Hidden Markov Anomaly Detection

Nico Görnitz NICO.GOERNITZ@TU-BERLIN.DE
Berlin Institute of Technology, 10587 Berlin, Germany

Mikio Braun MIKIO.BRAUN@TU-BERLIN.DE
Berlin Institute of Technology, 10587 Berlin, Germany

Marius Kloft KLOFT@HU-BERLIN.DE
Humboldt University of Berlin, 10099 Berlin, Germany

Abstract

We introduce a new anomaly detection methodology for data with latent dependency structure. As a particular instantiation, we derive a hidden Markov anomaly detector that extends the regular one-class support vector machine. We optimize the approach, which is non-convex, via a DC (difference of convex functions) algorithm, and show that the parameter ν can be conveniently used to control the number of outliers in the model. The empirical evaluation on artificial and real data from the domains of computational biology and computational sustainability shows that the approach can achieve significantly higher anomaly detection performance than the regular one-class SVM.

1. Introduction

In the age of big data, effective filtering methodologies for unlabeled data such as the framework of learning-based *anomaly detection* (Markou & Singh, 2003; Chandola et al., 2009) are gaining increasing interest by the machine learning community (Blanchard et al., 2010; Saligrama & Zhao, 2012; Kloft & Laskov, 2012; Görnitz et al., 2014). Learning-based anomaly detection methods are at the heart of several important applications areas, including, in computer security, the detection of yet unsignatured attacks and novel intrusions in computer networks (Jyothsna et al., 2011) and, in computational biology, the characterization of systematic anomalies in microarray analysis (Noto et al., 2014) and deep sequencing data (Kukita et al., 2013).

Prominent approaches to learning-based anomaly detection include prevalent kernel-based approaches such as the one-class support vector machine (OC-SVM) (Schölkopf et al., 2001) or support vector data description (Tax & Duin, 2004). Such methods use a kernel approach to learn a non-linear representation of the class membership that can be used to predict the anomaly score of new and yet unseen inputs. These methods are based on the fundamental assumption that the nominal inputs x_i are realized independently from a common probability distribution P, without exploiting any potential patterns contained in the outputs y_i. However, in many real-world applications, the data is associated with an inherently underlying output structure: e.g., in intrusion detection or speech and text recognition the data naturally admits a language and grammar structure (Rieck et al., 2010; Joachims et al., 2009); in bioinformatics, the annotation into exonic, intronic, and intergenic regions inherently underlies genomic data (Schweikert et al., 2009). It has been shown that methodologies exploiting such potential structure such as the structured support vector machine (SSVM) (Tsochantaridis et al., 2005a) can substantially help performance in such contexts (Rätsch & Sonnenburg, 2007; Joachims et al., 2009; Rieck et al., 2010).

In this paper, we propose a novel kernel-based framework for the detection of anomalies with underlying sequence structure, called *hidden Markov anomaly detection*. This approach can be introduced in a general way for data with latent dependency structure, and, for a specific choice of loss function and joint feature map, we obtain a hidden Markov analogue to the one-class SVM. Just like the original OC-SVM, the method has a parameter ν that controls the fraction of the outliers, and enjoys deep theoretical guarantees: we prove that a large deviation bound holds on the generalization error (measured with respect to the loss function).

The presented approach is general enough to work with many problems that can be addressed within the structured output prediction framework. In this paper though, we focus on using hidden Markov chain type structured output models, as common in label sequence learning, to facilitate the detection of anomalous sequences where changes follow a hidden Markov model. That way, we achieve much better performance compared to the standard OC-SVM, which treats each position independently, as our empirical evaluation shows.

Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

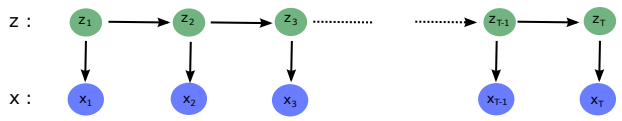

Figure 1. Factorization of hidden Markov models: the latent variables (z, green) can not be observed directly, instead, noisy observations (x, blue) and bindings between consecutive latent variables give rise to their current state.

The remainder of the paper is structured as follows: in Section 2 we describe the problem setting. In Section 3 we introduce the novel latent anomaly detection framework, leading to hidden Markov anomaly detection (Section 4.2), for which we develop an effective optimization algorithm. Our method is evaluated on controlled artificial data and two real-world data sets from bioinformatics and computational sustainable energy applications (Section 5). Section 6 concludes. Additional run time experiments and mathematical proofs are presented in the supplementary material.

2. Anomaly Detection

In *anomaly detection* (Chandola et al., 2009) we are given a set of input instances $x_1, \ldots, x_n \in \mathcal{X}$, which are commonly assumed to be realized from independent and identically distributed (i.i.d) random variables $X_1, \ldots, X_n \sim P$, where P is a potentially unknown measure of probability. The aim is to find a set containing the most typical instances under the measure P, and instances lying outside of the set are declared as anomalies. The task of anomaly detection can be formally phrased within the framework of density level set estimation (Tsybakov, 1997) as follows. Denoting by X another i.i.d. copy according to P, the theoretically optimal nominal set is $L_\nu := \{x \in \mathcal{X} : p(x) \geq b_\nu\}$ for $\nu \in]0, 1[$ and b_ν such that $P(X \notin L_\nu) = \nu$, which is called the ν *density level set* and can be interpreted as follows: L_ν contains the most likely inputs under the density p, while rare or untypical data ("anomalies") are modeled to lie outside of L_ν. The parameter ν indicates the fraction of outliers in the model.

The aim is to compute, based on the data $x_1, \ldots, x_n \in \mathcal{X}$, a good approximation of L_ν, that is, to determine a function $f : \mathcal{X} \to \mathbb{R}$ giving rise to an estimated density level set $\hat{L}_\nu := \{x \in \mathcal{X} : f(x) \geq 0\}$. It is desirable that $\hat{L}_\nu$ closely approximates the true density level set L_ν, i.e., $\hat{L}_\nu$ converges to L_ν in probability, that is,

$$P(\hat{L}_\nu \backslash L_\nu \cup L_\nu \backslash \hat{L}_\nu) \to 0 \text{ for } n \to \infty.$$

This implies that $\hat{L}_\nu$ has asymptotically probability mass ν, that is, $P(X \notin \hat{L}_\nu) \to \nu$ for $n \to \infty$. Classic approaches to anomaly detection include kernel-based ones (Müller et al., 2001) such as the one-class support vector machine (Schölkopf et al., 2001) (OC-SVM). The OC-SVM is one of the most prominent and successful anomaly detectors and employs linear models $f_{\boldsymbol{w},\rho}(x) = \langle \boldsymbol{w}, \phi(x) \rangle - \rho$, where the data is mapped into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ via a feature map $\phi : \mathcal{X} \to \mathcal{H}$. It subsequently separates a fraction of $1 - \nu$ many inputs from the origin with maximum margin:

$$\max_{\boldsymbol{w},\rho,\boldsymbol{\xi} \geq 0} \|\boldsymbol{w}\|^2 - \rho + \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i \qquad \text{(OC-SVM)}$$
$$\text{s.t.} \quad \xi_i \geq -f_{\boldsymbol{w},\rho}(x_i) \quad \forall\, i = 1, \ldots, n.$$

However, this approach does not exploit latent dependency structure of the data. Latent dependencies, however, are prevalent in many real-world applications, e.g., network intrusion detection (Rieck et al., 2010; Kloft & Laskov, 2011; Görnitz et al., 2009a;b; Görnitz et al., 2013; Kloft et al., 2008), speech and text recognition (Joachims et al., 2009), and gene finding (Rätsch & Sonnenburg, 2007). In this case, the i.i.d. postulate no longer holds: having a latent dependency structure means that there are unobserved latent variables Z_i such that, only when conditioned on Z_i, the X_i become conditionally independent. In other words, if the latent structures Z_i are unknown, the input variables X_i are dependent.

Many real-world problems are sequential by nature, with observations stemming from probability distributions according to a corresponding hidden state sequence. E.g. the goal in gene finding is the segmentation of the DNA input sequence into genic and intergenic regions where the observed sequence of nucleotides (A,C,G,T) changes its distribution whenever we enter or leave a genic region. In this paper, we focus on problems that exhibit sequence structure where observations change their distribution according to a corresponding latent state sequence and can be tackled with hidden Markov models (see Figure 1). We show that exploiting latent structure directly, leads to significant improvements over state-of-the-art methods. This problem can be solved in a generic way for latent dependencies, as we show in the next section.

3. Latent Anomaly Detection

In the problem setting of *latent anomaly detection*, we extend the expressiveness of the model given in Eqn. (OC-SVM) by considering models of the form $f_{\boldsymbol{w},\rho}(x) = \max_{z \in \mathcal{Z}} \langle \boldsymbol{w}, \Psi(x, z) \rangle + \delta(z) - \rho$, where $\Psi : \mathcal{X} \times \mathcal{Z} \to \mathcal{H}$ is a *joint feature map* into a reproducing kernel Hilbert space $\mathcal{H}$ that corresponds to a kernel function $k : (\mathcal{X} \times \mathcal{Z}) \times (\mathcal{X} \times \mathcal{Z}) \to \mathbb{R}$, and $\delta : \mathcal{Z} \to \mathbb{R}$ is a prior weight function of the instances $z \in \mathcal{Z}$. This is a principled way of approaching the encoding problem for arbitrary dependencies between x and z as it is common in the structured output literature (Tsochantaridis et al., 2005b). Albeit, it has been already used to encode hidden Markov and hidden semi-Markov models (Görnitz et al., 2011; Rätsch & Sonnenburg, 2007), it is not restricted to those and has been applied to Markov random fields (Nowozin & Lampert, 2010), weighted context-free grammars and taxonomies (Tsochantaridis et al., 2005b). Here, the maximization step for the latent variable z acts as a fre-

quentist's equivalent to marginalization in basic probability theory (Nowozin & Lampert, 2010).

Employing the above notation, we phrase the *primal optimization problem* of latent anomaly detection as follows:

Problem 1 (PRIMAL LATENT ANOMALY DETECTION OPTIMIZATION PROBLEM)**.** *Given a monotonically non-decreasing loss function $l : \mathbb{R} \to \mathbb{R}$, minimize, with respect to $\boldsymbol{w} \in \mathcal{H}$ and $\rho \in \mathbb{R}$,*

$$\frac{1}{2}\|\boldsymbol{w}\|^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n} l\Big(\rho - \max_{z\in\mathcal{Z}}\big(\langle \boldsymbol{w}, \Psi(x_i, z)\rangle + \delta(z)\big)\Big). \tag{P}$$

The interpretation of the above formulation is as follows. The loss function could be, e.g., $l(t) = \max(0, t)$, in which case the above detection method extends the one-class support vector machine (Schölkopf et al., 2001) to the latent domain (this is extensively discussed in the upcoming Section 4.2). Variants of this detection method can be obtained from the above general formulation by employing different loss functions, e.g., of logistic or exponential type ($l(t) = \log(1 + \exp(t))$ and $l(t) = \exp(t)$, respectively). It is important to note that, when contrasted to the classical kernel-based hypothesis model $f_{\boldsymbol{w},\rho}(\phi(x)) = \langle \boldsymbol{w}, \phi(x)\rangle - \rho$, the above detection method employs a latent hypothesis model of the form $f_{\boldsymbol{w},\rho}(x) = \max_{z\in\mathcal{Z}}\langle \boldsymbol{w}, \Psi(x, z)\rangle + \delta(z) - \rho$, which allows for additional flexibility.

3.1. Dual Optimization Problem

To obtain a dual representation of the Problem 1, we start by equivalently re-writing (P) as

$$\begin{aligned}\min_{\boldsymbol{w}\in\mathcal{H},\rho\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^n} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n} l(\xi_i)\\ \text{s.t.} \quad & \xi_i \geq \rho - \max_{z\in\mathcal{Z}}\big(\langle \boldsymbol{w}, \Psi(x_i, z)\rangle + \delta(z)\big),\ \forall i\end{aligned}$$

Denote, for all $\boldsymbol{\alpha} \in \mathbb{R}^n$ with $\boldsymbol{\alpha} \geq 0$,[1] the Lagrangian by

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}, \rho, \boldsymbol{\xi}, \boldsymbol{\alpha}) \ := \ & \frac{1}{2}\|\boldsymbol{w}\|^2 - \rho + \frac{1}{\nu n}\sum_{i=1}^{n} l(\xi_i) + \sum_{i=1}^{n}\alpha_i\\ & \Big(\rho - \xi_i - \max_{z\in\mathcal{Z}}\big(\langle \boldsymbol{w}, \Psi(x_i, z)\rangle + \delta(z)\big)\Big).\end{aligned}$$

By *weak* duality (e.g., Boyd & Vandenberghe, 2004, Chapter 5),

[1]For vectors $x \in \mathbb{R}^n$, we denote by $\boldsymbol{x} \geq 0$ as the component-wise inequalities $x_i \geq 0$, $i = 1, \ldots, n$.

$$\begin{aligned}\text{Eq. (P)} \ \geq \ & \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\geq 0}\ \min_{\boldsymbol{w}\in\mathcal{H},\rho\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^n} \mathcal{L}(\boldsymbol{w}, \rho, \boldsymbol{\xi}, \boldsymbol{\alpha}) \ =\\ & \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\geq 0}\Bigg(-\frac{1}{\nu n}\sum_{i=1}^{n}\max_{\xi_i\in\mathbb{R}}\Big(\alpha_i\nu n\xi_i - l(\xi_i)\Big)\\ & + \min_{\rho\in\mathbb{R}}\rho\Big(-1 + \sum_{i=1}^{n}\alpha_i\Big)\\ & \underbrace{-\max_{\substack{\boldsymbol{w}\in\mathcal{H},z_i\in\mathcal{Z}\\ i=1,\ldots,n}}\Big(\sum_{i=1}^{n}\alpha_i\big(\langle \boldsymbol{w}, \Psi(x_i, z_i)\rangle + \delta(z_i)\big) - \frac{1}{2}\|\boldsymbol{w}\|^2\Big)}_{(*)}\Bigg)\end{aligned}$$

Let $\boldsymbol{w}^{\boldsymbol{\alpha}}$ and $(z_i^{\boldsymbol{\alpha}})_{i=1,\ldots,n}$ be the maximizing arguments in $(*)$. Thus $\max_{z_i\in\mathcal{Z}}\langle \boldsymbol{w}^{\boldsymbol{\alpha}}, \Psi(x_i, z_i)\rangle + \delta(z_i) = \langle \boldsymbol{w}^{\boldsymbol{\alpha}}, \Psi(x_i, z_i^{\boldsymbol{\alpha}})\rangle + \delta(z_i^{\boldsymbol{\alpha}})$, and, moreover, $\max_{z_i\in\mathcal{Z}}\langle \boldsymbol{w}, \Psi(x_i, z_i)\rangle + \delta(z_i) \geq \langle \boldsymbol{w}, \Psi(x_i, z_i^{\boldsymbol{\alpha}})\rangle + \delta(z_i^{\boldsymbol{\alpha}})$ for all $\boldsymbol{w} \in \mathcal{H}$ and $i = 1, \ldots, n$. Hence, for all $\boldsymbol{\alpha} \in \mathbb{R}^n_+$,

$$(*) \ = \ \max_{\boldsymbol{w}\in\mathcal{H}}\sum_{i=1}^{n}\alpha_i\big(\langle \boldsymbol{w}, \Psi(x_i, z_i^{\boldsymbol{\alpha}})\rangle + \delta(z_i^{\boldsymbol{\alpha}})\big) - \frac{1}{2}\|\boldsymbol{w}\|^2,$$

from which it follows $\boldsymbol{w}^{\boldsymbol{\alpha}} = \sum_{i=1}^{n}\alpha_i\Psi(x_i, z_i^{\boldsymbol{\alpha}})$, and thus

$$(*) = \max_{\substack{z_i\in\mathcal{Z}\\ i=1,\ldots,n}}\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j k\Big((x_i, z_i), (x_j, z_j)\Big) + \sum_{i=1}^{n}\alpha_i\delta(z_i).$$

Hence,

$$\begin{aligned}& \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\geq 0}\ \min_{\boldsymbol{w}\in\mathcal{H},\rho\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^n} \mathcal{L}(\boldsymbol{w}, \rho, \boldsymbol{\xi}, \boldsymbol{\alpha})\\ & = \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\geq 0}\Bigg(-\frac{1}{\nu n}\sum_{i=1}^{n}\max_{\xi_i\in\mathbb{R}}\Big(\alpha_i\nu n\xi_i - l(\xi_i)\Big)\\ & + \min_{\rho\in\mathbb{R}}\rho\Big(-1 + \sum_{i=1}^{n}\alpha_i\Big)\\ & - \max_{\substack{z_i\in\mathcal{Z}\\ i=1,\ldots,n}}\Big(\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j k\Big((x_i, z_i), (x_j, z_j)\Big) + \sum_{i=1}^{n}\alpha_i\delta(z_i)\Big)\Bigg)\\ & \overset{(\dagger)}{=} \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\geq 0, \sum_{i=1}^{n}\alpha_i = 1}\Bigg(-\frac{1}{\nu n}\sum_{i=1}^{n} l^*(\alpha_i\nu n)\\ & - \max_{\substack{z_i\in\mathcal{Z}\\ i=1,\ldots,n}}\Big(\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j k\Big((x_i, z_i), (x_j, z_j)\Big) + \sum_{i=1}^{n}\alpha_i\delta(z_i)\Big)\Bigg)\end{aligned}$$

where for $(\dagger)$ we employ the notion of the Fenchel-Legendre convex conjugate function $f^*(\boldsymbol{a}) := \sup_{\boldsymbol{b}}\langle \boldsymbol{a}, \boldsymbol{b}\rangle - f(\boldsymbol{b})$ (Rifkin & Lippert, 2007) and exploit that the function $\boldsymbol{w} \mapsto \frac{1}{2}\|\cdot\|^2$ is self-conjugated; as well as we observe that $\min_{\rho\in\mathbb{R}}\rho\Big(-1 + \sum_{i=1}^{n}\alpha_i\Big) = 0$ if $\sum_{i=1}^{n}\alpha_i = 1$ and $-\infty$ else-wise, which enforces the constraint $\sum_{i=1}^{n}\alpha_i = 1$ when maximizing with respect to $\boldsymbol{\alpha}$. Thus we obtain the following dual optimization problem of (P).

Problem 2 (DUAL LATENT ANOMALY DETECTION OPTIMIZATION PROBLEM)**.** *Given a monotonically non-decreasing loss function $l : \mathbb{R} \to \mathbb{R}$, and denoting by the $l^* : \mathbb{R} \to \mathbb{R}$ the dual loss function, maximize, with respect to $\boldsymbol{\alpha} \in \mathbb{R}^n$ and subject to $\boldsymbol{\alpha} \geq 0$ and $\sum_{i=1}^n \alpha_i = 1$,*

$$-\min_{\substack{z_i \in \mathcal{Z} \\ i=1,\dots,n}} \left(\frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j k\big((x_i, z_i), (x_j, z_j)\big) + \sum_{i=1}^n \alpha_i \delta(z_i) \right) - \frac{1}{\nu n} \sum_{i=1}^n l^*(\alpha_i \nu n) \,. \tag{D}$$

Remark 1 (Dual complexity)**.** The minimization over $z \in \mathcal{Z}$ can be expanded into slack variables, so the above dual becomes a quadratically constrained program (QCQP) with $n \cdot |\mathcal{Z}|$ many quadratic constraints.

Remark 2 (Prediction function $f(x)$ and estimated density level set $\hat{L}_\nu$)**.** By the above dualization the prediction function can be written as

$$f(x) = \max_{z \in \mathcal{Z}} \left(\sum_{i=1}^n \alpha_i k\big((x_i, z_i^{\boldsymbol{\alpha}}), (x, z)\big) + \delta(z) \right) - \rho.$$

where ρ can be calibrated by line search such that exactly a fraction of $1 - \nu$ training points satisfy $f(x_i) \geq 0$. The corresponding estimated density-level set is given by $\hat{L}_\nu := \{x \in \mathcal{X} : f(x) \geq 0\}$.

3.2. Theoretical Analysis

For the theoretical analysis, we consider a slight variation of latent anomaly detection,

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathcal{H}} \quad & \frac{1}{n} \sum_{i=1}^n l\Big(1 - \max_{z \in \mathcal{Z}} \big(\langle \boldsymbol{w}, \Psi(x_i, z) \rangle + \delta(z) \big)\Big) \\ \text{s.t.} \quad & \|\boldsymbol{w}\| \leq C \,. \end{aligned} \tag{1}$$

For the important choice of $l(t) = \max(0, t)$ studied in Section 4.2, the above reformulation is equivalent to the original problem (P), in the sense for any choice of ν in (P), there exists a choice of $C > 0$ in (1) such that both problems have the same solution in the variable $\boldsymbol{w}$. This is shown in Supplementary Material D.

To analyze (1) theoretically, note that (1) corresponds to performing empirical risk minimization (ERM), $\hat{f} := \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n l(f(x_i))$ over the class $\mathcal{F} := \{f_{\boldsymbol{w}} = \big(x \mapsto 1 - \max_{z \in \mathcal{Z}} (\langle \boldsymbol{w}, \Psi(x, z) \rangle + \delta(z))\big) : \|\boldsymbol{w}\| \leq C\}$. In the following theorem, we show that the solution of (1) has asymptotically the same loss as the theoretically optimal quantity $f^* := \operatorname{argmin}_{f \in \mathcal{F}} E\, l(f(X))$.

Theorem 3 (LATENT ANOMALY DETECTION GENERALIZATION BOUND)**.** *The following generalization bound holds for the latent anomaly detection method (D.1). Let $l : \mathbb{R} \to \mathbb{R}$ be a non-negative and L-Lipschitz continuous loss function. Denote $A := \max_{z \in \mathcal{Z}} |\delta(z)|$ and $B := \max_{x \in \mathcal{X}, z \in \mathcal{Z}} \|\Psi(x, z)\|$. With probability at least $1 - \epsilon$ over the draw of the sample, the generalization error is bounded as:*

$$E\, l(\hat{f}) - E\, l(f^*) \;\leq\; 8L \frac{1 + A + BC\,|\mathcal{Z}|}{\sqrt{n}} + L(1 + A + BC) \sqrt{\frac{2 \log(2/\epsilon)}{n}} \,.$$

Proof. The full proof is shown in supplemental material D. □

Remark 3. While the present analysis considers a worst-case bound that is independent of the structure of the latent space $\mathcal{Z}$, it would be interesting to analyze the bound also for special choices of the joint feature map and discrete loss functions. Such an analysis was presented in Mcallester & Keshet (2011), who showed asymptotic consistency of the update direction of a perception-like structured prediction algorithm.

Remark 4. Note that the requirements on the loss function are, in particular, fulfilled by the loss $l(t) = \max(0, t)$, which is employed both by the one-class SVM and by the proposed hidden Markov anomaly detector that is introduced in Section 4.2 below. Indeed in that case, l is non-negative and Lipschitz continuous with constant $L = 1$.

4. Hidden Markov Anomaly Detection

In this section, we derive the proposed *hidden Markov anomaly detection* (HMAD) methodology that is capable of dealing with sequence data that exhibits latent state structure. We therefore need to settle for an appropriate loss function l and a joint feature map $\Psi(x, z)$.

4.1. Latent One-class SVM

Setting $l(t) := \max(0, t)$, we can derive a latent version of the one-class support vector machine (OC-SVM) (Schölkopf et al., 2001). Contrary to (Lampert & Blaschko, 2009), structures need not to be known. We derive the latent version of the OC-SVM as follows.

Problem 4 (PRIMAL LATENT OC-SVM OPTIMIZATION PROBLEM)**.** *Given the monotonically non-decreasing hinge loss function $l : \mathbb{R} \to \mathbb{R}$, $l(t) = \max(0, t)$, minimize, with respect to $\boldsymbol{w} \in \mathcal{H}$ and $\rho \in \mathbb{R}$,*

$$\frac{1}{2} \|\boldsymbol{w}\|^2 - \rho + \frac{1}{\nu n} \sum_{i=1}^n \max\Big(0, \rho - \max_{z \in \mathcal{Z}} \big(\langle \boldsymbol{w}, \Psi(x_i, z) \rangle + \delta(z) \big)\Big) \,. \tag{P'}$$

It is easy to check that the dual loss of $l(t) = \max(0, t)$ is the function $l^*(t) = 0$ if $0 \leq t \leq 1$ and ∞ else, and thus the corresponding dual optimization problem is as follows.

Problem 5 (DUAL LATENT ONE-CLASS SVM OPTIMIZATION PROBLEM)**.** *Given the monotonically non-decreasing hinge loss function $l : \mathbb{R} \to \mathbb{R}$, $l(t) = \max(0, t)$, and denoting by $l^* : \mathbb{R} \to \mathbb{R}$ the dual hinge loss function, maximize, with respect to $\boldsymbol{\alpha} \in \mathbb{R}^n$ and subject to $0 \leq \boldsymbol{\alpha} \leq \frac{1}{\nu n}$ and $\sum_{i=1}^n \alpha_i = 1$,*

$$- \min_{\substack{z_i \in \mathcal{Z} \\ i=1,\dots,n}} \left(\frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j k\big((x_i, z_i^{\boldsymbol{\alpha}}), (x_j, z_j^{\boldsymbol{\alpha}})\big) - \sum_{i=1}^{n} \alpha_i \delta(z_i^{\boldsymbol{\alpha}}) \right) \tag{D'}$$

4.2. Hidden Markov Anomaly Detection (HMAD)

In hidden Markov anomaly detection, we are interested in inferring the hidden state sequence $z = (z^1, \dots, z^T) \in \mathcal{Z}$, with single entries $z^t \in \mathcal{Y}$, associated with an observed feature sequence $x = (x^1, \dots, x^T)$, i.e., each element of the sequence is a feature vector $x^t = (x_l^t)_{l=1,\dots,d} \in \mathbb{R}^d$. Hidden Markov models have been introduced as a certain class of probability density functions P with chain-like factorization (Rabiner, 1989) and parameters $\boldsymbol{w}$:

$$P(x, z|\boldsymbol{w}) = \pi(x^1|z^1, \boldsymbol{w}) \prod_{t=2}^{T} \big(P(z^t|z^{t-1}, \boldsymbol{w}) P(x^t|z^t, \boldsymbol{w})\big) . \tag{2}$$

Based on the corresponding log-probability and conditioned on the inputs, $\log P(z|x) = \log \pi(z^1|x^1, \boldsymbol{w}) + \sum_{t=2}^{T} \log P(z^t|z^{t-1}, \boldsymbol{w}) + \log P(z^t|x^t, \boldsymbol{w})$, we introduce the matching, decomposable scoring function $G\colon \mathcal{X} \times \mathcal{Z} \times \mathcal{H} \to \mathbb{R}$ with

$$\log P(z|x) = G(x, z, \boldsymbol{w}) = \sum_{t=2}^{T} G^{\text{trans}}(z^t, z^{t-1}, \boldsymbol{w}) + \sum_{t=1}^{T} G^{\text{em}}(x^t, z^t, \boldsymbol{w}), \tag{3}$$

such that $G(x, z, \boldsymbol{w}) \propto \langle \boldsymbol{w}, \Psi(x, z) \rangle$. This motivates defining a joint feature map as follows:

Definition 1 (HIDDEN MARKOV JOINT FEATURE MAP)**.** *Given a feature map $\phi : \mathcal{X} \to \mathcal{F}$, define the* Hidden Markov joint feature map $\Psi : \mathcal{X} \times \mathcal{Z} \to \mathcal{H}$ *as*

$$\Psi(x, z) = \begin{pmatrix} (\sum_{t=2}^{T} \mathbf{1}[z^t = i \wedge z^{t-1} = j])_{i,j \in \mathcal{Y}}, \\ (\sum_{t=1}^{T} \mathbf{1}[z^t = i]\, \phi(x^t))_{i \in \mathcal{Y}} \end{pmatrix} .$$

To better understand the above feature map, observe that the weight vector $\boldsymbol{w} = (\boldsymbol{w}^{\text{em}}, \boldsymbol{w}^{\text{trans}})$ decomposes into a transition vector $\boldsymbol{w}^{\text{trans}} = (\boldsymbol{w}_{i,j}^{\text{trans}})_{i,j \in \mathcal{Y}}$ and an emission vector $\boldsymbol{w}^{\text{em}} = (\boldsymbol{w}_i^{\text{em}})_{i \in \mathcal{Y}}$, so the linear model becomes

$$\langle \boldsymbol{w}, \Psi(x, z) \rangle = \sum_{t=2}^{T} \sum_{i,j \in \mathcal{Y}} \mathbf{1}[z^t = i \ \wedge \ z^{t-1} = j]\, \boldsymbol{w}_{i,j}^{\text{trans}} + \sum_{t=1}^{T} \sum_{i \in \mathcal{Y}} \mathbf{1}[z^t = i] \langle \boldsymbol{w}_i^{\text{em}}, \phi(x^t) \rangle ,$$

which is reminiscent of the log probability associated with HMMs and given by (3).

Definition 2 (HIDDEN MARKOV ANOMALY DETECTION (HMAD))**.** Hidden Markov anomaly detection (HMAD) *is defined as the latent OC-SVM (Problem 4 and 5) together with the hidden Markov joint feature map (Definition 1).*

Note that thus, because of the specific form of the joint feature map occuring in HMAD, the problem of maximizing over the latent variables in Eqn. (P′) can be solved by finding the most probable state sequence of the corresponding hidden Markov model, which can be efficiently computed using, e.g., Viterbi's algorithm (Rabiner, 1989).

4.3. Properties

Similar to its non-structured counterpart, the structured one-class SVM enjoys interesting properties, as we show below. Recall that for an input x and prediction function f the following cases can occur:

1. $f(x) > 0$ (then x is strictly inside the density level set)
2. $f(x) = 0$ (then x is right at the boundary of the set)
3. $f(x) < 0$ (then x is outside of the density level set, i.e., x is an outlier)

The following theorem shows that the parameter ν controls the number of outliers.

Theorem 6. *The following statements hold for the structured one-class SVM and the induced decision function f:*

(a) The fraction of outliers (inputs x_i with $f(x_i) < 0$) is upper bounded by ν.
(b) The fraction of inputs lying strictly inside the density level set (inputs x_i with $f(x_i) > 0$) is upper bounded by $1 - \nu$.

The theorem is proven in Appendix E and shows that the quantity ν can be interpreted as the fraction of outliers predicted by the learning algorithm. In particular this shows, together with theoretical analysis of Section 3.2, that for well behaved problems (where there is no probability mass exactly on the decision region and where the true decision boundary is contained in the hypothesis set, e.g., via the use of universal kernels (Steinwart & Christmann, 2008)), the estimated density level set $\hat{L}_\nu$ asymptotically equals the truly underlying density level set L_ν: $P(\hat{L}_\nu \backslash L_\nu \cup L_\nu \backslash \hat{L}_\nu) \to 0$ for $n \to \infty$.

4.4. Optimization Algorithm

A first difficulty occurring when trying to solve the optimization problem (P′) consists in the function $g : (\boldsymbol{w}, \rho) \mapsto \rho - \max_{z \in \mathcal{Z}} \big(\langle \boldsymbol{w}, \Psi(x_i, z) \rangle + \delta(z) \big)$, which is concave and thus renders the optimization problem non-convex. However, note that any concave function $h : \mathbb{R} \to \mathbb{R}$ can be decomposed into convex and concave parts, $\max(0, h(x)) = \max(0, -h(x)) + h(x)$. Hence, putting $g(\boldsymbol{w}, \rho) = \rho - \max_{z \in \mathcal{Z}} \big(\langle \boldsymbol{w}, \Psi(x_i, z) \rangle + \delta(z)$, we can write Eq. (P′) $= \frac{1}{2} \|\boldsymbol{w}\|^2 - \rho + \frac{1}{\nu n} \sum_{i=1}^{n} \Big(\max\big(0, -g(\boldsymbol{w}, \rho)\big) + g(\boldsymbol{w}, \rho) \Big)$. The above decomposition consists of a convex term followed by a concave term, which admits the optimization framework of DC programming (difference of convex functions) (Tao & An, 1998). Although the function $-g$ is not differentiable, it admits, at any point $(\boldsymbol{w}_0, \rho_0) \in \mathcal{H} \times \mathbb{R}$, a subdifferential

$$\partial_{(\boldsymbol{w}_0,\rho_0)} g(\boldsymbol{w}_0,\rho_0) := \{\boldsymbol{v} \in \mathcal{H}\times\mathbb{R} : g(\boldsymbol{w},\rho) - g(\boldsymbol{w}_0,\rho_0) \geq \langle \boldsymbol{v}, (\boldsymbol{w},\rho) - (\boldsymbol{w}_0,\rho_0)\rangle, \forall (\boldsymbol{w},\rho) \in \mathcal{H}\times\mathbb{R}\}.$$

One can verify—using the sub-differentiability of the maximum operator—that, for any $z \in \mathcal{Z}$, the point $(\Psi(x_i, z), -1)$ is contained in the subdifferential $\partial_{(\boldsymbol{w}_0,\rho_0)} g(\boldsymbol{w}_0,\rho_0)$. Thus, we can linearly approximate, for any $z \in \mathcal{Z}$, via $g(x) \approx \langle \boldsymbol{w}, \Psi(x_i, z)\rangle + \delta(z) - \rho$. In the optimization algorithm we will thus construct a sequence of variables $(\boldsymbol{w}^t, \rho^t, z^t)$, $t = 1, 2, 3, \ldots$, where we use this approximation with z chosen as $z^t = \operatorname{argmax}_{z\in\mathcal{Z}} \left\langle \boldsymbol{w}^{t-1}, \Psi(x_i, z)\right\rangle + \delta(z)$, where $\boldsymbol{w}^{t-1}$ is conveniently computed by solving a regular one-class SVM problem. The resulting optimization algorithm is described in Algorithm 1.

Algorithm 1 Hidden Markov Anomaly Detection

input data $x_1, \ldots, x_n$
put $t = 0$ and initialize $\boldsymbol{w}^t$ (e.g., randomly)
repeat
  t:=t+1
  for $i = 1, \ldots, n$ **do**
    $z_i^t := \operatorname{argmax}_{z\in\mathcal{Z}} \langle \boldsymbol{w}^{t-1}, \Psi(\boldsymbol{x}_i, z)\rangle + \delta(z)$
    (i.e. use Viterbi algorithm)
  end for
  let $(\boldsymbol{w}^t, \rho^t)$ be the optimal arguments when solving one-class SVM with $\phi(\boldsymbol{x}_i) := \Psi(\boldsymbol{x}_i, z_i^t)$
until $\forall\, i = 1, \ldots, n:\ z_i^t = z_i^{t-1}$
Return optimal model parameters $\boldsymbol{w} := \boldsymbol{w}^t$, $\rho = \rho^t$, and $z_i := z_i^t \quad \forall\, i = 1, \ldots, N$

Despite the non-convex nature of the optimization problem, we found in our experiments that the algorithm tends to converge often faster than the standard column-generation approach of the supervised structured SVM (Tsochantaridis et al., 2005a), since no storage of constraints is necessary, which in turn leads to constant time and space complexity for each iteration of Algorithm 1.

5. Empirical Analysis

We conducted experiments for the scenario of label sequence learning where we have full access to the ground truth as well as two real-world scenarios from computational biology and computational sustainability. Our interest is to assess the anomaly *detection* performance of our hidden Markov anomaly detection (HMAD) method. As baseline methods that excel in one-class classification settings, we chose one-class support vector machines (OC-SVM) with appropriate kernels. For initialization, we randomly choose a vector $\boldsymbol{w}^0$ for each run of our algorithm which is sufficient, since no initialization of structures is needed, as those are deduced from the parameter vector.

5.1. Controlled Experiment

For the controlled experiments, we aim to gain insights into the behavior of our method. We investigate the anomaly detection performance for low to very high (up to 30%) fraction of anomalies. Furthermore, we are interested in the anomaly detection performance for an increasing amount of disorganization in the input sequences. Since HMAD exploits latent structure, it is not clear how it performs when less structure is present. Vanilla OC-SVMs does not exploit latent dependencies and should be unaffected by this. Additionally, we are interested in the runtime behavior for various training set sizes.

We generated Gaussian noise sequences of length 600 with unit variance for the nominal bulk of the data. Non-trivial anomalies (see Fig. 3) were induced as blocks of Gaussian noise with non-zero mean and a total, cumulative length of 120 per anomalous example. We vary either the fraction of anomalies in the training data set or the number of blocks, depending on the amount of structure that is modeled into the data (see Figure 3: from 120 sub-blocks of length 1 (100% disorganization) to a single block of length 120 (0% disorganization). We employ a binary state model consisting of 2 states and 4 possible transitions with an constant prior $\delta(\cdot)$. We report on the average area under the ROC curve (AUC) for the anomaly detection performance over 50 repetitions of the experiment. Since we know the underlying ground truth we can exactly compute the Bayes classifier,[2] which in our case lies within the set of linear classifiers, and serves as a hypothetical upper performance bound for the maximal achievable detection performance.

We compare the detection performance of our method to the one achieved by OC-SVMs with RBF kernels, histogram kernels, and linear kernels using l_1- and l_2-feature normalization, and optimal kernel parameters (1.0 for the RBF kernel, 8 for the histogram kernel, and l_1 for the linear kernel). The results of the anomaly detection experiment are shown in Figure 2 (left and center). As can be seen in the figure, our method achieves tremendously higher detection rates than the OC-SVMs using linear or RBF kernel, which perform similar bad as random guessing. Most competitive baseline methods are OC-SVMs with histogram kernels and optimal bin size (8 bins). There exists a strong relation between our method HMAD and Fisher kernels (Jebara et al., 2004) in the sense, that the same representation is used. Unlike Fisher kernels, our methodology includes the parameter optimization procedure, and therefore, given the same model parameters both methods are on par. For a more detailed comparison we refer the reader to Appendix B. Remarkably, our method achieves stable on-par performance with the Bayes classifier for all levels of disorganization, even when there is no structure to be exploited in the data (see Figure 2 center) and outperforms significantly all competitors for varying fraction of anomalies (see Figure 2 left).

[2] For data that is i.i.d. realized from a distribution (which is the case in our synthetic experiment), the Bayes classifier is defined as the classifier achieving the maximal accuracy among all measurable functions.

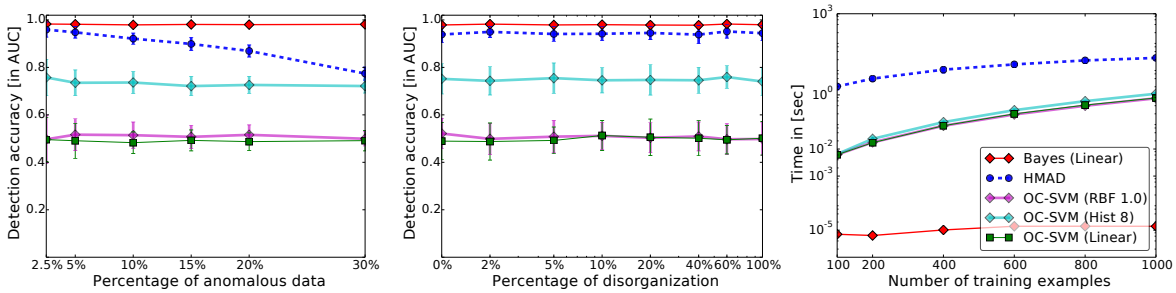

Figure 2. Results for the controlled experiment: (left) anomaly detection performance for various fractions of anomalies in the training set, (center) anomaly detection performance for increasing amount of disorganization, and (right) runtime behavior. All settings show results for our hidden Markov anomaly detection (HMAD) as well as a set of competitors (using optimal kernel parameters). Noticeable, the detection performance of HMAD is not affected by increasing amounts of disorganization in the input data (center).

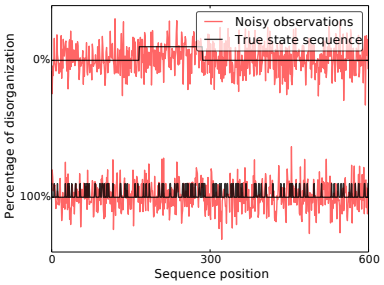

Figure 3. Examples of observation sequences for two extreme cases of our controlled experiments: even in the easy setting (top), the true state sequence is barely visible to the naked eye in the noisy observed sequence, while in the challenging setting (bottom) it is almost impossible for humans to extrapolate the truly underlying state sequence.

Exemplary, we depict two typical anomalous observation sequences of length 600 and anomalous block length 120 of the experiment in Fig. 3 for the 0% (top) and 100% disorganization (bottom) settings. As can be seen, anomalies are not trivially detectable. We also conducted runtime experiments (Fig. 2 right) to compare the runtime of our method HMAD against that of the baseline methods. We used the same two-state model as in the previous controlled experiment, but with training set size varying from 100 to 1000 examples. We used a fraction 10% of anomalies to ensure there is a sufficient number of anomalies in the data. As expected, absolute computational runtime is higher than for vanilla OC-SVMs. This is due to the iterative approach that includes Viterbi decoding of the sequences and solving a vanilla OC-SVM in each step. However, computational complexity grows with increasing number of examples comparable to OC-SVM which gives a total complexity of $\mathcal{O}(\text{OC-SVM}) + \mathcal{O}(c)$, where c is a constant.

5.2. Bioinformatics Application: Procaryotic Gene Prediction

In prokaryotes (mostly bacteria and archaea) gene structures consist of the protein coding region that starts by a start codon (one out of three specific 3-mers in many prokaryotes) followed by a number of codon triplets (of three nucleotides each) and is terminated by a stop codon (one out of five specific 3-mers in many prokaryotes) (Alberts et al., 2002). Genic regions are first transcribed to RNA and then translated into a protein. Since genes are separated from one another by intergenic regions, the problem of identifying genes can be posed as a label sequence learning task, were one assigns a label (out of intergenic, start, stop, exonic) to each position in the genome (Schweikert et al., 2009).

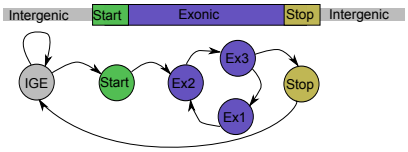

Figure 4. State model of prokayotic gene finding.

We downloaded the genome of the widely studied *escherichia coli* bacteria, which is publicly available.[3] Genomic sequences were cut between neighboring genes (splitting intergenic regions equally), such that a minimum distance of 6 nucleotides between genes was maintained. Intergenic regions have a minimum distance of 50 nucleotides to genic regions. Features were derived from the nucleotide sequence by transcoding it to a numerical representation of triplets. All examples have a minimum length of 500 nucleotides and do not exceed 1200 nucleotides.

For the OC-SVM we use matching spectrum kernels of order 1,2, and 3 (resp. 64, 4160, and 266.304 dimensions), while the SSVM and HMAD obtain a sequence of binary entries as input data. A description of the used state model, which is based on Görnitz et al. (2011), is given in Figure 4. Start and stop states use corresponding features that encode start and stop codons. Any other states is using all 64 binary input features. Furthermore, we choose $\delta(z)$ to have a slightly higher probability towards the intergenic state. For a more fair comparison, OC-SVM and HMAD are given the true fraction of anomalies which varies from 2.5% up to 30%. The training set contained 200 examples of intergenic and genic examples with a total length of >170.000 nucleotides, while the testing set contained 350 intergenic and 50 genic examples of length >330.000 nucleotides, rending this a computationally challenging experiment. The experiment was repeated 20 times where training and test set are drawn randomly.

[3] http://www.sanger.ac.uk... .../resources/downloads/bacteria/escherichia-coli.html

We further employ a simple feature selection procedure where the 8 most distinctive genic- and intergenic features are selected on a comparable labeled procaryote (*e. fergusonii*), which increased performance for OC-SVM by more than 10%. While performance for our HMAD remained unchanged, training and prediction times dropped down to 15% when compared to the full model.

The results in Figure 5 show a vastly superior performance of our method (HMAD) in terms of the detection accuracy: HMAD achieves a perfect AUC of 1.00 (which means: it exactly identifies every sequence containing a gene with zero error) for all outlier fractions, while the classical one-class SVM shows much worse performance with an AUC of 0.85 at best and 0.66 in the worst case. Using higher order spectrum kernels increases the detection performance only marginally. This result is remarkable as it has been reported that string kernels such as spectrum kernel achieve state of the art performance in this application (Schweikert et al., 2009).

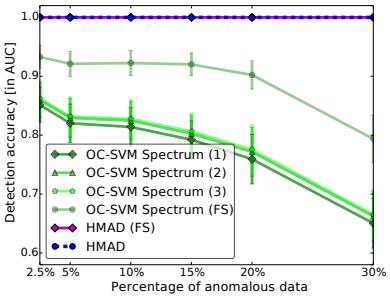

Figure 5. Detection performance for various fractions of outliers in terms of AUC for the procaryotic gene finding experiment. Clearly, the accuracy of our hidden Markov anomaly detection exceeds the vanilla one-class SVM performance even when using higher order (1,2 & 3 codons = 64, 4160 and 266.304 dimensions) spectrum kernels.

5.3. Computational Sustainability Application: Anomalous State Detection in Wind Turbines

In this anomaly detection task, the target objective is to discriminate between two different wind turbine states depending on the weather conditions. Such applications are important, for example, to monitor machines for failures or changes in underlying system state (Zaher et al., 2009). We used the wind turbine simulator FAST (Jonkman et al., 2005) to generate simulated sensor readings. The weather conditions, i.e., wind speed and turbulence are modeled by the wind turbulence simulator TurbSim (Jonkman & Buhl, 2012). We used 200 nominal and anomalous sequences of length 800, consisting of 5 time series of sensor data each. Nominal data consisted of a single wind speed and perturbation class setting, while the anomalous data contained a block of differing wind speed and perturbation class. From this data we selected half for training with various anomalous data fraction and the remaining for testing. The OC-SVM employs histogram kernels with 4, 8, and 16 bins and all methods are given the true fraction of outliers. As can be seen in Fig. 6 the detection performance of our method

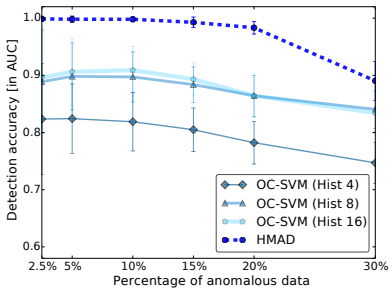

Figure 6. Detection performance for various fractions of outliers in terms of AUC for the computational sustainability experiment. Clearly, the accuracy of our hidden Markov anomaly detection exceeds the vanilla one-class SVM performance with histogram kernels albeit detection performance deteriorates with increasing amount of outliers in the training set.

vastly outperforms all OC-SVMs, which is the previously best known performing method on this data. Detection performances for all methods decrease with increasing amount of anomalies.

6. Conclusion

We proposed a novel methodology for latent anomaly detection on hidden Markov models, which combines ideas from structured output learning and kernel-based anomaly detection. Theoretical guarantees in the form of generalization error bounds underlie the proposed general latent anomaly detection framework, which we optimized using a DC approach. We empirically analyzed a specific instantiation of our approach, *hidden Markov anomaly detection (HMAD)*, on controlled artificial and real data from the domains of bioinformatics and computational sustainability.

The results show that the proposed HMAD significantly outperforms the original one-class SVM on real output sequence data. For gene finding, an increasingly important application where a large amount of pre-knowledge is incorporated, we showed that we can achieve a perfect detection rate (1.00 AUC), substantially outperforming the vanilla one-class SVM (0.66 AUC at 30%). Similar, for the studied computational energy sustainability application, the proposed method achieved almost optimal accuracy (>0.99 AUC), while the regular one-class SVM achieved only 0.92 AUC at best.

Finally and importantly, note that our approach is neither restricted to hidden Markov models nor to the setting of anomaly detection; it can be extended to tree- or graph-structured joint feature maps and to clustering and dimensionality reduction (e.g., hidden Markov PCA). A principal analysis of this general framework will be presented in forthcoming publications.

7. Acknowledgments

MK acknowledges support by the German Research Foundation through the grant KL 2698/2-1. NG was supported by BMBF ALICE II grant 01IB15001B.

References

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J.D. *Molecular Biology of the Cell*. Garland, 4th edition, 2002.

Bartlett, P.L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, November 2002.

Blanchard, Gilles, Lee, Gyemin, and Scott, Clayton. Semi-supervised novelty detection. *JMLR*, 11:2973–3009, 2010.

Boyd, Stephan and Vandenberghe, Lieven. *Convex Optimization*. Cambrigde University Press, 2004.

Chandola, Varun, Banerjee, Arindam, and Kumar, Vipin. Anomaly detection: A survey. *ACM Comput. Surv.*, 41 (3):1–58, 2009.

Görnitz, Nico, Kloft, Marius, and Brefeld, Ulf. Active and semi-supervised data domain description. In *Machine Learning and Knowledge Discovery in Databases*, pp. 407–422. Springer Berlin Heidelberg, 2009a.

Görnitz, Nico, Kloft, Marius, Rieck, Konrad, and Brefeld, Ulf. Active learning for network intrusion detection. In *Proceedings of the 2nd ACM workshop on Security and artificial intelligence*, pp. 47–54. ACM, 2009b.

Görnitz, Nico, Widmer, Christian, Zeller, Georg, Kahles, Andre, Sonnenburg, Soeren, and Raetsch, Gunnar. Hierarchical multitask structured output learning for large-scale sequence segmentation. In *NIPS*, pp. 1–10, 2011.

Görnitz, Nico, Widmer, Christian, Zeller, Georg, Kahles, André, Sonnenburg, Sören, and Rätsch, Gunnar. Hierarchical multitask structured output learning for large-scale sequence segmentation. In *NIPS*, 2011.

Görnitz, Nico, Kloft, Marius, Rieck, Konrad, and Brefeld, Ulf. Toward supervised anomaly detection. *Journal of Artificial Intelligence Research (JAIR)*, 46:235–262, 2013.

Görnitz, Nico, Porbadnigk, Anne K, Binder, Alexander, Sannelli, Claudia, Braun, Mikio, Müller, Klaus-Robert, and Kloft, Marius. Learning and evaluation in presence of non-iid label noise. In *AISTATS 2014*, pp. 293–302, 2014.

Jebara, Tony, Kondor, Risi, and Howard, Andrew. Probability product kernels. *JMLR*, 2004.

Joachims, Thorsten, Hofmann, Thomas, Yue, Yisong, and Yu, Chun-Nam. Predicting structured objects with support vector machines. *Communications of the ACM*, 52 (11):97–104, 2009.

Jonkman, B. J. and Buhl, M. L. *Turbsim User's Guide: Version 1. 50*. Nat. Renew. Energy Laboratory, 2012. ISBN 9781234541484. URL http://books.google.de/books?id=NAfOygAACAAJ.

Jonkman, J. M., Buhl, M. L., of Energy. Office of Energy Efficiency, United States. Dept., and Energy, Renewable. *FAST User's Guide: Updated August 2005*. NREL/TP. National Renewable Energy Laboratory, 2005. URL http://books.google.de/books?id=V_9pNwAACAAJ.

Jyothsna, V., Prasad, V. V. Rama, and Prasad, K. Munivara. Article: A review of anomaly based intrusion detection systems. *International Journal of Computer Applications*, 28(7):26–35, August 2011.

Kloft, M., Brefeld, U., Sonnenburg, S., and Zien, A. ℓ_p-norm multiple kernel learning. *Journal of Machine Learning Research*, 12:953–997, Mar 2011.

Kloft, Marius and Laskov, Pavel. Online anomaly detection under adversarial impact. 2011.

Kloft, Marius and Laskov, Pavel. Security analysis of online centroid anomaly detection. *J. Mach. Learn. Res.*, 13(1):3681–3724, December 2012. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2503308.2503359.

Kloft, Marius, Brefeld, Ulf, Düessel, Patrick, Gehl, Christian, and Laskov, Pavel. Automatic feature selection for anomaly detection. In *Proceedings of the 1st ACM workshop on Workshop on AISec*, pp. 71–76. ACM, 2008.

Kukita, Yoji, Uchida, Junji, Oba, Shigeyuki, Nishino, Kazumi, Kumagai, Toru, Taniguchi, Kazuya, Okuyama, Takako, Imamura, Fumio, and Kato, Kikuya. Quantitative identification of mutant alleles derived from lung cancer in plasma cell-free dna via anomaly detection using deep sequencing data. *PLoS One*, 8(11):e81468, 2013.

Lampert, Christoph H. and Blaschko, Matthew B. Structured prediction by joint kernel support estimation. *Machine Learning*, 77(2-3):249–269, 4 2009. doi: 10.1007/s10994-009-5111-0.

Markou, M. and Singh, S. Novelty detection: a review – part 1: statistical approaches. *Signal Processing*, 83: 2481–2497, 2003.

Mcallester, David and Keshet, Joseph. Generalization bounds and consistency for latent structural probit and ramp loss. *Nips*, pp. 1–8, 2011.

Mohri, Mehryar, Rostamizadeh, Afshin, and Talwalkar, Ameet. *Foundations of Machine Learning*. The MIT Press, 2012. ISBN 026201825X, 9780262018258.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. An introduction to kernel-based learning algorithms. *IEEE Neural Networks*, 12(2):181–201, May 2001.

Müller, K R, Mika, S, Rätsch, G, Tsuda, K, and Schölkopf, B. An introduction to kernel-based learning algorithms. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 12(2):181–201, 1 2001. doi: 10.1109/72.914517.

Noto, Keith, Brodley, Carla E., Majidi, Saeed, Bianchi, Diana W., and Slonim, Donna K. Csax: Characterizing systematic anomalies in expression data. In *RECOMB 18*, pp. 222–236, 2014.

Nowozin, Sebastian and Lampert, Christoph H. Structured learning and prediction in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 6(3-4): 185–365, 2010. doi: 10.1561/0600000033.

Rabiner, Lawrence R. A tutorial on hidden markov models and selected appications in speech recognition. *Proceedings of the IEEE*, 77(2), 1989.

Rätsch, Gunnar and Sonnenburg, Sören. Large scale hidden semi-markov svms. In Schölkopf, B, Platt, J, and Hoffman, T (eds.), *NIPS*, pp. 1161–1168. MIT Press, 2007.

Rieck, Konrad, Krueger, Tammo, Brefeld, Ulf, and Müller, Klaus-Robert. Approximate tree kernels. *JMLR*, 11: 555–580, 2010.

Rifkin, Ryan M. and Lippert, Ross A. Value regularization and fenchel duality. *JMLR*, 8:441–479, 2007.

Saligrama, Venkatesh and Zhao, Manqi. Local anomaly detection. In *AISTATS 2012*, pp. 969–983, 2012.

Schölkopf, B. and Smola, A.J. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A.J., and Williamson, R.C. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7): 1443–1471, 2001.

Schweikert, Gabriele, Zien, Alexander, Zeller, Georg, Behr, Jonas, Dieterich, Christoph, Ong, Cheng Soon, Philips, Petra, De Bona, Fabio, Hartmann, Lisa, Bohlen, Anja, et al. mgene: accurate svm-based gene finding with an application to nematode genomes. *Genome research*, 2009.

Steinwart, Ingo and Christmann, Andreas. *Support Vector Machines*. Springer, 1st edition, 2008. ISBN 0387772413.

Tao, Pham Dinh and An, Le Thi Hoai. A DC optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8(2):476–505, 1998.

Tax, David M.J. and Duin, Robert P.W. Support vector data description. *Machine Learning*, 54:45–66, 2004.

Tsochantaridis, Ioannis, Joachims, Thorsten, Hofmann, Thomas, and Altun, Yasemin. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, 6:1453–1484, December 2005a. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1046920.1088722.

Tsochantaridis, Ioannis, Joachims, Thorsten, Hofmann, Thomas, and Altun, Yasemin. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005b.

Tsybakov, A.B. On nonparametric estimation of density level sets. *Annals of Statistics*, 25:948–969, 1997.

Zaher, ASAE, McArthur, SDJ, Infield, DG, and Patel, Y. Online wind turbine fault detection through automated scada data analysis. *Wind Energy*, 12(6):574–593, 2009.

Supplementary Material

A. Comparison to Structured SVMs (SSVM)

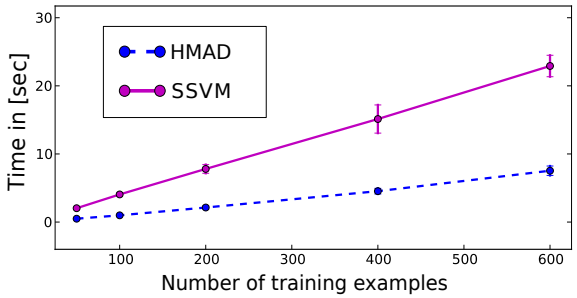

Figure A.1. Without the need for constraint generation, our hidden Markov anomaly detection easily outperforms the structured SVM.

We report run time comparisons of the structured output SVM (SSVM) and our hidden Markov anomaly detection in the same setting as in the controlled experiment in Section 5.1 in Figure A.1. Since the HMAD does not need to add constraints in each iteration, it easily outperforms the SSVM. However, it does require multiple iterations that include Viterbi decoding as well as solving a vanilla one-class SVM and therefore is slower than the OC-SVM (for a comparison see Fig. 2).

B. Comparison to Fisher Kernels

Fisher kernels (Jebara et al., 2004) have been proposed as a way of incorporating graphical models into the framework of kernel-based learning (Müller et al., 2001) and therefore benefit from the vast amount of kernel machines. A practical Fisher kernel is defined as the gradient of the log-likelihood of the probabilistic model with respect to its model parameters.

There is a strong connection of Fisher kernels and our HMAD, in the sense, that we use the same representation of graphical models. However, our method HMAD includes the parameter optimization procedure. Specifically, given the same model parameters learned by our method, the corresponding Fisher kernel employed in an one-class SVM leads to the same solution. Of course, learning the right model parameter is the key to good performance.

To cope with a variety of parameter learning settings and hence, have a realistic comparison against multiple parameter estimation methodologies for Fisher kernels, we use the very same model as in Section 4.2 and derive an upper and a lower bound for the maximum likelihood estimation for Fisher kernels. Here, a lower bound can be easily obtained by using random model parameters, whereas an upper bound uses the *ground truth* latent states information for parameter estimation.

The results in Fig. B.1 and Fig. B.2 show the range of possible solutions for the Fisher kernel (gray area) with the upper bound (red) and (unsurprisingly unstable) lower bound (magenta), in the same setting as in Section 5.1. Moreover, it shows that our method HMAD performs nearly as good as the upper bound in *absence* of any label information.

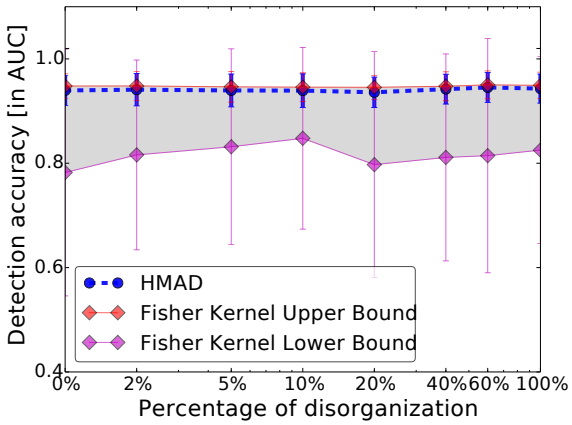

Figure B.1. Comparison for an increasing amount of disorganization of our method HMAD (blue) against a variety of Fisher kernels (gray area), including a lower bound (magenta) based on random model parameters and an upper bound (red) that was trained on *ground truth* data.

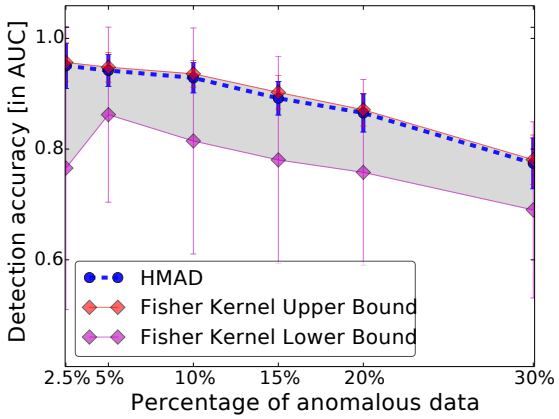

Figure B.2. Comparison for an increasing amount of anomalies of our method HMAD (blue) against a variety of Fisher kernels (gray area), including a lower bound (magenta) based on random model parameters and an upper bound (red) that was trained on *ground truth* data.

C. Sensibility to Number of Hidden States

To assess the stability of the found solution, we did experiments with an increasing number of hidden states for our proposed method HMAD in the same setting as in Section 5.1. The results in Fig. C.1 show, that our method is not sensible to the number of hidden states.

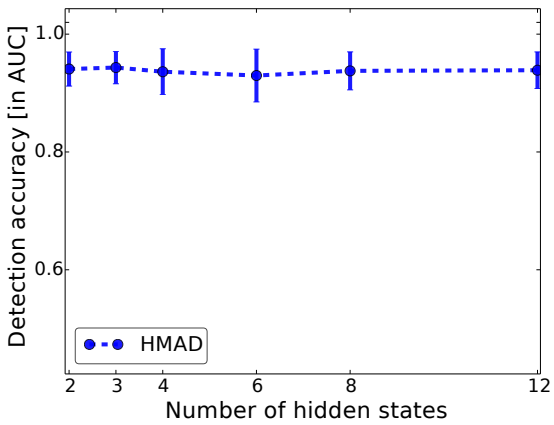

Figure C.1. Performance evaluation for an increasing number of hidden states of our method HMAD (blue).

D. Proofs of Results in Section 3.2

We show the equivalence of (1) and (P) for loss $l(t) = \max(0, t)$.

(P′)

Proof of equivalence of (1) *and* (P) *for* $l(t) = \max(0, t)$. First note that for loss $l(t) = \max(0, t)$ the problem (1) becomes the structured one-class SVM problem (P′) from Section 4.2. To see that (1) is equivalent to (P′), we employ a variable substitution $\tilde{\boldsymbol{w}} := \boldsymbol{w}/\rho^*$ in (1). This yields

$$\text{Eq. (P}') \;=\; -\rho^* + \rho^* \min_{\tilde{\boldsymbol{w}} \in \mathcal{H}} \left(\frac{1}{2} \|\tilde{\boldsymbol{w}}\|^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \max\left(0, 1 - \max_{z \in \mathcal{Z}} \langle \tilde{\boldsymbol{w}}, \Psi(x_i, z) \rangle + \delta(z)' \right) \right), \tag{D.1}$$

where $\delta(z)' = \delta(z)/\rho^*$ and ρ^* is optimal in (P′). Thus, in order to solve (D.1) (and thus (P′)), it is sufficient to solve

$$\min_{\boldsymbol{w} \in \mathcal{H}} \; \frac{1}{2} \|\boldsymbol{w}\|^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \max\Big(0, 1 - \max_{z \in \mathcal{Z}} \langle \boldsymbol{w}, \Psi(x_i, z) \rangle + \delta(z) \Big). \tag{D.2}$$

By Lemma 1 below, for each $\nu \in]0, 1]$, there exists a $C > 0$ such that (D.2) is, indeed, equivalent to (1). □

Lemma 1. *Let $D \subset \mathbb{R}^d$ be a set, let $f, g : D \to \mathbb{R}$ be arbitrary functions. Consider the optimization tasks*

$$\min_{\boldsymbol{x} \in D} \quad f(\boldsymbol{x}) + \sigma g(\boldsymbol{x}), \tag{D.3}$$

$$\min_{\boldsymbol{x} \in D : g(\boldsymbol{x}) \leq \tau} \quad f(\boldsymbol{x}). \tag{D.4}$$

Assume that the minima exist. Then we have that for each $\sigma > 0$ there exists $\tau > 0$ such that OP (D.3) *is equivalent to OP* (D.4)*, that is, each optimal solution x^* of one is an optimal solution of the other, and vice versa.*

Proof. The proof is similar to the one of Proposition 12 in (Kloft et al., 2011). Let be $\sigma > 0$ and $\boldsymbol{x}^*$ be the optimal of (D.3). We have to show that there exists a $\tau > 0$ such that $\boldsymbol{x}^*$ is optimal in (D.4). We set $\tau = g(\boldsymbol{x}^*)$. Suppose $\boldsymbol{x}^*$ is not optimal in (D.4), that is, it exists $\tilde{\boldsymbol{x}} \in D : g(\tilde{\boldsymbol{x}}) \leq \tau$ such that $f(\tilde{\boldsymbol{x}}) < f(\boldsymbol{x}^*)$. Then we have

$$f(\tilde{\boldsymbol{x}}) + \sigma g(\tilde{\boldsymbol{x}}) < f(\boldsymbol{x}^*) + \sigma \tau,$$

which by $\tau = g(\boldsymbol{x}^*)$ translates to

$$f(\tilde{\boldsymbol{x}}) + \sigma g(\tilde{\boldsymbol{x}}) < f(\boldsymbol{x}^*) + \sigma g(\boldsymbol{x}^*).$$

This contradicts the optimality of $\boldsymbol{x}^*$ in (D.3), and hence shows that $\boldsymbol{x}^*$ is optimal in (D.4), which was to be shown. □

Proof of Theorem 3. By (Bartlett & Mendelson, 2002) we have that, if l is L-Lipschitz and ranges in $[0, D]$, with probability at least $1 - \epsilon$ over the draw of the sample,

$$E\, l(\hat{f}) - E\, l(f^*) \leq 8 L R_n(\mathcal{F}) + \frac{l(0)}{n} + D \sqrt{\frac{2 \log(2/\epsilon)}{n}}, \tag{D.5}$$

where $R_n(\mathcal{F}) := \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i)$ is the *Rademacher complexity* of the class $\mathcal{F}$ and $\sigma_1, \ldots, \sigma_n$ denote i.i.d. Rademacher variables (random signs). For many learning algorithms $R_n(\mathcal{G})$ is of the order $O(1/\sqrt{n})$, when employing appropriate regularization, and thus so is (D.5). We will show that also the latent anomaly detection method of (1) enjoys this favorable rate, too: By definition of the Rademacher complexity of $\mathcal{F}$,

$$\begin{aligned} R_n(\mathcal{F}) \;&=\; E \max_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i) \\ &=\; E \max_{\boldsymbol{w} \in \mathcal{H} : \|\boldsymbol{w}\| \leq C} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \Big(1 - \max_{z \in \mathcal{Z}} \big(\langle \boldsymbol{w}, \Psi(X_i, z) \rangle + \delta(z) \big) \Big) \\ &=\; \underbrace{\Big(1 + \max_{z \in \mathcal{Z}} |\delta(z)|\Big) E\left[\left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \right| \right]}_{(*)} \\ &\quad + \underbrace{E \max_{\boldsymbol{w} \in \mathcal{H} : \|\boldsymbol{w}\| \leq C} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \max_{z \in \mathcal{Z}} \langle \boldsymbol{w}, \Psi(X_i, z) \rangle}_{(**)} \end{aligned}$$

We bound the two summands in the above expression separately: on one hand, by Jensen's inequality, $E\left|\frac{1}{n} \sum_{i=1}^{n} \sigma_i\right| \leq \sqrt{E \frac{1}{n^2} \sum_{i,j=1}^{n} \sigma_i \sigma_j} = \frac{1}{\sqrt{n}}$ because $E \sigma_i \sigma_j = 0$ when $i \neq j$, which shows

$(*) \leq \frac{1+A}{\sqrt{n}}$. To bound the second summand, note that $(**) \leq R_n(\mathcal{F}')$ with $\mathcal{F}'$ defined as $\mathcal{F}' := \left\{ f_{\boldsymbol{w}} = \left(x \mapsto \max_{z \in \mathcal{Z}} \langle \boldsymbol{w}, \Psi(x,z) \rangle \right) : \|\boldsymbol{w}\| \leq C \right\}$. Furthermore put $\mathcal{F}'' := \left\{ f_{\boldsymbol{w}} = \left(x \mapsto \max_{z \in \mathcal{Z}} f_z \right) : f_z \in \mathcal{F}_z, z \in \mathcal{Z} \right\}$ and $\mathcal{F}_z := \left\{ f_{\boldsymbol{w}} = \left(x \mapsto \langle \boldsymbol{w}, \Psi(x,z) \rangle \right) : \|\boldsymbol{w}\| \leq C \right\}$. Clearly, $\mathcal{F}' \subset \mathcal{F}''$ and thus $R_n(\mathcal{F}') \leq R_n(\mathcal{F}'')$. By Lemma 2 in the supplemental material, $R_n(\mathcal{F}'')$ is itself bounded by $R_n(\mathcal{F}'') \leq \sum_{z \in \mathcal{Z}} R_n(\mathcal{F}_z)$, and the terms $R_n(\mathcal{F}_z)$, for each $z \in \mathcal{Z}$ are known from (Bartlett & Mendelson, 2002) to be bounded as $R_n(\mathcal{F}_z) \leq \frac{B}{\sqrt{n}}$.[4] This shows $(**) \leq \frac{BC|\mathcal{Z}|}{\sqrt{n}}$. The result is then obtained from (D.5) by noting, that D can be chosen as $D := L(1+A+BC)$. □

In the proof of Theorem 3 above, we use the following result.

Lemma 2 (Lemma 8.1 in (Mohri et al., 2012))**.** *Let $\mathcal{F}_1, \ldots, \mathcal{F}_l$ be sets of functions $f : \mathcal{X} \to \mathbb{R}$, 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 (Mohri et al., 2012). 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 $l > 2$. For the complete proof see Section 8 in (Mohri et al., 2012). □

E. Proofs of Results in Section 4.3

Proof of Theorem 6. First observe that it holds $\alpha_i^* \max(0, f(x_i)) = 0$ for all $i = 1, \ldots, n$ in the optimal point of the Lagrangian saddle point problem.[5] This implies that we have $f(x_i) \leq 0$ if x_i is a *support vector* (that is, $\alpha_i^* > 0$) (Müller et al., 2001; Schölkopf & Smola, 2002). Since $\sum_{i=1}^{n} \alpha_i^* = 1$ and $\alpha_i^* \leq \frac{1}{\nu n}$ there must at least $\lceil \nu n \rceil$ many such points (the function $\lceil \cdot \rceil$ rounds a real number up to the next large integer). Hence there can be no more than $n - \lfloor \nu n \rfloor$ many points with $f(x_i) > 0$, which corresponds to a fraction of $\frac{n - \lfloor \nu n \rfloor}{n} \leq 1 - \nu$, and thus shows the assertion (b). Next observe that if we have $f(x_i) < 0$ then $\alpha_i^* = \frac{1}{\nu n}$ (to see this, note that if $\alpha_i^* < \frac{1}{\nu n}$ we could increase the objective of the Lagrangian by increasing α_i^*, which would contradict the optimality of α_i^*). Since $\sum_{i=1}^{n} \alpha_i^* = 1$ there can be no more than $\lfloor \nu n \rfloor$ many such points, which corresponds to a fraction of $\frac{\lfloor \nu n \rfloor}{n} \leq \nu$, thus showing the assertion (a). □

[4] Again this quickly follows from Jensen's inequality because $E\sigma_i\sigma_j = 0$ when $i \neq j$.

[5] For convex problems, this statement is known as the KKT condition *complementary slackness*. The argument holds, however, for the solution of the Lagrangian saddle point problem, regardless of whether or not the problem is convex, and for arbitrary objective and constraint functions.