Supplementary Information - Proofs of Theorems

Appendix A The adversarial distribution in $\mathbf{RSVM}_2(\sigma)$

In general there is no probability distribution $p(\bar{\boldsymbol{x}}|\boldsymbol{x})$ that attains the optimal value of the adversarial maximization problem (Equation 6 in the paper). However, the optimum is achieved as a limit over a sequence of distributions as described next. We shall assume *wlog* that the smallest ball that contains the set of vectors $\boldsymbol{w}$ is centered at zero and has radius $r = \max_{\bar{y}} \|\boldsymbol{w}_{\bar{y}}\|$ (Every set $\boldsymbol{w}$ can be shifted by translation to such a set, and the construction can be extended accordingly). In other words, we can assume $0 \in conv\{\boldsymbol{w}_{\bar{y}} : \|\boldsymbol{w}_{\bar{y}}\|_2 = r\}$, with $r = \max_{\bar{y}} \|\boldsymbol{w}_{\bar{y}}\|$. There are $\lambda_{\bar{y}}$ that satisfies.

$$\sum \lambda_{\bar{y}} \boldsymbol{w}_{\bar{y}} = 0$$

with $\sum \lambda_{\bar{y}} = 1$ and $\lambda_{\bar{y}} \geq 0$. Additionally $\lambda_{\bar{y}} \neq 0$ iff $\|\boldsymbol{w}_{\bar{y}}\|_2 = r$. We proceed to define a set of distributions $p_\gamma(\bar{\boldsymbol{x}}|\boldsymbol{x})$, each parameterized by a value $\gamma > 0$. The distribution $p_\gamma(\bar{\boldsymbol{x}}|\boldsymbol{x})$ has non zero mass only on $L+1$ points, and is defined as follows:

$$\begin{array}{rcl} p_\gamma(\boldsymbol{x}|\boldsymbol{x}) & = & 1-\gamma \\ p_\gamma(\sigma \dfrac{\boldsymbol{w}_{\bar{y}}}{r\gamma} + \boldsymbol{x}|\boldsymbol{x}) & = & \lambda_{\bar{y}}\gamma \quad , \quad \forall \bar{y} \end{array}$$

It is easy to see that $p_\gamma(\bar{\boldsymbol{x}}|\boldsymbol{x})$ is a valid distribution. To see that $p_\gamma(\bar{\boldsymbol{x}}|\boldsymbol{x})$ satisfies the constraints in $\mathcal{S}_{\ell_2}$ note that:

$$E_{p_\gamma(\bar{\boldsymbol{x}}|\boldsymbol{x})}(\bar{\boldsymbol{x}}) = (1-\gamma)\boldsymbol{x} + \sum_{\lambda_y \neq 0} \sigma \lambda_{\bar{y}} \frac{\boldsymbol{w}_{\bar{y}}}{r} + \lambda_{\bar{y}}\gamma \boldsymbol{x} = \boldsymbol{x}$$

$$E_{p_\gamma(\bar{\boldsymbol{x}}|\boldsymbol{x})}(\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2) = \sum_{\lambda_y \neq 0} \lambda_y \sigma\gamma \frac{\|\boldsymbol{w}_y\|_2}{r\gamma} = \sigma$$

Finally, we want to show that as $\gamma \to 0$ we obtain the optimal value of the adversarial problem. To show this, note that when γ is sufficiently small the loss $\ell(\frac{\sigma \boldsymbol{w}_{\bar{y}}}{r\gamma} + \boldsymbol{x}, y; \boldsymbol{w})$ is given by

$$\ell(\frac{\boldsymbol{w}_{\bar{y}}}{r\gamma} + \boldsymbol{x}, y; \boldsymbol{w}) = e_{\bar{y},y} + \Delta \boldsymbol{w}_{\bar{y}}^T \left(\sigma \frac{\boldsymbol{w}_{\bar{y}}}{r\gamma} + \boldsymbol{x} \right)$$

We can now write the loss corresponding to $p_\gamma(\bar{\boldsymbol{x}}|\boldsymbol{x})$ and take the limit $\gamma \to 0$.

$$E_{p_\gamma(\bar{\boldsymbol{x}}|\boldsymbol{x})}(\ell(\bar{\boldsymbol{x}}, \bar{y}; \boldsymbol{w})) =$$

$$(1-\gamma)\ell(\boldsymbol{x}, \bar{y}; \boldsymbol{w}) + \gamma \sum_{\lambda_{\bar{y}} \neq 0} \lambda_{\bar{y}} \left(e_{\bar{y},y} + \Delta \boldsymbol{w}_{\bar{y}}^T \left(\sigma \frac{\boldsymbol{w}_{\bar{y}}}{r\gamma} + \boldsymbol{x} \right) \right)$$

$$\xrightarrow{\gamma \to 0} \ell(\bar{\boldsymbol{x}}, \bar{y}; \boldsymbol{w}) + \sigma \sum_{\lambda_y \neq 0} \lambda_y \Delta \boldsymbol{w}_{\bar{y}}{}^T \frac{\boldsymbol{w}_y}{r} = \ell(\boldsymbol{x}, \bar{y}; \boldsymbol{w}) + \sigma \|\boldsymbol{w}_y\|_2$$

We obtained the optimum value of the adversarial maximization problem (see Theorem 3.1), and thus the limit of p_γ corresponds to the optimal adversary.

Incase $\boldsymbol{w}$ is not centered around zero, note that by our proof the optimal adversary is given by

$$E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}(\ell(\boldsymbol{x}, y, \boldsymbol{w})) = \sigma \min_\beta \max_{\bar{y}} \|\boldsymbol{w}_{\bar{y}} - \beta\| + \ell(\boldsymbol{x}, y, \boldsymbol{w})$$

The β that solves this optimization problem, will be the center of the smallest ball containing the set of vectors $\boldsymbol{w}_{\bar{y}}$. The construction of p_γ is similiar. Note also that the optimization problem we are considering (eq. 13) will converge to a set $\boldsymbol{w}$, centered at zero.

Appendix B

Appendix B.1 Proof of Theorem 3.5

Theorem. *$RSVM_2^2(\sigma)$ is equivalent to the problem*

$$\min_{w_y; \alpha_i; \boldsymbol{\beta}_i; \gamma_i} \frac{1}{n} \sum_i \alpha_i \sigma + \alpha_i \|\boldsymbol{x}_i\|^2 + \boldsymbol{x}_i{}^T \boldsymbol{\beta}_i + \gamma_i \quad s.t.$$

$$\forall \bar{y} \begin{bmatrix} \alpha Id & \frac{1}{2}\left(\boldsymbol{\beta}_i - \Delta^i \boldsymbol{w}_{\bar{y}}\right) \\ \frac{1}{2}\left(\boldsymbol{\beta}_i - \Delta^i \boldsymbol{w}_{\bar{y}}\right)^T & \gamma_i - e_{y_i,\bar{y}} \end{bmatrix} \succeq 0.$$

Proof. Our starting point is,

$$\begin{array}{ll} \max\limits_{p \in \mathcal{P}} & E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}[\ell(\bar{\boldsymbol{x}}; y; \boldsymbol{w})] \\ \text{s.t.} & E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}[\bar{\boldsymbol{x}}] = \boldsymbol{x} \ \ , \ E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}[\|\bar{\boldsymbol{x}} - \boldsymbol{x}\|_2^2] = \sigma \end{array}$$

which is equivalent to,

$$\begin{array}{ll} \max\limits_{p \in \mathcal{P}} & E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}[\ell(\bar{\boldsymbol{x}}; y; \boldsymbol{w})] \\ \text{s.t.} & E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}[\bar{\boldsymbol{x}}] = \boldsymbol{x} \ , \ E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}[\|\bar{\boldsymbol{x}}\|_2^2] = \sigma + \|\boldsymbol{x}\|_2^2 \end{array}$$

As before, given a labeled example $(\boldsymbol{x}; y)$, we define $\Delta \boldsymbol{w}_{\bar{y}} = \boldsymbol{w}_{\bar{y}} - \boldsymbol{w}_y$. The dual of the last problem is

$$\begin{array}{ll} \min. & \alpha\sigma + \alpha \|\boldsymbol{x}\|_2^2 + \boldsymbol{\beta}^T \boldsymbol{x} + \gamma \\ \text{s.t.} & \alpha \bar{\boldsymbol{x}}^T \bar{\boldsymbol{x}} + \boldsymbol{\beta}^T \bar{\boldsymbol{x}} + \gamma \geq e_{\bar{y},y} + \Delta \boldsymbol{w}_y^T \bar{\boldsymbol{x}} \quad \forall \bar{y} \forall \bar{\boldsymbol{x}} \end{array}$$

In the above, each constraint is quadratic in $\bar{\boldsymbol{x}}$, where α is the coefficient of the quadratic term. We note that $\alpha > 0$, since otherwise, the constraints will be violated. Hence we replace the infinitely many constraints with a constraint on the point that achieves the minimum value and get the equivalent problem:

$$\begin{array}{ll} \min. & \alpha\sigma + \alpha \|\boldsymbol{x}\|^2 + \boldsymbol{\beta}^T \boldsymbol{x} + \gamma \\ \text{s.t.} & \gamma - e_{y,\bar{y}} - \frac{(\boldsymbol{\beta} - \Delta \boldsymbol{w}_{\bar{y}})^T (\boldsymbol{\beta} - \Delta \boldsymbol{w}_{\bar{y}})}{4\alpha} \geq 0 \ , \forall \bar{y} \end{array}$$

Moving to the Schur complement we obtain the following problem,

$$\begin{array}{ll} \text{min.} & \alpha\sigma + \alpha\|\boldsymbol{x}\|^2 + \boldsymbol{\beta}^T\boldsymbol{x} + \gamma \\ \text{s.t.} & \left[\begin{array}{cc} \alpha Id & \frac{1}{2}\left(\boldsymbol{\beta} - \Delta\boldsymbol{w}_{\bar{y}}\right) \\ \frac{1}{2}\left(\boldsymbol{\beta} - \Delta\boldsymbol{w}_{\bar{y}}\right)^T & \gamma - e_{y,\bar{y}} \end{array}\right] \succeq 0 \ , \forall\bar{y}. \end{array} \tag{1}$$

and the result is immediate. □

Note that the dual of Eq. 1 is

$$\begin{array}{ll} \underset{b_y;c_y;a_y}{\text{max.}} & \sum_{\bar{y}} \Delta\boldsymbol{w}_{\bar{y}} b_{\bar{y}} + \sum_{y\neq\bar{y}} c_{\bar{y}} \\ \text{s.t.} & \sum_y \left[\begin{array}{cc} a_{\bar{y}}Id & b_{\bar{y}} \\ {b_{\bar{y}}}^T & c_{\bar{y}} \end{array}\right] = \left[\begin{array}{cc} \sigma + \|\boldsymbol{x}\|^2 & \boldsymbol{x} \\ \boldsymbol{x}^T & 1 \end{array}\right] \\ & \left[\begin{array}{cc} a_{\bar{y}}Id & b_{\bar{y}} \\ {b_{\bar{y}}}^T & c_{\bar{y}} \end{array}\right] \succeq 0 \end{array}$$

Conceptually, we construct a probability distribution over the labels $p(\cdot|\boldsymbol{x})$ such that for each point $\frac{1}{c_y}b_y$ we give probability c_y. (The positivity constraint implies that $c_y = 0$ will entail $b_y = 0$ hence there is no problem in dividing by c_y). The constraints then can be interpreted as

$$E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}[\bar{\boldsymbol{x}}] = \boldsymbol{x} \text{ and } E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}[\|\bar{\boldsymbol{x}} - \boldsymbol{x}\|^2] \leq \sigma.$$

The expected loss $E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}[\ell(\bar{\boldsymbol{x}}; y; \boldsymbol{x})]$ has exactly the optimal value of the problem. Furthermore, the optimal value is an upper bound on the expected loss for a probability that satisfies these constraints. Hence p is the desired probability. At first sight, it seems that this optimization problem requires us to solve for each example point $\boldsymbol{x}$ an SDP with complexity that scales with the dimension of $\boldsymbol{x}$. In fact, the complexity of each SDP problem (i.e. minimization of $\{\alpha_i, \boldsymbol{\beta}_i, \gamma_i\}$ for a given $\boldsymbol{w}$) can be reduced to scale with the number of classes. Intuitively, this follows from the fact that there is no point in putting adversarial noise on the space orthogonal to the space spanned by $\boldsymbol{w}$, hence the adversarial problem can be solved in that space.

Appendix B.2 Proof of Theorem 3.6

Theorem. *If $y \in \{1, -1\}$ is binary, $RSVM_2^2(\sigma)$ is equivalent to the problem*

$$\min_{\boldsymbol{w}} \frac{1}{n}\sum_i \frac{\sqrt{\sigma\|\boldsymbol{w}\|^2 + \left(1 - y\boldsymbol{w}^T\boldsymbol{x}\right)^2} + \left(1 - y\boldsymbol{w}^T\boldsymbol{x}\right)}{2} \tag{2}$$

Proof. For simplicity we let $y \in \{1, -1\}$, and as noted above, we also have $\frac{1}{2}\boldsymbol{w}_1 = -\frac{1}{2}\boldsymbol{w}_{-1} = \boldsymbol{w}$. The problem of Eq. 1 becomes,

$$\begin{array}{ll} \text{min.} & \alpha\sigma + \alpha\|\boldsymbol{x}\|^2 + \boldsymbol{\beta}^T\boldsymbol{x} + \gamma \\ \text{s.t.} & \left[\begin{array}{cc} \alpha Id & \frac{1}{2}\left(\boldsymbol{\beta} + y\boldsymbol{w}\right) \\ \frac{1}{2}\left(\boldsymbol{\beta} + y\boldsymbol{w}\right)^T & \gamma - 1 \end{array}\right] \succeq 0 \\ & \left[\begin{array}{cc} \alpha Id & \frac{1}{2}\boldsymbol{\beta} \\ \frac{1}{2}\boldsymbol{\beta}^T & \gamma \end{array}\right] \succeq 0 \end{array}$$

As noted above, the variables of the dual problem belong to the simplex (and thus define a probability distribution). Since we assume $\sigma > 0$ this probability measure can not be degenerate, hence the dual variables are not zero. By complementary slackness, at the optimal values, the matrices used in the constraints of the last problem cannot be of a full rank. Thus, using the Schur complement again we obtain,

$$\begin{aligned} \gamma - \frac{1}{4\alpha}\boldsymbol{\beta}^T\boldsymbol{\beta} &= 0 \\ \gamma - 1 - \frac{1}{4\alpha}(\boldsymbol{\beta} + y\boldsymbol{w})^T(\boldsymbol{\beta} + y\boldsymbol{w}) &= 0. \end{aligned}$$

We rewrite the problem and get,

$$\begin{array}{ll} \text{min.} & \alpha\left(\sigma + \|\boldsymbol{x} + \frac{1}{2\alpha}\boldsymbol{\beta}\|^2\right) \\ \text{s.t.} & \alpha = -\frac{2y\boldsymbol{\beta}^T\boldsymbol{w} + \|\boldsymbol{w}\|^2}{4}, \end{array}$$

which is equivalent to the problem

$$\begin{array}{ll} \text{min.} & \alpha\left(\sigma + \frac{\left(w^T x + \frac{1}{2\alpha}\boldsymbol{\beta}^T\boldsymbol{w}\right)^2}{\|\boldsymbol{w}\|^2}\right) \\ \text{s.t.} & y\boldsymbol{\beta}^T\boldsymbol{w} = -2\alpha + \frac{1}{2}\|\boldsymbol{w}\|^2 \end{array}$$

We plug the value of $y\boldsymbol{\beta}^T\boldsymbol{w}$ into the objective and get,

$$\text{min.} \frac{\alpha}{\|\boldsymbol{w}\|^2}\left(\|\boldsymbol{w}\|^2\sigma + \left(y\boldsymbol{w}^T\boldsymbol{x} - 1 + \frac{1}{4\alpha}\|w\|^2\right)^2\right). \tag{3}$$

Setting to zero the derivative of the last problem with respect to α we get,

$$\|\boldsymbol{w}\|^2\sigma + \left(y\boldsymbol{w}^T\boldsymbol{x} - 1\right)^2 - \left(\frac{1}{4\alpha}\|\boldsymbol{w}\|^2\right)^2 = 0$$

Yielding,

$$\frac{\|\boldsymbol{w}\|^2}{\alpha} = 4\sqrt{\|\boldsymbol{w}\|^2\sigma + \left(1 - y\boldsymbol{w}^T x\right)^2}.$$

Plugging the last result back into Eq. 3 yields the desired result. □