A Supplementary Material for "Robust Cost Sensitive Support Vector Machine"

1 Proposition 1

The supplementary material is devoted to the full version proof of Proposition 1. The proposition is given as follows.

Proposition 1 *Let* $\mathcal{U}_i = \{\boldsymbol{\delta}_i |\ \|\boldsymbol{\delta}_i\|_q \leq \gamma_i\}$ *for* $i = 1, \ldots, m$, *and suppose the regularizer* $r(\boldsymbol{w})$ *takes the form* $\sum_{k=1}^{l} \eta_k \|\boldsymbol{w}\|_{p_k}^{d_k}$ *where* $\eta_k \geq 0, d_k, p_k \in \mathbb{N}$. *Then the following robust classification problem*

$$\begin{array}{ll} \min\limits_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{\zeta}} & r(\boldsymbol{w}) + \sum_{i=1}^{m} C_i \zeta_i \\ \text{s.t.} & \min\limits_{\boldsymbol{\delta}_i \in \mathcal{U}_i} \ y_i(\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}_i + \boldsymbol{\delta}_i) + b) \geq 1 - \zeta_i, \\ & \zeta_i \geq 0 \quad i = 1, \cdots, m, \end{array} \tag{1}$$

is equivalent to the following regularized classification problem

$$\begin{array}{ll} \min\limits_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{\zeta'}} & r'(\boldsymbol{w}) + R\|\boldsymbol{w}\|_p + \sum_{i=1}^{m} C'_i \zeta'_i \\ \text{s.t.} & y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geq 1 - \zeta'_i, \\ & \zeta'_i \geq 0 \quad i = 1, \cdots, m, \end{array} \tag{2}$$

where p *denotes the dual norm of* q *and the regularizer* $r'(\boldsymbol{w})$ *takes the form* $\sum_{k=1}^{l} \eta'_k \|\boldsymbol{w}\|_{p_k}^{d_k}$. *Here, parameters* η'_k, R *and costs* C'_i *are assigned according to* (1).

When we say the two classification problem is "equivalent", we mean they produce the same optimal hyperplane $\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$. This result tells us that robust classification problems where each data has different uncertainty set sizes are equivalent to solving a regularized classification problem.

2 Proof of Proposition 1

We first observe that we can rewrite (1) into a standard (non-robust) optimization problem, since each perturbation in the data are uncorrelated. (1) is equivalent to

the following problem.

$$\begin{array}{ll} \min\limits_{\mathbf{w},\mathbf{b},\boldsymbol{\zeta}} & r(\mathbf{w}) + \sum_{i=1}^{m} C_i\zeta_i \\ \text{s.t.} & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - \gamma_i\|\mathbf{w}\|_p \geq 1 - \zeta_i, \\ & \zeta_i \geq 0 \quad i = 1, \cdots, m, \end{array} \tag{3}$$

where $\|\cdot\|_p$ is the dual norm of $\|\cdot\|_q$.

To show equivalence between (2) and (3), we create an identical optimal hyperplane $\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$ for (2) and (3) through setting the η'_k of the regularizer $r'(\mathbf{w})$, parameter R and costs C'_i appropriately.

In doing so, we take a look at the Lagrange function of the two problems and derive the KKT conditions. The Lagrange functions L_{rob} and L_{reg} for (3) and (2) are

$$L_{rob}(\mathbf{w}, b, \boldsymbol{\zeta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = r(\mathbf{w}) + \sum_{i=1}^{m} C_i\zeta_i - \sum_{i=1}^{m} \alpha_i\big(y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - \gamma_i\|\mathbf{w}\|_p - 1 + \zeta_i\big) - \sum_{i=1}^{m} \beta_i\zeta_i,$$

$$L_{reg}(\mathbf{w}, b, \boldsymbol{\zeta}', \boldsymbol{\alpha}', \boldsymbol{\beta}') = r'(\mathbf{w}) + \sum_{i=1}^{m} C'_i\zeta'_i + \sum_{i=1}^{m} \alpha'_i\big(y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1 + \zeta'_i\big) - \sum_{i=1}^{m} \beta'_i\zeta'_i,$$

where $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $(\boldsymbol{\alpha}', \boldsymbol{\beta}')$ are the dual variables associated with problem (3) and (2) respectively. Then, we obtain the KKT conditions as follows.

(I) KKT Conditions for the Robust Classifier (3)

- $\frac{\partial r(\mathbf{w})}{\partial w_j} + (\sum_{i=1}^{m} \alpha_i\gamma_i)\frac{|w_j|^{p-2}}{\|\mathbf{w}\|_p^{p-1}} w_j = \sum_{i=1}^{m} \alpha_i y_i x_{ij}$
- $\sum_{i=i}^{m} \alpha_i y_i = 0$
- $\boldsymbol{\alpha} + \boldsymbol{\beta} = \mathbf{C}$
- $\boldsymbol{\alpha} \geq 0, \boldsymbol{\beta} \geq 0$
- $\alpha_i(y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - \gamma_i\|\mathbf{w}\|_p - 1 + \zeta_i\big) = 0$
- $\beta_i\zeta_i = 0$
- $y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - \gamma_i\|\mathbf{w}\|_p \geq 1 - \zeta_i$
- $\zeta_i \geq 0$

(II) KKT Conditions for the Regularized Classifier (2)

- $\frac{\partial r'(\mathbf{w})}{\partial w_j} + R\frac{|w_j|^{p-2}}{\|\mathbf{w}\|_p^{p-1}} w_j = \sum_{i=1}^{m} \alpha'_i y_i x_{ij}$
- $\sum_{i=1}^{m} \alpha'_i y_i = 0$
- $\boldsymbol{\alpha}' + \boldsymbol{\beta}' = \mathbf{C}'$
- $\boldsymbol{\alpha}' \geq 0, \boldsymbol{\beta}' \geq 0$
- $\alpha'_i(y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1 + \zeta'_i) = 0$
- $\beta'_i\zeta'_i = 0$
- $y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \zeta'_i$
- $\zeta'_i \geq 0$

Where w_j, x_{ij} denotes the j-th element of $\mathbf{w}, \mathbf{x}_i$ and $\mathbf{C}, \mathbf{C}'$ denotes the vector with costs C_i, C'_i respectively. The first four items are the stationarity conditions and

dual feasibility. The following two items are the complementary slackness, and the last two items are the primal feasibility.

Let $(\mathbf{w}_{rob}, b_{rob}, \boldsymbol{\zeta}_{rob}, \boldsymbol{\alpha}_{rob}, \boldsymbol{\beta}_{rob})$ and $(\mathbf{w}_{reg}, b_{reg}, \boldsymbol{\zeta}'_{reg}, \boldsymbol{\alpha}'_{reg}, \boldsymbol{\beta}'_{reg})$ be points satisfying the above KKT conditions (I) and (II) respectively. Since both problems are convex, any solution satisfying the KKT conditions is the optimal solution. Therefore, to prove the proposition, we show that we can construct (2) to have an identical optimal hyperplane as $\mathbf{w}_{rob}^{\mathrm{T}}\mathbf{x} + b_{rob} = 0$ that satisfies the KKT conditions (II), by appropriately setting the η'_k of the regularizer $r'(\mathbf{w})$, parameter R and costs C'_i.

Let $\gamma = \max_i \gamma_i$. Then for an appropriate regularizer $r'(\mathbf{w})$, parameter R and costs C'_i, (2) will have an optimal solution satisfying $(\mathbf{w}_{reg}, b_{reg}) = (\tau\mathbf{w}_{rob}, \tau b_{rob})$, where τ is defined as $\tau = \frac{1}{1+\gamma\|\mathbf{w}_{rob}\|_p}$. Since hyperplanes are invariant under scaling of the parameters, this will be our desired solution. We explain the motivation of τ shortly after and give a detailed proof of this statement.

We consider two cases; for data with uncertainty set size γ and data with uncertainty set size smaller than γ. Take a look at the following figure.

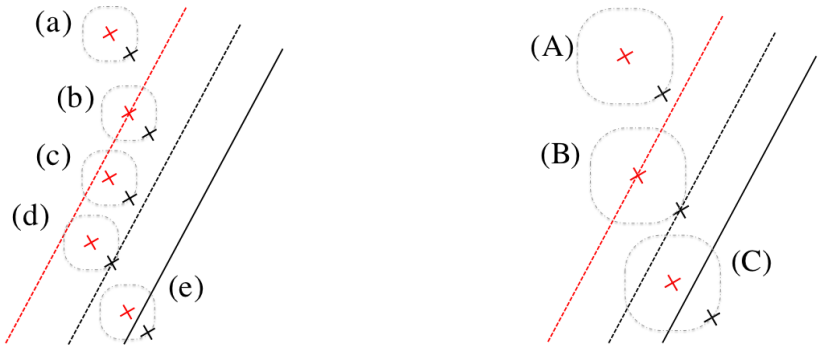

Figure 1: (Left) Data with uncertainty set size $\gamma_i < \gamma$. (Right) Data with uncertainty set size $\gamma_i = \gamma$. Black and red are used to illustrate the results of robust and regularized classification respectively. The bold lines represent the optimal hyperplanes $\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$ and the dashed lines represent the margin hyperplanes $|\mathbf{w}^{\mathrm{T}}\mathbf{x} + b| = 1$ for the robust and regularized problems. Finally, the dashed grey circles represents the uncertainty set of each data.

The reason for assigning τ in the above way was so the black dashed line gets translated exactly by γ to the red dashed line. The description of the letters in parenthesis is summarized in the following Table 1. Margin Errors (ME) are data with $\zeta > 0$, Margin Support Vectors (MSV) are data with $\zeta = 0$ and $\alpha > 0$, and Others denotes data other than ME and MSV i.e., those data that are correctly classified and not on the margin hyperplane.

By focusing on the different types of data depicted in Table 1, we derive a method of constructing a pair $(\mathbf{w}, b, \boldsymbol{\zeta}, \boldsymbol{\alpha}', \boldsymbol{\beta}')$ satisfying the KKT conditions (II) where $(\mathbf{w}, b) = (\tau\mathbf{w}_{rob}, \tau b_{rob})$ using $(\mathbf{w}_{rob}, b_{rob}, \boldsymbol{\zeta}_{rob}, \boldsymbol{\alpha}_{rob}, \boldsymbol{\beta}_{rob})$. The following Table

		Regularized Problem		
		Margin Errors	Margin Support Vectors	Others
Robust Problem	ME	(C), (e)	-	-
	MSV	(d)	(B)	-
	Others	(c)	(b)	(A), (a)

Table 1: Description of different types of data in Figure 1. "Others" stands for those data that are correctly classified and not on the margin hyperplane.

2 summarizes how we assign the costs C_i for each types of data and what the pairs $(\boldsymbol{\zeta}', \boldsymbol{\alpha}', \boldsymbol{\beta}')$ evaluate to.

Table 2: Relationship between the assigned costs and the optimal solutions.

Types	Costs	Regularized Problem			Robust Problem		
	C_i'	ζ_i'	α_i'	β_i'	ζ_i	α_i	β_i
(A)	C_i	0	0	C_i	0	0	C_i
(B)	C_i	0	α_i	β_i	0	α_i	β_i
(C)	C_i	$\tau\zeta_i$	C_i	0	ζ_i	C_i	0
(a)	C_i	0	0	C_i	0	0	C_i
(b)	0	0	0	0	0	0	C_i
(c)	0	0	0	0	0	0	C_i
(d)	α_i	$\tau(\gamma - \gamma_i)\|\mathbf{w}_{rob}\|_p$	α_i	0	0	α_i	β_i
(e)	C_i	$\tau\zeta_i + \tau(\gamma - \gamma_i)\|\mathbf{w}_{rob}\|_p$	C_i	0	ζ_i	C_i	0

We show the above $\{(\zeta_i', \alpha_i', \beta_i')\}_{i=1}^m$ satisfies the KKT conditions (II). Since $\{(\zeta_i, \alpha_i, \beta_i)\}_{i=1}^m$ satisfies the KKT conditions in (I), it is straightforward to see that $\boldsymbol{\alpha}' + \boldsymbol{\beta}' = \mathbf{C}'$, $\boldsymbol{\alpha}' \geq 0$, $\boldsymbol{\beta}' \geq 0$, $\beta_i'\zeta_i' = 0$ and $\zeta_i' \geq 0$ holds. Furthermore, since $\boldsymbol{\alpha}' = \boldsymbol{\alpha}$, we obtain $\sum_{i=1}^m y_i\alpha_i' = 0$.

Next we consider the remaining complementary condition $\alpha_i'(y_i(\mathbf{w}_{reg}^{\mathrm{T}}\mathbf{x}_i + b_{reg}) - 1 + \zeta_i') = 0$. Since this holds when $\alpha_i = 0$, we only consider the case of $\alpha_i \neq 0$. Then we can write ζ_i' as $\tau\zeta_i + \tau(\gamma - \gamma_i)\|\mathbf{w}_{rob}\|_p$. Recalling that $(\mathbf{w}_{reg}, \mathbf{b}_{reg}) = (\tau\mathbf{w}_{rob}, \tau\mathbf{b}_{rob})$ and $\tau = \frac{1}{1+\gamma\|\mathbf{w}_{rob}\|_p}$ we obtain the following equation.

$$
\begin{aligned}
& y_i(\mathbf{w}_{rob}^{\mathrm{T}}\mathbf{x}_i + b_{rob}) - \gamma_i\|\mathbf{w}_{rob}\|_p - 1 + \zeta_i = 0 \\
\Leftrightarrow \quad & y_i(\mathbf{w}_{reg}^{\mathrm{T}}\mathbf{x}_i + b_{reg}) - \tau\gamma_i\|\mathbf{w}_{rob}\|_p - \tau + \tau\zeta_i = 0 \\
\Leftrightarrow \quad & y_i(\mathbf{w}_{reg}^{\mathrm{T}}\mathbf{x}_i + b_{reg}) - \tau(\gamma_i\|\mathbf{w}_{rob}\|_p + 1) - \tau(\gamma - \gamma_i)\|\mathbf{w}_{rob}\|_p + \zeta_i' = 0 \\
\Leftrightarrow \quad & y_i(\mathbf{w}_{reg}^{\mathrm{T}}\mathbf{x}_i + b_{reg}) - 1 + \zeta_i' = 0.
\end{aligned}
$$

The remaining primal feasibility condition $y_i(\mathbf{w}_{reg}^{\mathrm{T}}\mathbf{x}_i + b_{reg}) \geq 1 - \zeta_i'$ is obtained in the same manner.

Finally, we assign a suitable regularizer $r'(\mathbf{w})$ and parameter R, so the first item in the KKT conditions (Ⅱ) holds. Since $r(\mathbf{w})$ takes the form $\sum_{k=1}^{l} \eta_k \|\mathbf{w}\|_{p_k}^{d_k}$, by substituting $\mathbf{w}_{rob} = \mathbf{w}_{reg}/\tau$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}'$ into the first item in the KKT conditions (I), we obtain

$$\sum_{k=1}^{l} \eta_k \tau^{2-d_k} w_{reg,j} \frac{|w_{reg,j}|^{p_i-2} \|\mathbf{w}_{reg}\|_{p_i}^{a_i}}{\|\mathbf{w}_{reg}\|_{p_i}^{p_i}} + (\sum_{i=1}^{m} \alpha_i' \gamma_i) \frac{|w_{reg,j}|^{p-2}}{\|\mathbf{w}_{reg}\|_p^{p-1}} w_{reg,j} = \sum_{i=1}^{m} \alpha_i' y_i \mathbf{x}_{ij},$$

where $w_{reg,i}$ denotes the i-th element in $\mathbf{w}_{reg}$. Observe that by assigning $r'(\mathbf{w}) = \sum_{k=1}^{l} \eta_k \tau^{2-d_k} \|\mathbf{w}\|_{p_k}^{d_k}$ and $R = \sum_{i=1}^{m} \alpha_i \gamma_i$, we obtain the first item in the KKT conditions (Ⅱ). Thus, every KKT conditions (Ⅱ) has been derived from the KKT conditions (I).

We showed that the values assigned according to Table 2 satisfy the KKT conditions (Ⅱ). Therefore, a robust classification problem (1) is equivalent to a regularized classification problem (2) if the regularizer $r'(\mathbf{w})$, parameter R and costs C_i' are assigned in the above manner. Hence proving the proposition. ■

2.1 Comments on Proposition 1

Proposition 1 actually holds for uncertainty sets defined with a quadratic norm $\|\cdot\|_A$ as well, where A is a positive semidefinite matrix. In this case, the dual norm p is given as $\|\cdot\|_{A^{-1}}$. The proof of this follows in the same manner. Quadratic norms induces, ellipsoid shaped uncertainty sets.

We point out that (3) was generically constructed so that its KKT condition would equal to that of (2).