it is known [22] that, with probability at least $1-\delta$ over the draw of the sample,

$$\mathbb{E}[\widehat{g}_n] - \mathbb{E}[g^*] \;\leq\; 4R_n(\mathcal{G}) + b\sqrt{\frac{2\log(2/\delta)}{n}}. \qquad (3)$$

Here, $R_n(\mathcal{G}) := \mathbb{E}\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i g(x_i)$ is the *Rademacher complexity*, where $\sigma_1,\ldots,\sigma_n$ are i.i.d. Rademacher variables (random signs). Usually $R_n(\mathcal{G})$ is of the order $O(1/\sqrt{n})$, when we employ appropriate regularization, and thus so is (3). We will show that also LATENTSVDD enjoys this favorable rate:

Theorem 2 (Generalization bound for LATENTSVDD)**.** *Let* $g^* \in \operatorname{argmin}_{g\in\mathcal{G}_{\text{LATENTSVDD}}} \mathbb{E}[g]$ *and* $\widehat{g}_n \in \operatorname{argmin}_{g\in\mathcal{G}_{\text{LATENTSVDD}}} \frac{1}{n}\sum_{i=1}^{n} g(x_i)$. *Assume there is a real number* $B>0$ *such that* $\mathbb{P}(\|\Psi(\boldsymbol{x}_i,\boldsymbol{z})\|\leq B)=1$. *Denote the cardinality of* $\mathcal{Z}$ *by* $|\mathcal{Z}|$. *Then, the following generalization bound holds:*

$$\mathbb{E}[\widehat{g}_n] - \mathbb{E}[g^*] \;\leq\; 4|\mathcal{Z}|\frac{\lambda + B\sqrt{\lambda}}{\sqrt{n}} + B\sqrt{\frac{2\log(2/\delta)}{n}}.$$

Sketch of Proof. For the proof, we proceed in three steps: first, we prove a Rademacher bound for the classic SVDD (cf. the lemma below). Next, we use Lemma 8.1 in [23] to conclude a Rademacher bound for LATENTSVDD. Finally, we conclude the claimed result by (3). □

The complete proof of Theorem 2 is shown in the supplemental material. It builds on the following generalization bound for the classic SVDD, which is also proved in the supplement.

Lemma 3 (Rademacher bound for SVDD)**.** *Put* $\mathcal{F}_{\text{SVDD}}(\boldsymbol{z}) := \Big\{ f_{\boldsymbol{c},\Omega} = \big(\boldsymbol{x} \mapsto \Omega + 2\langle \boldsymbol{c}, \Psi(\boldsymbol{x}_i,\boldsymbol{z})\rangle - \|\Psi(\boldsymbol{x}_i,\boldsymbol{z})\|^2\big) \;:\; 0 \leq \|\boldsymbol{c}\|^2 + \Omega \leq \lambda \Big\}$ *and* $\mathcal{G}_{\text{SVDD}}(\boldsymbol{z}) := l \circ \mathcal{F}_{\text{SVDD}}(\boldsymbol{z})$ *with* $l(t) := \max(0,-t)$. *Assume there is a real number* $B>0$ *such that* $\mathbb{P}(\|\Psi(\boldsymbol{x}_i,\boldsymbol{z})\|\leq B)=1$. *Then the Rademacher complexity of* $\mathcal{G}_{\text{SVDD}}$ *is bounded as follows:*

$$R(\mathcal{G}_{\text{SVDD}}(\boldsymbol{z})) \;\leq\; \frac{\lambda + B\sqrt{\lambda}}{\sqrt{n}}.$$

Now, while for simplicity the above analysis is based on the assumption of i.i.d. observations, it should be clear, that the very same analysis can also be performed asserting rather "slight violations" of the independence assumption. This is closer to the scenario considered in this paper, where we might face dependent noise in the labels. In particular, we can use the result of Theorem 1 in [24] to prove an analogue of our main result under the assumption of ϕ-mixing data (a formal mathematical relaxation of the i.i.d. assumption). As a result the convergence rate can be slightly or even considerably slower than $O(\sqrt{1/n})$), depending on the "degree of violation" of the independence assumption.

4 Experiments

We examine the effectiveness of the proposed 4-step methodology for learning and evalution in presence of non-i.i.d. label noise on both controlled synthetic and real-world data from the domain of EEG-based brain-computer interfacing (BCI), an popular application domain in the neurosciences. While in the synthetic scenario we have complete control over the truly underlying labels, we rely on indirect evidence for quantifying the results for the EEG data. To this end, we use both quantitative (KTA and AUC scores) and qualitative measures (visual inspection), as described earlier in Section 2.4. However, we compare these quantitative measures with the underlying ground truth in the controlled synthetic setup, so that we can analyze and "align" these measures on the controlled data. In all experiments, we measure the classification performance (before and after denoising) using Linear Discriminant Analysis (LDA) with shrinkage of the covariance matrix, which is the state of the art in single-trial classification of EEG data for event-related potentials [25].

4.1 Latent Space Structure and Model Parameters of Proposed LatentSVDD for Experiments

In Section 2 we have generally described LATENTSVDD in terms of a joint feature map $\Psi(\boldsymbol{x},\boldsymbol{z})$ [12]. For all experiments, we specifically employ a variant of the joint feature map with latent space $\mathcal{Z} := \{1,\ldots,K\}$ that is similar to the multi-class joint feature map [12]: let be $\Lambda(\boldsymbol{z}) = \{\delta(\boldsymbol{z}_1,\boldsymbol{z}),\delta(\boldsymbol{z}_2,\boldsymbol{z}),\ldots,\delta(\boldsymbol{z}_K,\boldsymbol{z})\} \in \{0,1\}^K$ and $a\otimes b$ be the direct tensor product of vectors a and b. Given a data point $\boldsymbol{x}$, we define our joint feature map as $\Psi(\boldsymbol{x},\boldsymbol{z}) = \phi(\boldsymbol{x})\otimes\Lambda(\boldsymbol{z})$. In our experiments we restricted the number of possible latent states to 12. We observed in the experiments, that increasing the number of states beyond 12 usually hardly changes the results (it just leads to additional "unused" latent states). This indicates that 12 is a reasonable choice for the data we employed in the experiments.

4.2 Baseline Methods

We compare our approach against baseline methods which excel in varying areas, such as:

1. dealing with i.i.d. label noise
2. semi-supervised and manifold learning
3. unsupervised learning

Furthermore, it is worth mentioning that most supervised learning algorithms are able to handle, or explicitly assuming, i.i.d. label noise. For instance, *Relevant Dimensionality Estimation* (RDE, [26, 27]), is a state-of-the-art kernel-based learning method for denoising labels. RDE estimates the number of leading kernel principal components required to reconstruct the signal (non-noise) part of the data. By projecting the observed label $\boldsymbol{y}$ to the

first d kPCA [28] components $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$ one obtains the denoised label $\widehat{\boldsymbol{y}} := \sum_{i=1}^{d} \boldsymbol{u}_i \boldsymbol{u}_i^\top \boldsymbol{y}$ RDE works well if the underlying label noise is independent. However, we expect non-i.i.d. label noise which should be indicated by a drop of performance of RDE when compare against LATENTSVDD. Another standard approach for label denoising is *label propagation*, a technique that arose in the general context of semi-supervised Learning (e.g., [29, 30, 31, 32]). Another baseline approach is given by the vanilla support vector data description [9] method (without a latent variable). This method infers a model of normality for each class. This way we can obtain a denoised set of labels depending on whether or not a data point is contained within the normality ball output by SVDD.

4.3 Controlled Synthetic Experiment

We designed a synthetic experiment that resembles the EEG setting that we have in mind, which is characterized by unbalanced classes and small sample sizes. To this end, we independently sampled 300 positive and 80 negative instances from two-dimensional Gaussian distributions. To further increase the complexity of the problem, we we added 250 Gaussian noise dimensions. This results in data having a similar dimensionality as usually encountered in EEG experiments. Then, we introduce systematic label noise by randomly flipping the label of positive instances that have $x_2 < -0.5$ and that of negative instances with $x_2 > +0.5$. This results in roughly 35% of the labels being flipped. Note that this leads to rather systematic label noise, which is to be contrasted to uniform label noise, which could, e.g., result from randomly flipping *all* labels.

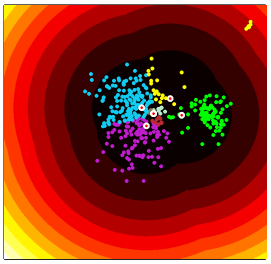

Figure 1: Anomaly scores and latent variable tessellation for LATENTSVDD.

Performance is measured by sampling 25 times from the data pool and dividing into 75% training data and 25% test data. Since (shrinkage) LDA estimates its optimal parameter analytically, no model selection and hence no validation data set was necessary. The experiment was repeated 50 times for several levels of systematic label noise (ranging from 0% up to 100%). The average results over the 50 repetitions of the experiment are shown in Figure 2. For the four compared denoising approaches, we report the results in terms of the AUC achieved by LDA when testing on the original truly underlying labels (left-hand figure), while the black-dashed line shows the accuracy on the originally observed (i.e., undenoised) labels. The kernel-target-alignment scores associated with the denoised labels are shown in the center figure. The right-hand figure shows the 0-1 error of the denoised labels with respect to the truly underlying labels.

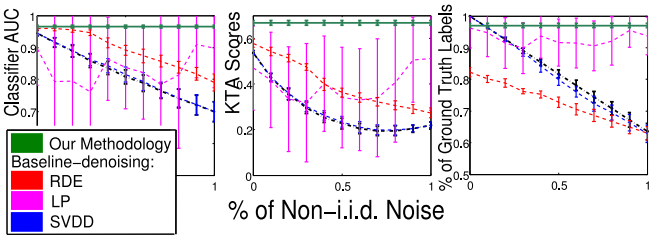

Figure 2: Accuracy in terms of AUC for all methods tested on the underlying true labels (left). Kernel target alignment scores for the denoised labels (center). And fraction of correctly inferred labels given the underlying true labels (right).

We observe that the vanilla SVDD's performance drops for settings with increasing label noise. This might result from the fact that the SVDD infers a model of normality for each class separately, thus ignoring the coupling induced by the latent noise structure. Label propagation leads to intrinsically unstable solutions, which can be concluded from its large error bars. Generally, it seems to be much more sensitive to random perturbations in the label noise than its competitors. In low-noise settings, the RDE baseline has the highest KTA scores and correspondingly also high AUC values, but—of all methods compared—the lowest truly underlying performance (lowest agreement between true and the denoised labels), which can be observed from the right-hand side of Figure 3. This indicates that RDE overfits to the training labels. Finally, the proposed multi-step methodology based on LATENTSVDD is less affected by variations in the label noise level, and overall achieves the highest accuracies, while having small error bars, that is, it is more stable than the compared methods. Furthermore, we observed that in average only 6 iterations are needed for convergence of the optimization algorithm. Figure 1 shows exemplary contour plots of anomaly scores and tessellation induced by the latent variable as output by LATENTSVDD.

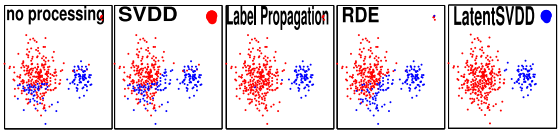

Figure 3: Exemplary denoising results for all methods by 40% systematic label noise. The dot size codes for the anomaly scores as returned by SVDD and LATENTSVDD.

4.4 EEG Experiment

We evaluated our proposed learning methodology on the data of an EEG-BCI experiment, for which we recorded 20 participants. The results are presented in this section.

4.4.1 Setting

Motivation & Neuroscientific Background In our EEG experiment, we address the question of whether or not the brain of a participant processed a response error. Conventionally, the EEG data would be analyzed based on the *be-*

havioral response of the participant, grouping all trials together where the behavioral response is de facto correct or wrong (= behavioral labels). However, having committed a mistake *behaviorally* does not equate having processed it *neurally* [5]. While the *neural* processing is what we are really interested in, these neural labels are unknown, as no ground truth is available. We used LATENTSVDD for finding these neural labels in a data-driven way, with the goal of dividing the EEG trials: those where an error was processed neurally, and those where none was processed.

When participants recognize having committed a response error, two specific components are evoked in the event-related potential (ERP) of the EEG signal: an error negativity (N_e) and an error positivity (P_e). Out of these, only the P_e has been attributed to error or post-error processing itself [33]. Therefore, we focus on the P_e in the following, which is characterized by a centro-parietal maximum 200–500ms after feedback [34, 35, 36, 37].

Paradigm & Methods In our experiment, 20 participants were asked to perform a fast-paced d2 test [38], a common test of visual selective attention. In this test, participants are presented two types of visual stimuli and are asked to distinguish between these two stimuli by pressing the corresponding button: the right hand should be used for the target stimulus (20% of trials), the left hand for the non-target stimulus (80% of trials). In total, each participant assessed 300 stimuli under time pressure. Feedback was given 500 ms after each response, both on reaction time and correctness. Brain activity was recorded with multi-channel EEG amplifiers (BrainAmp DC by Brain Products, Munich, Germany) with 119 Ag/AgCl electrodes placed according to an extended international 10-10 system, sampled at 1000 Hz and band-pass filtered between 0.05 Hz and 200 Hz.

We examined the neural response that was elicited by receiving feedback. For this, the EEG data was divided into epochs of 500 ms, starting from the onset of feedback. These epochs were baseline corrected (based on the 200 ms interval prior to feedback) and artifact rejection was performed. As features for LATENTSVDD and classification, we calculated 9 features per epoch. For this purpose, the interval [0 500 ms] was divided in 10 non-overlapping intervals of 50 ms length. We then calculated the mean signal in each of these intervals and subsequently, the gradient between these means. In order to test class separability, we classified the EEG data using shrinkage LDA, sampling 30 times from the data set and dividing the data set into 75% training data and 25% test data. Classification was run using (a) behavioral labels, (b) the 'neural' labels suggested by LATENTSVDD, and, for comparison, those derived by SVDD, LP and RDE. We expect the 'neural' classes to be better separable than before (higher AUC values) and to have a better matching of labels and data (higher KTA scores), compared to using behavioral labels (correct vs. incorrect responses).

4.4.2 Results

Class Re-Assignment and Anomalous Trials On average, LATENTSVDD flipped the labels for 35.94% of all trials. This resulted in a *neural* error rate of 31.18%, compared the lower *behavioral* error rate (18.05%). Based on the anomaly score that LATENTSVDD returns for each trial, we rejected a small percentage of trials for each participant (cf section 2.2.). For the majority of participants, there are only few trials with high anomaly scores, with a steep drop-off compared to the other trials (cf Figure 4). Visual inspection revealed that the results also make sense neuroscientifically: the rejected trials show typical artifacts (eye blinks, voltage drifts with respect to all electrodes or a single electrode) that have escaped the conventional artifact rejection run prior to applying LATENTSVDD, as well as trials with unusually high amplitudes.

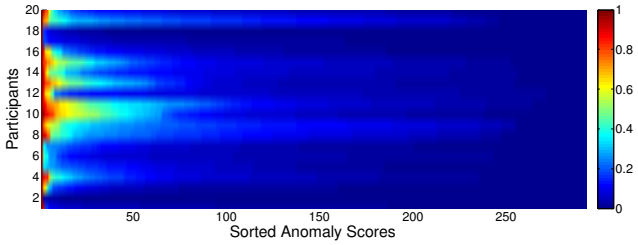

Figure 4: Sorted anomaly scores for each data point of each participant.

Quantitative Assessment We quantified the benefits of LATENTSVDD using KTA scores and linear classification (LDA). Both measures confirm that the labels assigned by LATENTSVDD allow a much better separation of the data than behavioral labels for all 20 participants. As can be seen in Figure 5.B, LATENTSVDD renders the classes clearly more distinct from each other, reflected in higher AUC values (0.95 ± 0.02 versus 0.60 ± 0.08). This is accompanied by substantially higher KTA score for all participants. As can be seen in Figure 5.A, LATENTSVDD is also superior compared to other denoising methods (SVDD, LP, RDE). SVDD and LP lag far behind, both in AUC and KTA scores. In fact, applying these methods even makes separability of classes worse than before (no method: 0.60 ± 0.08, SVDD: 0.59 ± 0.07, LP: 0.54 ± 0.17). In contrast, RDE proves to be a close competitor to LATENTSVDD. However, our approach shows better results for this EEG experiment, with a mean AUC score of 0.95 ± 0.02 (RDE: 0.90 ± 0.04) and a mean KTA score of 0.3911 (RDE: 0.2842).

Neuroscientific Assessment While AUC and KTA scores help quantify the positive effect of LATENTSVDD, the results are also neurophysiologically sound. In the following, we discuss this for our methodology at the example of participant 5. The different steps of our methodology are visualized in Figure 6. Each plot shows the same data

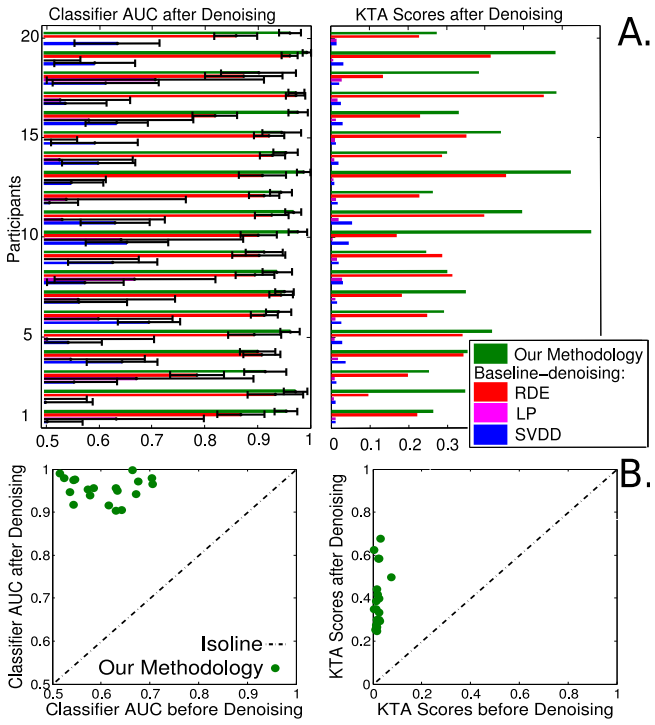

Figure 5: AUC and KTA results for all participants of the experiment.

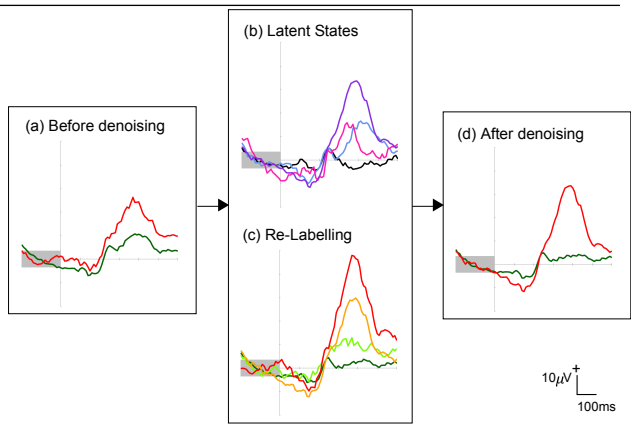

Figure 6: Time course at electrode Cz: (a) before denoising (behavioral labels), (b) latent brain states revealed by LATENTSVDD, (c) resulting re-assignment of labels, (d) after denoising.

(time course at electrode Cz), yet grouped in different classes. The conventional approach is shown on the far left (a), the superior results retained by LATENTSVDD on the far right (d), with classes that are clearly better separable. Initially (Figure 6(a), classes show great similarity (correct responses in green, erroneous responses in red). Our methodology reveals four latent brain states (Figure 6(b)). The state with the highest amplitude (purple) corresponds to typical error processing, with a clear positive component P_e. A clear positivity also occurs in the blue and pink state, yet less pronounced and with different latencies. In contrast, no error has been processed in the black state. Based on the latent variable, a subset of trials is then re-assigned (Figure 6(c)). Red and green indicate labels that are retained, orange and light green signify trials where the labels were switched (orange to red, light green to green). As can be seen, the re-assignment makes sense intuitively. Finally, Figure 6(d) shows the denoised data, which reveals a more pronounced error positivity P_e (red) than before. While the latent states themselves are highly subject-specific, we find similar results, i.e. the recovery of a stronger P_e component than before, for all other participants.

5 Conclusion

Finding the true label for data with systematic, non-i.d.d. label noise is a common challenge in experimental disciplines such as the neurosciences. We proposed a 4-step methodology for learning and evaluation in presence of non-i.i.d. label noise, in the heart of which lies a novel learning algorithm—LATENTSVDD—that allows to capture the hidden state of the label noise. We optimized the associated objective function by a DC-type algorithm, of which we prove convergence, and we derived an equivalent Fenchel dual criterion. Our approach enjoys deep learning-theoretical guarantees with the usual $O(1/\sqrt{n})$ convergence rate.

Our method achieves the most competitive performance when labels, which we demonstrate in a series of controlled synthetic and real-world experiments. The core of which is an extensive case study of EEG-BCI data recorded during an attention test, where we observed that the labels denoised by the proposed methodology lead to substantial better separability of the data (assessed with linear classification; rise in the mean AUC from 0.60 to 0.95 for EEG data). Visual inspection of the data by a domain expert shows that the class assignments output by our methodology are neurophysiologically plausible, leading to more easily interpretable brain states that subsequently allow for a better and more meaningful experimental evaluation.

The joint feature map construction in principle could also allow for a more complex encoding of structure such as e.g. trees or hidden markov models. It will be interesting to extend our novel methodology for multi-modal neuroimaging data [39, 40], and furthermore explore applications beyond the neurosciences.

Acknowledgments

We gratefully thank Benjamin Blankertz, Andres Munoz Medina, and Mehryar Mohri for valuable comments and helpful suggestions. This work was supported by the German Bundesministerium für Bildung und Forschung (BMBF FKZ 01GQ0850 and 01IB001A), the German Science Foundation (DFG MU 987/6-1, RA 1894/1-1), the European Community's 7th Framework Programme under the PASCAL2 Network of Excellence (ICT-216886), and the BK21 program of NRF.

References

[1] V. Vapnik, *Statistical learning theory*. New York: John Wiley, 1998.

[2] G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, eds., *Toward Brain-Computer Interfacing*. Cambridge, MA: MIT Press, 2007.

[3] B. Blankertz, M. Tangermann, C. Vidaurre, S. Fazli, C. Sannelli, S. Haufe, C. Maeder, L. E. Ramsey, I. Sturm, G. Curio, and K.-R. Müller, "The Berlin Brain-Computer Interface: Non-medical uses of BCI technology," *Front Neuroscience*, vol. 4, p. 198, 2010. Open Access.

[4] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller, "Introduction to machine learning for brain imaging," *Neuroimage*, vol. 56, pp. 387–399, 2011.

[5] A. K. Porbadnigk, M. S. Treder, B. Blankertz, J.-N. Antons, R. Schleicher, S. Möller, G. Curio, and K.-R. Müller, "Single-trial analysis of the neural correlates of speech quality perception," *Journal of Neural Engineering*, vol. 10(5), p. 056003, 2013.

[6] A. K. Porbadnigk, N. Görnitz, M. Kloft, and K.-R. Müller, "Decoding brain states by supervising unsupervised learning," *Journal of Computing Science and Engineering*, vol. 7(2), pp. 112–121, 2013.

[7] V. Vapnik, *The nature of statistical learning theory*. New York: Springer Verlag, 1995.

[8] W. Polonik, "Measuring mass concentration and estimating density contour clusters – an excess mass approach," *Annals of Statistics*, vol. 23, pp. 855–881, 1995.

[9] D. M. Tax and R. P. Duin, "Support vector data description," *Machine Learning*, vol. 54, pp. 45–66, 2004.

[10] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms," *IEEE Neural Networks*, vol. 12, pp. 181–201, May 2001.

[11] B. Schölkopf and A. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2002.

[12] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large Margin Methods for Structured and Interdependent Output Variables," *Journal of Machine Learning Research*, vol. 6, pp. 1453–1484, 2005.

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, pp. 1627–1645, Sept. 2010.

[14] B. K. Sriperumbudur and G. R. G. Lanckriet, "On the convergence of the concave-convex procedure," in *NIPS* (Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, eds.), pp. 1759–1767, Curran Associates, Inc., 2009.

[15] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, no. 4, pp. 915–936, 2003.

[16] E. Yen, N. Peng, P.-W. Wang, and S.-D. Lin, "On convergence rate of concave-convex procedure," *Proceedings of the NIPS 2012 Optimization Workshop*, 2012.

[17] G. Rätsch, S. Mika, B. Schölkopf, and K.-R. Müller, "Constructing boosting algorithms from SVMs: An application to one-class classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1184–1199, Sept. 2002.

[18] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola, "On kernel target alignment," in *Advances in Neural Information Proccessing Systems (NIPS)*, vol. 14, pp. 367–737, 2001.

[19] C. Cortes, M. Mohri, and A. Rostamizadeh, "Algorithms for learning kernels based on centered alignment," *J. Mach. Learn. Res.*, vol. 13, pp. 795–828, Mar. 2012.

[20] C. Cortes and M. Mohri, "Confidence intervals for the area under the roc curve," in *NIPS*, 2004.

[21] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller, and A. Zien, "Efficient and accurate lp-norm multiple kernel learning," in *Advances in Neural Information Processing Systems 22* (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, eds.), pp. 997–1005, MIT Press, 2009.

[22] P. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, Nov. 2002.

[23] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. The MIT Press, 2012.

[24] M. Mohri and A. Rostamizadeh, "Rademacher complexity bounds for non-i.i.d. processes," in *NIPS* (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), pp. 1097–1104, Curran Associates, Inc., 2008.

[25] B. Blankertz, S. Lemm, M. S. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components – a tutorial," *Neuroimage*, vol. 56, pp. 814–825, 2011.

[26] M. L. Braun, J. Buhmann, and K.-R. Müller, "On relevant dimensions in kernel feature spaces," *Journal of Machine Learning Research*, vol. 9, pp. 1875–1908, Aug 2008.

[27] G. Montavon, M. L. Braun, T. Krueger, and K.-R. Müller, "Analyzing local structure in kernel-based learning: Explanation, complexity and reliability assessment," *IEEE Signal Processing Magazine, Special Issue on Kernel-Based Learning for Signal Processing*, vol. 30(4), pp. 62–74, 2013.

[28] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput*, vol. 10, no. 5, pp. 1299–1319, 1998.

[29] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," tech. rep., Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

[30] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *International Conference on Machine Learning (ICML)*, vol. 20, p. 912, 2003.

[31] N. Görnitz, M. Kloft, and U. Brefeld, "Active and semi-supervised data domain description," in *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)* (W. L. Buntine, M. Grobelnik, D. Mladenic, and J. Shawe-Taylor, eds.), pp. 407–422, 2009.

[32] N. Görnitz, M. Kloft, K. Rieck, and U. Brefeld, "Toward Supervised Anomaly Detection.," *Journal of Artificial Intelligence Research (JAIR)*, vol. 46, pp. 235–262, 2013.

[33] M. Falkenstein, J. Hoormann, S. Christ, and J. Hohnsbein, "ERP components on reaction errors and their functional significance: a tutorial," *Biol Psychol*, vol. 51, no. 2-3, pp. 87–107, 2000.

[34] J. Hohnsbein, M. Falkenstein, and J. Hoormann, "Error processing in visual and auditory choice reaction tasks," *Journal of Psychophysiology*, vol. 3, p. 320, 1998.

[35] M. Falkenstein, J. Hohnsbein, J. Hoormann, and L. Blanke, "Effects of errors in choice reaction tasks on the ERP under focused and divided attention," in *Psychophysiological Brain Research* (C. Brunia, A. Gaillard, and A. Kok, eds.), pp. 192–195, Tilburg University Press, Tilburg, 1990.

[36] W. Gehring, M. Coles, D. Meyer, and E. Donchin, "The error-related negativity: an event-related brain potential accompanying errors," *Psychophysiology*, vol. 27, p. S34, 1990.

[37] W. Gehring, B. Goss, M. Coles, D. Meyer, and E. Donchin, "A neural system for error detection and compensation," *Psychological Science*, vol. 4, pp. 385–390, 1993.

[38] R. Brickenkamp and E. Zillmer, *D2 Test of Attention*. Göttingen, Germany: Hogrefe & Huber, 1998.

[39] F. Bießmann, F. C. Meinecke, A. Gretton, A. Rauch, G. Rainer, N. Logothetis, and K.-R. Müller, "Temporal kernel canonical correlation analysis and its application in multimodal neuronal data analysis," *Machine Learning*, vol. 79, no. 1-2, pp. 5—27, 2009.

[40] F. Bießmann, S. M. Plis, F. C. Meinecke, T. Eichele, and K.-R. Müller, "Analysis of multimodal neuroimaging data," *Biomedical Engineering, IEEE Reviews in*, vol. 4, pp. 26–58, 2011.

[41] M. Talagrand, "Concentration of measure and isoperimetric inequalities in product spaces," *Publications Mathmatiques de L'IHS*, vol. 81, pp. 73–205, 1995.

Supplemental Material

A Proofs

When bounding the Rademacher complexity for Lipschitz continuous loss classes (such as the hinge loss or the squared loss), the following lemma is often very helpful.

Lemma A.1 (Talagrand's lemma [41])**.** *Let $l : \mathbb{R} \to \mathbb{R}$ be a loss function that is L-Lipschitz continuous and $l(0) = 0$. Let $\mathcal{F}$ be a hypothesis class of real-valued functions and denote its loss class by $\mathcal{G} := l \circ \mathcal{F}$. Then the following inequality holds:*

$$R_n(\mathcal{G}) \leq 2LR_n(\mathcal{F}).$$

We can use the above result to prove Lemma 3.

Proof of Lemma 3. Since the LATENTSVDD loss function is 1-Lipschitz with $l(0) = 0$, by Lemma A.1, it is sufficient to bound $R(\mathcal{F}_{\text{SVDD}}(\boldsymbol{z}))$. To this end, it holds

$$\begin{aligned} R(\mathcal{F}_{\text{SVDD}}(\boldsymbol{z})) &\overset{\text{def.}}{=} \mathbb{E}\Big[\sup_{\boldsymbol{c},\Omega:0\leq\|\boldsymbol{c}\|^2+\Omega\leq\lambda} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\big(\Omega \\ &\quad + 2\langle \boldsymbol{c}, \Psi(\boldsymbol{x}_i,\boldsymbol{z})\rangle - \|\Psi(\boldsymbol{x}_i,\boldsymbol{z})\|^2\big)\Big] \\ &\leq \mathbb{E}\left[\sup_{\Omega:-\lambda\leq\Omega\leq\lambda} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\Omega\right] \\ &\quad + 2\mathbb{E}\left[\sup_{\boldsymbol{c}:\|\boldsymbol{c}\|^2\leq\lambda} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\langle \boldsymbol{c}, \Psi(\boldsymbol{x}_i,\boldsymbol{z})\rangle\right)\right] \\ &\quad + \underbrace{\mathbb{E}\left[-\frac{1}{n}\sum_{i=1}^{n}\sigma_i\|\Psi(\boldsymbol{x}_i,\boldsymbol{z})\|^2\right]}_{=0 \text{ (by symmetry of } \sigma_i)}. \end{aligned} \tag{A.3.1}$$

Note that the term to the right is zero because the Rademacher variables are random signs, independent of $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$. The term to the left can be bounded as follows:

$$\begin{aligned} \mathbb{E}\left[\sup_{\Omega:-\lambda\leq\Omega\leq\lambda} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\Omega\right] &= \lambda\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\right|\right] \\ &\overset{(*)}{\leq} \lambda\sqrt{\mathbb{E}\left[\frac{1}{n^2}\sum_{i,j=1}^{n}\sigma_i\sigma_j\right]} \\ &= \frac{\lambda}{\sqrt{n}}. \end{aligned} \tag{A.3.2}$$

where for $(*)$ we employ Jensen's inequality. Moreover, applying the Cauchy-Schwarz inequality and Jensen's inequality, respectively, we obtain

$$\begin{aligned} &\mathbb{E}\left[\sup_{\boldsymbol{c}:\|\boldsymbol{c}\|^2\leq\lambda} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\langle \boldsymbol{c}, \Psi(\boldsymbol{x}_i,\boldsymbol{z})\rangle\right)\right] \\ &\overset{\text{C.-S.}}{\leq} \mathbb{E}\left[\sup_{\boldsymbol{c}:\|\boldsymbol{c}\|^2\leq\lambda} \|\boldsymbol{c}\|\left\|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\Psi(\boldsymbol{x}_i,\boldsymbol{z})\right\|\right] \\ &\overset{\text{Jensen}}{\leq} \sqrt{\lambda\mathbb{E}\left[\frac{1}{n^2}\sum_{i,j=1}^{n}\sigma_i\sigma_j\langle\Psi(\boldsymbol{x}_i,\boldsymbol{z}),\Psi(\boldsymbol{x}_j,\boldsymbol{z})\rangle\right]} \\ &= \sqrt{\lambda\frac{1}{n^2}\sum_{i=1}^{n}\|\Psi(\boldsymbol{x}_i,\boldsymbol{z})\|^2} \\ &\leq B\sqrt{\frac{\lambda}{n}} \end{aligned} \tag{A.3.3}$$

because $\mathbb{P}(\|\Psi(\boldsymbol{x}_i,\boldsymbol{z})\| \leq B) = 1$. Hence, inserting the results (A.3.2) and (A.3.3) into (A.3.1), yields the claimed result, that is,

$$\begin{aligned} R(\mathcal{G}_{\text{SVDD}}(\boldsymbol{z})) &\overset{\text{Lemma A.1}}{\leq} R(\mathcal{F}_{\text{SVDD}}(\boldsymbol{z})) \\ &\leq \frac{\lambda}{\sqrt{n}} + B\sqrt{\frac{\lambda}{n}} = \frac{\lambda + B\sqrt{\lambda}}{\sqrt{n}}. \end{aligned} \tag{A.3.4}$$

□

Next, we invoke the following result, taken from [23] (Lemma 8.1).

Lemma A.2. *Let $\mathcal{F}_1,\ldots,\mathcal{F}_l$ be hypothesis sets in $\mathbb{R}^{\mathcal{X}}$, and let $\mathcal{F} := \{\max(f_1,\ldots,f_l\} : f_i \in \mathcal{F}_i, i \in \{1,\ldots,l\}\}$. Then,*

$$R_n(\mathcal{F}) \leq \sum_{j=1}^{l} R_n(\mathcal{F}_j).$$

Sketch of proof [23]. The idea of the proof is to write $\max(h_1,h_2) = \frac{1}{2}(h_1 + h_2 + |h_1 - h_2|)$, and then to show that

$$\mathbb{E}\left[\sup_{h_1\in\mathcal{F}_1,h_2\in\mathcal{F}_2} \frac{1}{n}\sum_{i=1}^{n}|h_1(x_i) - h_2(x_i)|\right] \leq R_n(\mathcal{F}_1) + R_n(\mathcal{F}_2).$$

This proof technique also generalizes to $l > 2$. □

We can use Lemma A.2 and Lemma 3, to conclude the main theorem of this paper, that is, Theorem 2, which establishes generalization guarantees of the usual order $O(1/\sqrt{n})$ for the proposed LATENTSVDD method.

Proof of Theorem 2. First observe that, because l is 1-Lipschitz,

$$R_n(\mathcal{G}_{\text{LATENTSVDD}}) \leq R_n(\mathcal{F}_{\text{LATENTSVDD}}).$$

Next, note that we can write

$$R_n(\mathcal{F}_{\text{LATENTSVDD}}) \;=\; \left\{ \max_{\boldsymbol{z}\in\mathcal{Z}} (f_{\boldsymbol{z}}) : f_{\boldsymbol{z}} \in \mathcal{F}_{\text{SVDD}}(\boldsymbol{z}) \right\}.$$

Thus, by Lemma 2 and Lemma 4,

$$\begin{aligned} R_n(\mathcal{F}_{\text{LATENTSVDD}}) \;&\leq\; |\mathcal{Z}| \max_{z\in\mathcal{Z}} R_n(\mathcal{F}_{\text{SVDD}}(\boldsymbol{z})) \\ &\leq\; |\mathcal{Z}| \frac{\lambda + B\sqrt{\lambda}}{\sqrt{n}}. \end{aligned}$$

Moreover, observe that the loss function in the definition of $\mathcal{G}_{\text{LATENTSVDD}}$ can only range in the interval $[0, B]$. Thus, Theorem 2 in the main paper gives the claimed result, that is,

$$\begin{aligned} \mathbb{E}[\widehat{g}_n] - \mathbb{E}[g^*] \;&\leq\; 4R_n(\mathcal{G}_{\text{LATENTSVDD}}) + B\sqrt{\frac{2\log(2/\delta)}{n}} \\ &\leq\; 4|\mathcal{Z}| \frac{\lambda + B\sqrt{\lambda}}{\sqrt{n}} + B\sqrt{\frac{2\log(2/\delta)}{n}}. \end{aligned}$$

□