Maximum Margin Multi-Label Structured Prediction Supplemental Material

Christoph H. Lampert
IST Austria (Institute of Science and Technology Austria)
Am Campus 1, 3400 Klosterneuburg, Austria
`http://www.ist.ac.at/∼chl` `chl@ist.ac.at`

1 Generalization Properties of MLSP

We provide the proof of Theorem 1 (Section 3.2) of the original manuscript.

Let $G_w(x) := \{y \in \mathcal{Y} : f_w(x,y) > 0\}$ for $f_w(x,y) = \langle w, \psi(x,y)\rangle$. We assume $|\mathcal{Y}| < r$ and $\|\psi(x,y)\| < s$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, and $\lambda(Y,y) \leq \Lambda$ for all $(Y,y) \in \mathbb{P}(\mathcal{Y}) \times \mathcal{Y}$. For any distribution, Q_w, over weight vectors, that can depend on w, we denote by $L(Q_w, P)$ the expected Δ_{max}-risk for P-distributed data,

$$L(Q_w, P) = \mathbb{E}_{\bar{w}\sim Q_w}\big\{\mathcal{R}_{P,\Delta_{max}}(G_{\bar{w}})\big\} = \mathbb{E}_{\bar{w}\sim Q_w,(x,Y)\sim P}\big\{\Delta_{max}(Y, G_{\bar{w}}(x))\big\}. \tag{1}$$

Theorem 1. *With probability at least $1-\sigma$ over the sample S of size n, the following inequality holds simultaneously for all weight vectors w.*

$$L(Q_w,D) \leq \frac{1}{n}\sum_{i=1}^{n}\ell(x^i,Y^i,f) + \frac{||w||^2}{n} + \Big(\frac{s^2||w||^2\ln(rn/||w||^2) + \ln\frac{n}{\sigma}}{2(n-1)}\Big)^{1/2} \tag{2}$$

for $\ell(x^i,Y^i,f) := \max_{y\in\mathcal{Y}}\big\{\lambda(Y^i,y)[\![v^i_y f(x^i,y) < 1]\!]\big\}$, where v^i is the binary indicator vector of Y^i.

Proof. The argument follows [1, Section 11.6], using the PAC-Bayesian bound

$$L(Q_w,D) \leq L(Q_w,S) + \sqrt{\frac{KL(Q_w,\pi) + \ln\frac{n}{\sigma}}{2(n-1)}}, \tag{3}$$

where π denotes a prior distribution on w, which we set as zero-mean Gaussian, $\pi(w) \propto \exp(-\frac{1}{2}||w||^2)$. We choose Q_w as a Gaussian centered at αw, $Q_w(\bar{w}) \propto \exp(-\frac{1}{2}||\bar{w}-\alpha w||^2)$. Then, the KL divergence between Q_w and π is just $\alpha^2||w||^2/2$. Analyzing the sample risk $L(Q_w,S)$ can be done for each training instance due to i.i.d. sampling. We denote by $\bar{Y}^i$ the predicted output for x^i with respect to $\bar{w}$. The proof is complete if we show

$$\mathbb{E}_{\bar{w}\sim Q_w}\Delta_{max}(Y^i,\bar{Y}^i) \leq \ell(x^i,Y^i,f_w) + \frac{\|w\|^2}{n}. \tag{4}$$

The claim follows then by inserting (4) and the expression for $KL(Q,P)$ into the PAC bound (3). □

To show inequality (4), we use

$$\mathbb{E}_{\bar{w}}\Delta_{max}(Y^i,\bar{Y}^i) = \mathbb{E}_{\bar{w}}\max_y\big\{\lambda(Y^i,y)[\![v^i_y f(x^i,y;\bar{w}) \leq 0]\!]\big\} \tag{5}$$

$$\leq \max_y\big\{\lambda(Y^i,y)[\![v^i_y f(x^i,y,w) < 1]\!]\big\}$$

$$+ P_{\bar{w}}\big\{v^i_y f(x^i,y,w) \geq 1 \wedge v^i_y f(x^i,y,\bar{w}) \leq 0\big\}, \tag{6}$$

where for the second term in (6) we used $\lambda(Y^i, y) \leq 1$.The first term is equivalent to the first term in Equation (4). The following lemma shows that the second term is bounded by $\frac{||w||^2}{n}$.

Lemma 1. *For $\bar{w} \sim Q_w$ with parameter $\alpha = (2s^2 \ln(rn/||w||^2))^{1/2}$, the implication*

$$v_y^i f(x^i, y; w) \geq 1 \quad \Rightarrow \quad v_y^i f(x^i, y; \bar{w}) > 0 \tag{7}$$

holds simultaneously for all $y \in \mathcal{Y}$ with probability at least $1 - ||w||^2/n$.

Proof. For any $\epsilon \geq 0$ and any ϕ with $||\phi|| = 1$,

$$P_{\bar{w} \sim Q_w}\big\{\langle(\alpha w - \bar{w}), \phi\rangle \geq \epsilon\big\} \leq \exp(-\epsilon^2/2). \tag{8}$$

Consider a fixed y with $v_y^i f(x^i, y, w) \geq 1$. Because $f(x^i, y, w) = \langle w, \psi(x^i, y)\rangle$, we can use inequality (8) to bound the probability that $\bar{w}$ causes y to be predicted with reverse sign. Writing $m^i(w) := v_y^i f(x^i, y, w)$ and $m^i(\bar{w}) := v_y^i f(x^i, y, \bar{w})$ we obtain

$$P_{\bar{w} \sim Q_w}\{m^i(\bar{w}) \leq \alpha m^i(w) - \epsilon||\psi(x^i, y)||\} \quad \leq \quad \exp(-\epsilon^2/2). \tag{9}$$

Because of $m^i(w) \geq 1$ this implies

$$P_{\bar{w} \sim Q_w}\big\{m^i(\bar{w}) \leq \alpha - \epsilon||\psi(x^i, y)||\big\} \quad \leq \quad \exp(-\epsilon^2/2). \tag{10}$$

Setting $\epsilon = \alpha/||\psi(x^i, y)||$ and using $\|\psi(x^i, y)\| < s$, we obtain

$$P_{\bar{w} \sim Q_w}\{m^i(\bar{w}) \leq 0\} \quad \leq \quad \exp(-\alpha^2/2s^2) \quad = \quad \frac{\|w\|^2}{rn}, \tag{11}$$

where the last equality is just the result of setting α to the stated value. The claim of Lemma 1 follows from this by taking a union bound over $\mathcal{Y}$, because $|\mathcal{Y}| < r$. □

References

[1] D. McAllester. Generalization bounds and consistency for structured labeling. In G. Bakır, T. Hofmann, B. Schölkopf, A.J. Smola, and B. Taskar, editors, *Predicting Structured Data*. MIT Press, 2007.

2 Detailed Toy Example

To give a more detailed impression of the difference between $\mathbb{P}$-SSVM, MLSP and SSVM training, we provide the complete log of training the difference techniques on an ad-hoc toy example consisting of only one training example with one output label.

Let $\mathcal{X} = (\mathbb{R}^2)^4$ be the set of 4-tuples in $\mathbb{R}^2$. For $x \in \mathcal{X}$ we write $x = (x_1, \dots, x_4)$ for the vortex coordinates, i.e. $x_i \in \mathbb{R}^2$ for $i = 1, \dots, 4$. Let $\mathcal{Y} = \{\texttt{0}, \texttt{1}\}^4$ be the set of binary labelings of such tuples, which we write as 4-letter strings. We form a training set with a single example $x^t = ((0,0),(0,1),(1,0),(1,1))$ with label set $Y^t = \{y^t\}$ for $y^t = \texttt{0001}$. The joint feature function we form as the center coordinates of the points labels by 0 and 1 respectively and an additional bias term.

$$\Psi(x,y) = \Big(\frac{1}{|\{i : y_i = \texttt{0}\}|}\sum\nolimits_{\{i:y_i=\texttt{0}\}} x_i,\ \ \frac{1}{|\{i : y_i = \texttt{1}\}|}\sum\nolimits_{\{i:y_i=\texttt{1}\}} x_i,\ \ 1\Big), \in \mathbb{R}^{2\times 2\times 1} \mathrel{\hat{=}} \mathbb{R}^5 \tag{12}$$

where entries corresponding to empty sums are defined as 0. The following table shows the resulting feature representations $\Psi(x^t, y)$ for $y \in \mathcal{Y}$ as column vectors.

y	`0000`	`0001`	`0010`	`0011`	`0100`	`0101`	`0110`	`0111`	`1000`	`1001`	`1010`	`1011`	`1100`	`1101`	`1110`	`1111`
$\Psi(x^t, y)$	$\begin{bmatrix}0.50\\0.50\\0.00\\0.00\\1.00\end{bmatrix}$	$\begin{bmatrix}0.33\\0.33\\1.00\\1.00\\1.00\end{bmatrix}$	$\begin{bmatrix}0.33\\0.67\\1.00\\0.00\\1.00\end{bmatrix}$	$\begin{bmatrix}0.00\\0.50\\1.00\\0.50\\1.00\end{bmatrix}$	$\begin{bmatrix}0.67\\0.33\\0.00\\1.00\\1.00\end{bmatrix}$	$\begin{bmatrix}0.50\\0.00\\0.50\\1.00\\1.00\end{bmatrix}$	$\begin{bmatrix}0.50\\0.50\\0.50\\0.50\\1.00\end{bmatrix}$	$\begin{bmatrix}0.00\\0.00\\0.67\\0.67\\1.00\end{bmatrix}$	$\begin{bmatrix}0.67\\0.67\\0.00\\0.00\\1.00\end{bmatrix}$	$\begin{bmatrix}0.50\\0.50\\0.50\\0.50\\1.00\end{bmatrix}$	$\begin{bmatrix}0.50\\1.00\\0.50\\0.00\\1.00\end{bmatrix}$	$\begin{bmatrix}0.00\\1.00\\0.67\\0.33\\1.00\end{bmatrix}$	$\begin{bmatrix}1.00\\0.50\\0.00\\0.50\\1.00\end{bmatrix}$	$\begin{bmatrix}1.00\\0.00\\0.33\\0.67\\1.00\end{bmatrix}$	$\begin{bmatrix}1.00\\1.00\\0.33\\0.33\\1.00\end{bmatrix}$	$\begin{bmatrix}0.00\\0.00\\0.50\\0.50\\1.00\end{bmatrix}$

As misclassification cost we use $\lambda(Y^t, y) := 1$ for $y = y^t$ and the normalized Hamming distance, $\lambda(Y^t, y) := \frac{1}{4}\sum_i y_i \neq y_i^t$, otherwise, see the following table.

y	`0000`	`0001`	`0010`	`0011`	`0100`	`0101`	`0110`	`0111`	`1000`	`1001`	`1010`	`1011`	`1100`	`1101`	`1110`	`1111`
$\lambda(Y^t, y)$	0.25	1.00	0.50	0.25	0.50	0.25	0.75	0.50	0.50	0.25	0.75	0.50	0.75	0.50	1.00	0.75

Tables 1–3 show in detail the process of training margin-rescaled versions of MLSP, SSVM and $\mathbb{P}$-SSVM with hard margin (no slack variables) in this situation. The tables are to be read as follows: Each row corresponds to one iteration of the working set training, starting with an empty working set and $w = 0$. In each iteration, we list the values of $f(x^t, y) = \langle w, \Psi(x^t, y)\rangle$ for all $y \in \mathcal{Y}$. From this and the values of $\lambda(Y^t, y)$, we identify the most violating contraint of the current optimization problem. For MLSP and SSVM these correspond to single outputs, which we state in the column y^t_{viol}. For $\mathbb{P}$-SSVM a constraint corresponds to a labels set. The entries of this Y^t_{viol} are marked by bold scores of the corresponding $f(x^t, y)$. From y^t_{viol} and Y^t_{viol}, respectively, we compute the new constraint to be added to the working set. Re-solving the optimization problem yields a new w, and a new iteration starts. The procedure is repeated until no violated constraint can be identified anymore.

In the captions, we report the output of multi-label prediction using the learned weight vector on the training example. For MLSP and $\mathbb{P}$-SSVM this is by construction the correct label set Y^t. For SSVM it is not, as the SSVM objective enforces only the correct ranking of labels, i.e. $f(x^t, y^t) > f(x^t, y)$ for $y \neq y^t$, but does not provide a value where to threshold. Comparing the training protocols in more details, one sees that MLSP in each iteration only considers the worst mistake, label, i.e. the label with a score farest from the desired one. For each example, it adds a constraint that enforces the score of this label to lead to a correct decision. In Table 1, the first pick is the positive label, y^t, which is subsequently guaranteed to lie in the predicted output. In the subsequent iterations, MLSP picks negative labels, $y \neq y^t$, that were incorrectly assigned high scores. In each case it adds a constraint that force the score of this label to lie below the margin determined by $\lambda(Y^i, y)$. The algorithm terminates before all labels have been considered in this way, because the sharing of information through the joint feature map Ψ. It causes the scores of some labels to be dragged down automatically when constraints for similar labels are added. For example, from iteration 3 to 4, the score of the label `0100` is reduced sufficiently after a constraint for `0111` was added.

SSVM training proceeds in similar way, but only enforces the correct ordering of label scores relatively to each other. This is an easier problem, which can be interpreted as the reason why SSVM training converges after fewer iterations. However, the scoring function $f(x, y)$ learned is not well suited to multi-label prediction. The table also shows that the bias entry of $\Psi(x, y)$ plays no role in SSVM training. All constraints added have a 0 as their last entry.

$\mathbb{P}$-SSVM forms new constraints for the working set as the sum of feature vectors of all currently mispredicted labels. The margin it enforces has the size of the sum of the per-label costs. The results are rather large entries in the constraint matrix (considering that only $|\mathcal{Y}| = 16$ in this example). In later iterations, Y^t_{viol} has fewer entries, which leads to the constraints more similar what MLSP and SSVM use. Note that the weight vector found by $\mathbb{P}$-SSVM in Table 3 is identical to the one by MLSP, but this is only a coincidence of the example chosen.

it.	w	$f(x^t, y)$																y^t_{viol}	new constraint
		0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111		
1	(0.00 , 0.00 ,0.00,0.00, 0.00)	0.00	**0.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0001	$\langle w, (0.33 , 0.33 , 1.00 , 1.00 , 1.0)\rangle \geq 1.0$
2	(0.10 , 0.10 ,0.31,0.31, 0.31)	0.41	1.00	0.72	0.83	0.72	0.83	0.72	0.72	0.45	0.72	0.62	0.72	0.62	0.72	**0.72**	0.62	1110	$\langle w, (-1.00,-1.00,-0.33,-0.33,-1.0)\rangle \geq 1.0$
3	(-0.75,-0.75,0.75,0.75,-0.00)	-0.75	1.00	0.00	0.75	0.00	0.75	0.00	**1.00**	-1.00	0.00	-0.75	0.00	-0.75	-0.00	-1.00	0.75	0111	$\langle w, (0.00 ,-0.00,-0.67,-0.67,-1.0)\rangle \geq 0.5$
4	(0.38 , 0.38 ,1.88,1.88,-3.00)	-2.62	1.00	-0.75	**0.00**	-0.75	0.00	-0.75	-0.50	-2.50	-0.75	-1.50	-0.75	-1.50	-0.75	-1.00	-1.12	0011	$\langle w, (0.00 ,-0.50,-1.00,-0.50,-1.0)\rangle \geq 0.25$
5	(0.63 , 0.13 ,1.63,2.13,-3.00)	-2.63	1.00	-1.08	-0.25	-0.42	**0.25**	-0.75	-0.50	-2.50	-0.75	-1.75	-1.08	-1.25	-0.42	-1.00	-1.13	0101	$\langle w, (-0.50,-0.00,-0.50,-1.00,-1.0)\rangle \geq 0.25$
6	(0.75 , 0.75 ,2.25,2.25,-4.00)	-3.25	1.00	-1.00	-0.25	-1.00	-0.25	-1.00	-1.00	-3.00	-1.00	-1.75	-1.00	-1.75	-1.00	-1.00	-1.75		

Table 1: MLSP training. Output on training set: $G(x^t) = \{\texttt{0001}\}$.

it.	w	$f(x^t, y)$																y^t_{viol}	new constraint
		0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111		
1	(0.00 , 0.00 ,0.00,0.00,0.00)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**	0.00	1110	$\langle w, (-0.67,-0.67,0.67,0.67,0.0)\rangle \geq 1.0$
2	(-0.38,-0.38,0.38,0.38,0.00)	-0.38	0.50	0.00	0.38	0.00	0.38	0.00	0.50	-0.50	0.00	-0.38	0.00	-0.38	-0.00	-0.50	**0.38**	1111	$\langle w, (0.33 , 0.33 ,0.50,0.50,0.0)\rangle \geq 0.75$
3	(0.00 , 0.00 ,0.75,0.75,0.00)	0.00	1.50	0.75	1.13	0.75	1.13	0.75	1.00	0.00	0.75	0.38	0.75	0.38	0.75	0.50	0.75		

Table 2: SSVM training. Output on training set: $G(x^t) = \{\texttt{0001}, \texttt{0010}, \texttt{0011}, \texttt{0100}, \texttt{0101}, \texttt{0110}, \texttt{0111}, \texttt{1001}, \texttt{1010}, \texttt{1011}, \texttt{1100}, \texttt{1101}, \texttt{1110}, \texttt{1111}\}$.

it.	w	$f(x^t, y)$																Y^t_{viol}	new constraint
		0000	0001	0010	0011	0100	0101	0110	0111	1000	1001	1010	1011	1100	1101	1110	1111		
1	(0.00 , 0.00 , 0.00 , 0.00 , 0.00)	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	*(bold)*	$\langle w, (-6.83,-6.83,-5.50,-5.50,-14.0)\rangle \geq 9.0$
2	(-0.18,-0.18,-0.14,-0.14,-0.36)	-0.54	**-0.76**	-0.68	-0.66	-0.68	-0.66	**-0.68**	-0.55	-0.59	-0.68	**-0.69**	-0.68	**-0.69**	-0.68	**-0.81**	**-0.50**		$\langle w, (-2.67,-2.67,-0.83,-0.83,-4.00)\rangle \geq 5.0$
3	(-0.42,-0.42,-0.13,-0.13,-0.63)	-1.05	**-1.18**	-1.19	-1.04	-1.19	-1.04	-1.19	-0.81	-1.20	-1.19	-1.33	-1.19	-1.33	-1.19	-1.56	-0.76		$\langle w, (0.33 , 0.33 , 1.00 , 1.00 , 1.00)\rangle \geq 1.0$
4	(-0.77,-0.77, 1.16 , 1.16 ,-0.80)	-1.57	1.00	**-0.41**	**0.55**	**-0.41**	**0.55**	**-0.41**	**0.74**	-1.83	-0.41	-1.38	**-0.41**	-1.38	**-0.41**	-1.57	**0.35**		$\langle w, (-3.00,-3.00,-5.17,-5.17,-9.00)\rangle \geq 4.5$
5	(-0.69,-0.69, 1.76 , 1.76 ,-2.06)	-2.75	1.00	-0.99	**0.23**	-0.99	**0.23**	-0.99	**0.28**	-2.98	-0.99	-2.21	-0.99	-2.21	-0.99	-2.26	**-0.30**		$\langle w, (-0.50,-0.50,-2.67,-2.67,-4.00)\rangle \geq 1.75$
6	(0.63 , 0.63 , 1.76 , 1.76 ,-2.94)	-2.31	1.00	-0.55	**0.01**	-0.55	**0.01**	**-0.55**	-0.60	-2.10	-0.55	-1.11	-0.55	-1.11	-0.55	**-0.50**	-1.18		$\langle w, (-2.00,-2.00,-2.33,-2.33,-4.00)\rangle \geq 2.25$
7	(0.27 , 0.27 , 1.98 , 1.98 ,-3.15)	-2.88	1.00	-0.89	**-0.04**	-0.89	**-0.04**	-0.89	-0.51	-2.79	-0.89	-1.75	-0.89	-1.75	-0.89	-1.28	-1.17		$\langle w, (-0.50,-0.50,-1.50,-1.50,-2.00)\rangle \geq 0.5$
8	(0.75 , 0.75 , 2.25 , 2.25 ,-4.00)	-3.25	1.00	-1.00	-0.25	-1.00	-0.25	-1.00	-1.00	-3.00	-1.00	-1.75	-1.00	-1.75	-1.00	-1.00	-1.75		

Table 3: $\mathbb{P}$-SSVM training. Output on training set: $G(x^t) = \{\texttt{0001}\}$.