Supplementary Material

A. Comparison to Structured SVMs (SSVM)

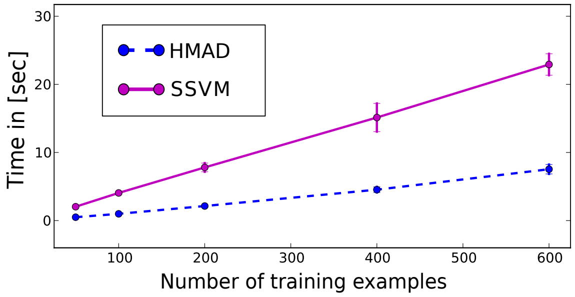

Figure A.1. Without the need for constraint generation, our hidden Markov anomaly detection easily outperforms the structured SVM.

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.

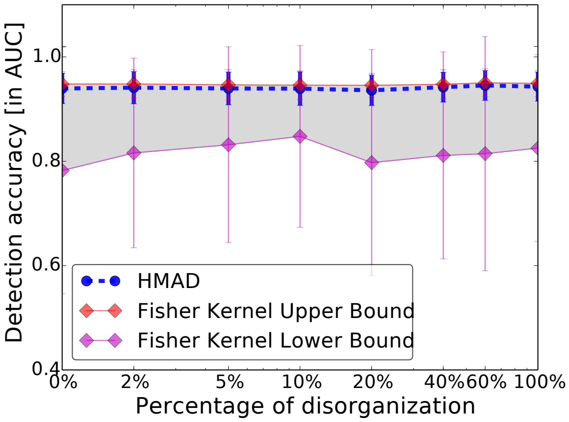

Figure B.1. Comparison for an increasing amount of disorganization of our method HMAD (blue) against a variety of Fisher kernels (gray area), including a lower bound (magenta) based on random model parameters and an upper bound (red) that was trained on *ground truth* data.

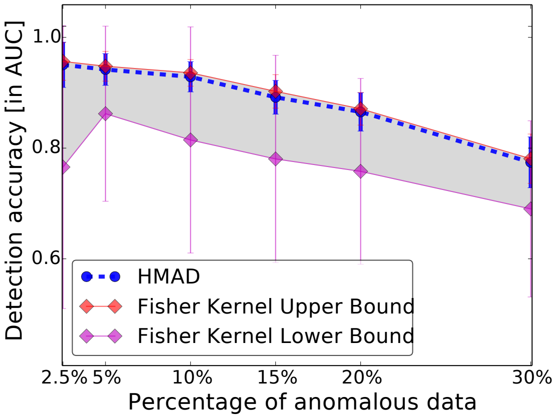

Figure B.2. Comparison for an increasing amount of anomalies of our method HMAD (blue) against a variety of Fisher kernels (gray area), including a lower bound (magenta) based on random model parameters and an upper bound (red) that was trained on *ground truth* data.

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.

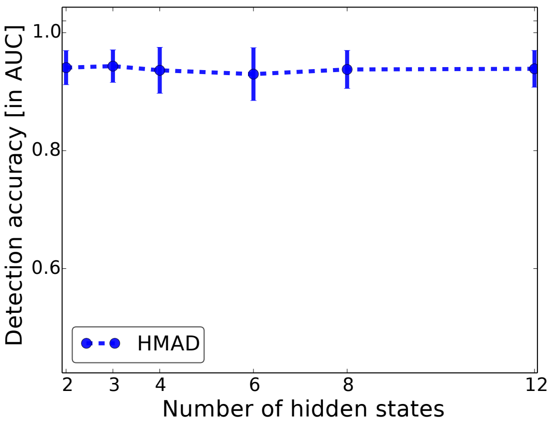

Figure C.1. Performance evaluation for an increasing number of hidden states of our method HMAD (blue).

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

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

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*

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

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

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 $\tilde{\boldsymbol{x}} \in D : g(\tilde{\boldsymbol{x}}) \le \tau$ 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,

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

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 $\mathcal{F}$,

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

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