Supplementary Information: Ellipsoidal Multiple Instance Learning

Gabriel Krummenacher GABRIEL.KRUMMENACHER@INF.ETHZ.CH
Department of Computer Science, ETH Zurich, Switzerland

Cheng Soon Ong CHENGSOON.ONG@UNIMELB.COM.AU
National ICT Australia, Victoria Research Laboratory, Melbourne, Australia

Joachim M. Buhmann JBUHMANN@INF.ETHZ.CH
Department of Computer Science, ETH Zurich, Switzerland

A. Proofs for Lemmas 3 and 4

Proof of Lemma 3. We make use of the following multivariate Chebyshev's inequality to proof Lemma 3

Theorem 1. *(Marshall & Olkin, 1960), (Bertsimas & Popescu, 2001), (Lanckriet et al., 2002)*

$$\begin{aligned} \sup_{\mathbf{y}\sim(\bar{\mathbf{y}},\Sigma_{\mathbf{y}})} \Pr[\mathbf{y} \in \mathcal{S}] &= \frac{1}{1+d^2}, \text{with} \\ d^2 &= \inf_{\mathbf{y}\in\mathcal{S}} (\mathbf{y}-\bar{\mathbf{y}})^\top \Sigma_{\mathbf{y}}^{-1} (\mathbf{y}-\bar{\mathbf{y}}) \end{aligned} \tag{1}$$

Where $\mathbf{y}$ is a random vector, the supremum is over all distributions for $\mathbf{y}$ with mean $\bar{\mathbf{y}}$ and covariance matrix $\Sigma_{\mathbf{y}}$ and $\mathcal{S}$ is a given convex set.

Now, setting $\mathcal{S} = \{\mathbf{x}_i | y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \leq 1 - \xi_i\}$ we get the claimed equality. □

Proof of Lemma 4. We follow the proof in (Lanckriet et al., 2002) to find a closed form expression for $\inf_{\mathbf{x}_i|y_i(\langle \mathbf{w},\mathbf{x}_i\rangle+b)\leq 1-\xi_i} (\mathbf{x}_i - \mathbf{q}_i)^\top \mathbf{\Sigma}_i^{-1} (\mathbf{x}_i - \mathbf{q}_i)$.

If $y_i(\langle \mathbf{w}, \mathbf{q}_i\rangle + b) \leq 1 - \xi_i$ then we can just set $\mathbf{x}_i = \mathbf{q}_i$ and the infimum becomes 0.

To show the other case of $y_i(\langle \mathbf{w}, \mathbf{q}_i\rangle + b) \geq 1 - \xi_i$ we write $d^2 = \inf_{\langle \mathbf{c},\mathbf{k}\rangle \geq f} \langle \mathbf{k}, \mathbf{k}\rangle$, where $\mathbf{k} = \mathbf{\Sigma}_i^{-1/2}(\mathbf{x}_i - \mathbf{q}_i)$, $\mathbf{c}^\top = -y_i \mathbf{w}^\top \mathbf{\Sigma}_i^{1/2}$ and $f = y_i(\langle \mathbf{w}, \mathbf{q}_i\rangle + b) - 1 + \xi_i \geq 0$. We form the Lagrangian:

$$L(\mathbf{k}, \lambda) = \langle \mathbf{k}, \mathbf{k}\rangle + \lambda(f - \langle \mathbf{c}, \mathbf{k}\rangle)$$

and maximize it with respect to the dual variable $\lambda \geq 0$ and minimize with respect to the primal variable $\mathbf{k}$. At the optimum we get $2\mathbf{k} = \lambda \mathbf{c}$ and $f = \langle \mathbf{c}, \mathbf{k}\rangle$. So, $\lambda = \frac{2f}{\langle \mathbf{c},\mathbf{c}\rangle}$ such that indeed $\lambda \geq 0$ because $f > 0$. Also, $\mathbf{k} = \frac{f\mathbf{c}}{\langle \mathbf{c},\mathbf{c}\rangle}$. This yields

$$\frac{(y_i(\langle \mathbf{w}, \mathbf{q}_i\rangle + b) - 1 + \xi_i)^2}{\mathbf{w}^\top \mathbf{\Sigma}_i \mathbf{w}}$$

Combining both cases $y_i(\langle \mathbf{w}, \mathbf{q}_i\rangle + b) \leq 1 - \xi_i$ and $y_i(\langle \mathbf{w}, \mathbf{q}_i\rangle + b) \geq 1 - \xi_i$ we get the right hand side of Lemma 4:

$$\frac{\max(0, y_i(\langle \mathbf{w}, \mathbf{q}_i\rangle + b) - 1 + \xi_i)^2}{\mathbf{w}^\top \mathbf{\Sigma}_i \mathbf{w}}$$

□

B. Derivation of SOCP

First we write $\sqrt{\mathbf{w}^\top \mathbf{P}_i \mathbf{w}}$ as $\|\mathbf{A}_i \mathbf{w}\|$ with $\mathbf{P}_i = \mathbf{A}_i^\top \mathbf{A}_i$.

Then we replace the hinge-loss type part of the objective function in Equation (21) in the main article with the following constraints, by introducing slack variables ξ_i:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \ & \frac{\lambda}{2}\|\mathbf{w}\|^2 - \sum_{y_+} \left(\left\langle \mathbf{w}, \frac{\mathbf{P}_i \mathbf{w}_k}{\sqrt{\mathbf{w}_k^\top \mathbf{P}_i \mathbf{w}_k}} + \mathbf{q}_i \right\rangle + b \right) + \sum_{i=1}^{B} \xi_i \\ \text{s.t.} \ & \|\mathbf{A}_i \mathbf{w}\| + \mathbf{w}^\top \mathbf{q}_i + b \leq \xi_i - 1, \quad \forall i\colon y_i = -1 \\ & \|\mathbf{A}_i \mathbf{w}\| + \mathbf{w}^\top \mathbf{q}_i + b \leq \xi_i + 1, \quad \forall i\colon y_i = +1 \\ & \mathbf{0} \leq \boldsymbol{\xi} \end{aligned} \tag{2}$$

Where $\sum_{y_+}$ means sum over all i for which $y_i = +1$.

Next replace the remaining objective function with θ

and add it as a constraint:

$$\begin{aligned}
&\min_{\theta,\mathbf{w},b,\boldsymbol{\xi}} \theta \\
\text{s.t.}\quad & \frac{\lambda}{2}\|\mathbf{w}\|^2 - \sum_{y+}\left(\left\langle \mathbf{w}, \frac{\mathbf{P}_i\mathbf{w}_k}{\sqrt{\mathbf{w}_k^\top\mathbf{P}_i\mathbf{w}_k}} + \mathbf{q}_i\right\rangle + b\right) + \sum_{i=1}^{B}\xi_i \le \theta \\
& \|\mathbf{A}_i\mathbf{w}\| + \mathbf{w}^\top\mathbf{q}_i + b \le \xi_i - 1, \quad \forall i\colon y_i = -1 \\
& \|\mathbf{A}_i\mathbf{w}\| + \mathbf{w}^\top\mathbf{q}_i + b \le \xi_i + 1, \quad \forall i\colon y_i = +1 \\
& \mathbf{0} \le \boldsymbol{\xi}
\end{aligned} \tag{3}$$

Finally we see that this quadratic constraint is equivalent to the SOC constraint in Equation 22 in the main article.

C. Distance between an ellipsoid and a hyperplane

Proof of Proposition 1. We would like to minimize the squared distance between a point x on the hyperplane, and a point z on the ellipsoid. This can be expressed as the following constrained optimisation problem:

$$\begin{aligned}
\min_{x,z}\quad & \|z-x\|^2 \\
\text{s.t.}\quad & (z-q)^\top P^{-1}(z-q) = 1 \\
& w^\top x + b = 0
\end{aligned}$$

We form the Lagrangian, using multiplier η for the ellipsoidal constraint and γ for the hyperplane.

$$\mathcal{L}(x,z,\eta,\gamma) = \|z-x\|^2 + \eta(z-q)^\top P^{-1}(z-q) - \eta + \gamma w^\top x + \gamma b \tag{4}$$

Taking the gradient of Equation (4) with respect to x and z respectively, and setting it to zero gives

$$2(z-x) = \gamma w \tag{5}$$

$$2(z-x) + 2\eta P^{-1}(z-q) = 0 \tag{6}$$

By substituting Equation (5) into Equation (6), we obtain that

$$z = -\frac{\gamma}{2\eta}Pw + q \tag{7}$$

and using this in Equation (5) gives

$$x = -\frac{\gamma}{2\eta}Pw + q - \frac{\gamma}{2}w \tag{8}$$

By substituting Equation (7) and (8) into the Lagrangian (Equation (4)) we obtain an expression only in the dual variables.

$$\begin{aligned}
\mathcal{L}(\eta,\gamma) &= -\frac{\gamma^2}{4}\|w\|^2 + \eta\left(\frac{\gamma}{2\eta}Pw\right)^\top P^{-1}\left(\frac{\gamma}{2\eta}Pw\right) \\
&\quad -\eta + \gamma w^\top\left(-\frac{\gamma}{2\eta}Pw + q - \frac{\gamma}{2}w\right) + \gamma b \\
&= -\frac{\gamma^2}{4}\|w\|^2 - \frac{\gamma^2}{4\eta}w^\top Pw - \eta + \gamma w^\top q + \gamma b
\end{aligned}$$

We would like to maximize the dual with respect to η and γ, and this point is achieved at the stationary points

$$\frac{\partial\mathcal{L}}{\partial\gamma} = -\frac{\gamma}{2}\|w\|^2 - \frac{\gamma}{2\eta}w^\top Pw + w^\top q + b = 0 \tag{9}$$

and

$$\frac{\partial\mathcal{L}}{\partial\eta} = \frac{\gamma^2}{4\eta^2}w^\top Pw - 1 = 0 \tag{10}$$

Equation (10) implies

$$\eta = \pm\frac{\gamma}{2}\sqrt{w^\top Pw} \tag{11}$$

Substituting the expression for η (Equation (11)) into the stationary condition for γ (Equation (9)) gives

$$-\frac{\gamma}{2}\|w\|^2 \pm \sqrt{w^\top Pw} + w^\top q + b = 0 \tag{12}$$

Observe from Equation (5) that the distance from the ellipsoid to the hyperplane is given by $\frac{\gamma}{2}\|w\|$ which from Equation (12) is given by

$$\frac{\gamma}{2}\|w\| = \frac{1}{\|w\|}\left(\pm\sqrt{w^\top Pw} + w^\top q + b\right)$$

When the ellipsoid intersects the hyperplane, we would like the point on the ellipsoid furthest away from the hyperplane, which is given by the solution of the following constrained optimisation problem.

$$\begin{aligned}
\max_{x,z}\quad & \|z-x\|^2 \\
\text{s.t.}\quad & (z-q)^\top P^{-1}(z-q) = 1 \\
& w^\top x + b = 0
\end{aligned}$$

Since the only difference is from finding the minimum to finding the maximum, the above derivation remains identical and the theorem follows.

□

D. Gradients

The gradient of the smooth hinge loss with respect to $\mathbf{w}$ and b is given respectively by

$$\frac{\partial}{\partial\mathbf{w}}\ell_\delta(\mathbf{x}_i, y_i, \mathbf{w}, b) = \begin{cases} \frac{1-y_i f(x_i)}{\delta}\cdot -y_i\frac{\partial}{\partial\mathbf{w}}f(\mathbf{x}_i) & \text{if } \Phi \\ -y_i\cdot\frac{\partial}{\partial\mathbf{w}}f(\mathbf{x}_i) & \text{if } \Psi \\ 0 & \text{if } \Omega \end{cases} \tag{13}$$

$$\frac{\partial}{\partial b}\ell_\delta(\mathbf{x}_i, y_i, \mathbf{w}, b) = \begin{cases} \frac{1-y_i f(x_i)}{\delta}\cdot -y_i\frac{\partial}{\partial b}f(\mathbf{x}_i) & \text{if } \Phi \\ -y_i\cdot\frac{\partial}{\partial b}f(\mathbf{x}_i) & \text{if } \Psi \\ 0 & \text{if } \Omega \end{cases} \tag{14}$$

Where

$$\Phi \equiv 1 - \delta < y_i \cdot f(\mathbf{P}_i; \mathbf{q}_i) \leq 1$$
$$\Psi \equiv y_i \cdot f(\mathbf{P}_i; \mathbf{q}_i) \leq 1 - \delta$$
$$\Omega \equiv y_i \cdot f(\mathbf{P}_i; \mathbf{q}_i) > 1$$

And where the gradient of the ellipsoid predictor $f(\mathbf{q}; \mathbf{P}) = \sqrt{\mathbf{w}^\top \mathbf{P} \mathbf{w}} + \mathbf{w}^\top \mathbf{q} + b$ is given by Equation (15) and Equation (16).

$$\frac{\partial}{\partial \mathbf{w}} f(\mathbf{q}; \mathbf{P}) = \mathbf{q} + \frac{\mathbf{P}\mathbf{w}}{\sqrt{\mathbf{w}^\top \mathbf{P} \mathbf{w}}} \tag{15}$$
$$\frac{\partial}{\partial b} f(\mathbf{q}; \mathbf{P}) = 1 \tag{16}$$

References

Bertsimas, Dimitris and Popescu, Ioana. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15: 780–804, 2001.

Lanckriet, Gert R. G., Ghaoui, Laurent El, Bhattacharyya, Chiranjib, and Jordan, Michael I. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3:555–582, 2002.

Marshall, A.W. and Olkin, I. Multivariate Chebyshev inequalities. *The Annals of Mathematical Statistics*, 31(4):1001–1014, 1960.