On Robustness and Regularization of Structural Support Vector Machines: Supplementary Material

MohamadAli Torkamani ALI@CS.UOREGON.EDU
Daniel Lowd LOWD@CS.UOREGON.EDU
Computer and Information Science Department, University of Oregon

Proof of Lemma 4.1:

Proof. We form $\boldsymbol{\delta}_{\tilde{\boldsymbol{y}}}^{\mathcal{C}}(\boldsymbol{x},\boldsymbol{y},\tilde{\boldsymbol{x}})$ (from Eq. (7) in the paper):

$$\begin{aligned}&\boldsymbol{\delta}^{\mathcal{C}} = \boldsymbol{\delta}_{\tilde{\boldsymbol{y}}}^{\mathcal{C}}(\boldsymbol{x},\boldsymbol{y},\tilde{\boldsymbol{x}}) = \\ &\boldsymbol{\phi}_{\mathcal{C}}(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}) - \boldsymbol{\phi}_{\mathcal{C}}(\tilde{\boldsymbol{x}},\boldsymbol{y}) - \boldsymbol{\phi}_{\mathcal{C}}(\boldsymbol{x},\tilde{\boldsymbol{y}}) - \boldsymbol{\phi}_{\mathcal{C}}(\boldsymbol{x},\boldsymbol{y}) = \\ &\sum_{(c_x,c_y)\in\mathcal{C}} (\prod_{i\in c_x}\tilde{\boldsymbol{x}}_i - \prod_{i\in c_x}\boldsymbol{x}_i)(\prod_{i\in c_y}\tilde{\boldsymbol{y}}_i - \prod_{i\in c_y}\boldsymbol{y}_i) \qquad (1)\end{aligned}$$

For an individual elements of the vector $\boldsymbol{\delta}$ as expanded in (1), we can apply Hölder's inequality to the right-hand side:

$$\begin{aligned}&|\boldsymbol{\delta}^{\mathcal{C}}| \le \\ &(\sum_{c_x\in\mathcal{C}} |\prod_{i\in c_x}\tilde{\boldsymbol{x}}_i - \prod_{i\in c_x}\boldsymbol{x}_i|^p)^{\frac{1}{p}}(\sum_{c_y\in\mathcal{C}} |\prod_{i\in c_y}\tilde{\boldsymbol{y}}_i - \prod_{i\in c_y}\boldsymbol{y}_i|^q)^{\frac{1}{q}}\end{aligned}$$

where $\frac{1}{p}+\frac{1}{q}=1$. Since $|\prod_{i\in c_y}\tilde{\boldsymbol{y}}_i - \prod_{i\in c_y}\boldsymbol{y}_i|^q \le 1$, we will have: $\sum_{c_y\in\mathcal{C}}|\prod_{i\in c_y}\tilde{\boldsymbol{y}}_i - \prod_{i\in c_y}\boldsymbol{y}_i|^q \le |\mathcal{C}|$, therefore:

$$|\boldsymbol{\delta}^{\mathcal{C}}| \le |\mathcal{C}|^{\frac{1}{q}}(\sum_{c_x\in\mathcal{C}} |\prod_{i\in c_x}\tilde{\boldsymbol{x}}_i - \prod_{i\in c_x}\boldsymbol{x}_i|^p)^{\frac{1}{p}}$$

After applying Lemma A and raising both sides of the inequality to the power of p, we will have:

$$\begin{aligned}|\boldsymbol{\delta}^{\mathcal{C}}|^p &\le |\mathcal{C}|^{\frac{p}{q}}(\alpha\sum_{c_x\in\mathcal{C}}\sum_{i\in c_x}|\tilde{\boldsymbol{x}}_i - \boldsymbol{x}_i|^p) \\ \Rightarrow \quad \frac{|\boldsymbol{\delta}^{\mathcal{C}}|^p}{\alpha|\mathcal{C}|^{\frac{p}{q}}} &\le \sum_{c_x\in\mathcal{C}}\sum_{i\in c_x}|\tilde{\boldsymbol{x}}_i - \boldsymbol{x}_i|^p \qquad (2)\end{aligned}$$

where $\alpha = \max_{c_x\in\mathcal{C}}|c_x|^{(p-1)}$, and $|c_x|$ is the number of variables in c_x. □

The proof of Lemma 4.1 depends on the following lemma:
Lemma A. For any sequence $a_1,\dots,a_n,b_1,\dots,b_n$, such that $0 \le a_i, b_j \le 1$, we have $|\prod_{i=1}^n a_i - \prod_{i=1}^n b_i|^p \le n^{(p-1)}\sum_{i=1}^n |a_i-b_i|^p$.

Proof. For $n = 1$, the inequality is trivial. Let $u_1 = \prod_{i=1}^{\lfloor n/2\rfloor} a_i$, $u_2 = \prod_{i=1}^{\lfloor n/2\rfloor} b_i$, $v_1 = \prod_{i=\lfloor n/2\rfloor+1}^{n} a_i$, and $v_2 = \prod_{i=\lfloor n/2\rfloor+1}^{n} a_i$. Also it is a known fact that $|f+g|^p \le 2^{p-1}(|f|^p+|g|^p)\ \ g,f\in\mathbf{R}$. We have:

$$\begin{aligned}|\prod_{i=1}^n a_i - \prod_{i=1}^n b_i|^p &= |u_1v_1 - u_2v_2|^p \\ &= |u_1v_1 - u_1v_2 + u_1v_2 - u_2v_2|^p \\ &\le 2^{p-1}(|u_1v_1-u_1v_2|^p + |u_1v_2-u_2v_2|^p) \\ &= 2^{p-1}(u_1^p|v_1-v_2|^p + v_2^p|u_1-u_2|^p) \\ &\le 2^{p-1}(|v_1-v_2|^p + |u_1-u_2|^p)\end{aligned}$$

by recursive application of the above procedure, the products can be decomposed at most $\log_2 n$ times. Therefore,

$$\begin{aligned}|\prod_{i=1}^n a_i - \prod_{i=1}^n b_i|^p &\le 2^{(p-1)\log_2 n}\sum_{i=1}^n |a_i-b_i|^p \\ &= n^{p-1}\sum_{i=1}^n|a_i-b_i|^p\end{aligned}$$

□

Proof of Corollary 4.3:

Proof. We begin with the result of Theorem 4.3, where $\frac{1}{B(d\alpha_i)^{\frac{1}{p}}|\mathcal{C}_i|^{\frac{1}{q}}}$ is the coeficient of variations in the feature corresponding to clique $\mathcal{C}_i$. Since $p=1$ then $q=\infty$, and $\alpha_i = \max_{c_x\in\mathcal{C}_i}|c_x|^{(p-1)} = 1$:

$$\begin{aligned}\frac{1}{B(d\alpha_i)^{\frac{1}{p}}|\mathcal{C}_i|^{\frac{1}{q}}} &= \frac{1}{Bd|\mathcal{C}_i|^{\frac{1}{\infty}}} \\ &= \frac{1}{Bd}\end{aligned}$$

Also in (2), set $p = 1$ and $q = \infty$.

$$|\boldsymbol{\delta}^{\mathcal{C}}| \leq (\sum_{c_x \in \mathcal{C}} |\prod_{i \in c_x} \tilde{\boldsymbol{x}}_i - \prod_{i \in c_x} \boldsymbol{x}_i|) \max_{c_y \in \mathcal{C}} |\prod_{i \in c_y} \tilde{\boldsymbol{y}}_i - \prod_{i \in c_y} \boldsymbol{y}_i|$$

Since, $\max_{c_y \in \mathcal{C}} |\prod_{i \in c_y} \tilde{\boldsymbol{y}}_i - \prod_{i \in c_y} \boldsymbol{y}_i| = 1$, we will be using a tighter upper-bound. $\square$

Proof of Proposition 5.7:

Proof. We prove the case when regularization function is $\|\boldsymbol{w}\| = \|\boldsymbol{w}\|_\infty$ (the proofs for $\|\boldsymbol{M}^{-1}\boldsymbol{w}\|_\infty$ and $\|\boldsymbol{M}^{-1}\boldsymbol{w}\|_1$ are very similar, but for simplicity we chose this case). Recall that the optimization program of the robust structural SVM is:

$$\begin{aligned} &\underset{\boldsymbol{w},\xi}{\text{minimize}}\; c_1 f(\boldsymbol{w}) + c_2\|\boldsymbol{w}\|_\infty + \xi \quad \text{subject to} \\ &\xi \geq \max_{\tilde{\boldsymbol{y}}}\; \boldsymbol{w}^T(\boldsymbol{\phi}(\boldsymbol{x},\tilde{\boldsymbol{y}}) - \boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y})) + \Delta(\boldsymbol{y},\tilde{\boldsymbol{y}}) \end{aligned} \tag{3}$$

It can be re-written as:

$$\begin{aligned} &\underset{\boldsymbol{w},\xi,t}{\text{minimize}}\; c_1 f(\boldsymbol{w}) + c_2 t + \xi \quad \text{subject to} \\ &\xi \geq \max_{\tilde{\boldsymbol{y}}}\; \boldsymbol{w}^T(\boldsymbol{\phi}(\boldsymbol{x},\tilde{\boldsymbol{y}}) - \boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y})) + \Delta(\boldsymbol{y},\tilde{\boldsymbol{y}}) \\ &w_i \leq t,\; -w_i \leq t \;\; \forall w_i \end{aligned}$$

In vector form we can write these constraints as: $\boldsymbol{w} \leq \boldsymbol{1}t$ and $-\boldsymbol{w} \leq \boldsymbol{1}t$. Clearly, there are two vectors $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$ for which:

$$\begin{aligned} \boldsymbol{w} + \boldsymbol{s_1} = \boldsymbol{1}t \quad &\Rightarrow \quad \boldsymbol{w} = \boldsymbol{1}t - \boldsymbol{s_1} \\ -\boldsymbol{w} + \boldsymbol{s_2} = \boldsymbol{1}t \quad &\Rightarrow \quad \boldsymbol{w} = \boldsymbol{s_2} - \boldsymbol{1}t \end{aligned}$$

Let $\gamma = [\boldsymbol{s_1}^T \;\; \boldsymbol{s_2}^T \;\; t]^T$, $m = \dim \boldsymbol{w}$, $\boldsymbol{I_{s_1}} = [\boldsymbol{I}_{m\times m} \; \underline{\boldsymbol{0}}_{m\times m} \; \boldsymbol{0}_{m\times 1}]$, $\boldsymbol{I_{s_2}} = [\underline{\boldsymbol{0}}_{m\times m} \; \boldsymbol{I}_{m\times m} \; \boldsymbol{0}_{m\times 1}]$, and $\boldsymbol{I_t} = [\underline{\boldsymbol{0}}_{1\times m} \; \underline{\boldsymbol{0}}_{1\times m} \; 1]$. (i.e. $\boldsymbol{s_1} = \boldsymbol{I_{s_1}}\gamma$, $\boldsymbol{s_2} = \boldsymbol{I_{s_2}}\gamma$, $t = \boldsymbol{I_t}\gamma$). By substitution:

$$\begin{aligned} \boldsymbol{w} &= \boldsymbol{1}\boldsymbol{I_t}\gamma - \boldsymbol{I_{s_1}}\gamma = (\boldsymbol{1}\boldsymbol{I_t} - \boldsymbol{I_{s_1}})\gamma \\ \boldsymbol{w} &= \boldsymbol{I_{s_2}}\gamma - \boldsymbol{1}\boldsymbol{I_t}\gamma = (\boldsymbol{I_{s_2}} - \boldsymbol{1}\boldsymbol{I_t})\gamma \end{aligned}$$

which implies $(\boldsymbol{1}\boldsymbol{I_t} - \boldsymbol{I_{s_1}})\gamma = (\boldsymbol{I_{s_2}} - \boldsymbol{1}\boldsymbol{I_t})\gamma$, therefore: $(2*\boldsymbol{1}\boldsymbol{I_t} - \boldsymbol{I_{s_1}} - \boldsymbol{I_{s_2}})\gamma = \boldsymbol{0}$, or equivalently $\gamma \in \mathcal{N}(2*\boldsymbol{1}\boldsymbol{I_t} - \boldsymbol{I_{s_1}} - \boldsymbol{I_{s_2}})$, where $\mathcal{N}(.)$ returns the null-space of the input matrix. Let columns of matrix $\boldsymbol{B}$ span $\mathcal{N}(2*\boldsymbol{1}\boldsymbol{I_t} - \boldsymbol{I_{s_1}} - \boldsymbol{I_{s_2}})$, also let $\gamma = \boldsymbol{B}\boldsymbol{\lambda}$, we will have $\boldsymbol{w} = (\boldsymbol{1}\boldsymbol{I_t} - \boldsymbol{I_{s_1}})\boldsymbol{B}\boldsymbol{\lambda}$. Let $\boldsymbol{A} = \boldsymbol{B}^T(\boldsymbol{1}\boldsymbol{I_t} - \boldsymbol{I_{s_1}})^T$ and $\boldsymbol{b} = \boldsymbol{I_t}^T$, then we can rewrite Problem (3) as:

$$\begin{aligned} &\underset{\boldsymbol{\lambda}\geq 0,\xi}{\text{minimize}}\; c_1 f(\boldsymbol{A}^T\boldsymbol{\lambda}) + c_2\boldsymbol{b}^T\boldsymbol{\lambda} + \xi \quad \text{subject to} \\ &\xi \geq \max_{\tilde{\boldsymbol{y}}}\; \boldsymbol{\lambda}^T\boldsymbol{A}(\boldsymbol{\phi}(\boldsymbol{x},\tilde{\boldsymbol{y}}) - \boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y})) + \Delta(\boldsymbol{y},\tilde{\boldsymbol{y}}) \end{aligned}$$

Note that since $(2*\boldsymbol{1}\boldsymbol{I_t} - \boldsymbol{I_{s_1}} - \boldsymbol{I_{s_2}})\boldsymbol{B} = \underline{\boldsymbol{0}}$, we will have $(\boldsymbol{1}\boldsymbol{I_t} - \boldsymbol{I_{s_1}})\boldsymbol{B} = (\boldsymbol{I_{s_2}} - \boldsymbol{1}\boldsymbol{I_t})\boldsymbol{B}$, and A can be transpose of any of them. $\square$