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)

\[
\sup_{y \sim \Sigma_y} \Pr\{y \in S\} = \frac{1}{1 + d^2}, \text{ with } d^2 = \inf_{y \in S} (y - \bar{y})^\top \Sigma_y^{-1} (y - \bar{y})
\]

(1)

Where \( y \) is a random vector, the supremum is over all distributions for \( y \) with mean \( \bar{y} \) and covariance matrix \( \Sigma_y \) and \( S \) is a given convex set.

Now, setting \( S = \{x_i | y_i((\langle w, x_i \rangle + b) \leq 1 - \xi_i\} \) 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 \( \inf_{x_i,|y_i((\langle w, x_i \rangle + b) \leq 1 - \xi_i}(x_i - q_i)^\top \Sigma_i^{-1}(x_i - q_i) \).

If \( y_i((\langle w, q_i \rangle + b) \leq 1 - \xi_i \) then we can just set \( x_i = q_i \) and the infimum becomes 0.

To show the other case of \( y_i((\langle w, q_i \rangle + b) \geq 1 - \xi_i \) we write \( d^2 = \inf_{c | \langle c, k \rangle \geq f}(\langle c, k \rangle), \) where \( k = \Sigma_i^{-1/2}(x_i - q_i) \), \( c^\top = -y_i w^\top \Sigma_i^{1/2} \) and \( f = y_i((\langle w, q_i \rangle + b) - 1 + \xi_i \geq 0. \)

We form the Lagrangian:

\[
L(k, \lambda) = \langle c, k \rangle + \lambda(f - \langle c, k \rangle)
\]

and maximize it with respect to the dual variable \( \lambda \geq 0 \) and minimize with respect to the primal variable \( k \).

At the optimum we get \( 2k = \lambda c \) and \( f = \langle c, k \rangle \). So, \( \lambda = \frac{-f}{\langle c, c \rangle} \) such that indeed \( \lambda \geq 0 \) because \( f > 0 \). Also, \( k = \frac{f}{\langle c, c \rangle} \). This yields

\[
\frac{(y_i((\langle w, q_i \rangle + b) - 1 + \xi_i)^2}{w^\top \Sigma_i w}
\]

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

\[
\max(0, y_i((\langle w, q_i \rangle + b) - 1 + \xi_i)^2}{w^\top \Sigma_i w}
\]

B. Derivation of SOCP

First we write \( \sqrt{w^\top P_i w} \) as \( \|A_i w\| \) with \( P_i = A_i^\top A_i \).

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 \( \xi_i \):

\[
\min_{w, \xi} \frac{\lambda}{2} \|w\|^2 - \sum_{y_i} \left( \langle w, \frac{P_i w_k}{\sqrt{w_i^\top P_i w_k}} + q_i \rangle + b \right) + \sum_{i=1}^{B} \xi_i,
\]

s.t. \( \|A_i w\| + w^\top q_i + b \leq \xi_i - 1, \quad \forall i: y_i = -1 \)
\( \|A_i w\| + w^\top q_i + b \leq \xi_i + 1, \quad \forall i: y_i = +1 \)
\( 0 \leq \xi \)

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

Next replace the remaining objective function with \( \theta \)
and add it as a constraint:
\[
\min_{\theta, w, b, \ell} \quad \theta
\s.t. \quad \frac{\lambda}{2} \|w\|^2 - \sum_{y_i} \left( \left\langle w, \frac{P_i w_k}{\sqrt{w_k P_i w_k}} + q_i \right\rangle + b \right) + \sum_{i=1}^{B} \xi_i \leq 0
\]

By substituting Equation (5) into Equation (6), we get
\[
\|
A, w\| + w^\top q_i + b \leq \xi_i - 1, \quad \forall i: y_i = -1
\]
\[
\|
A, w\| + w^\top q_i + b \leq \xi_i + 1, \quad \forall i: y_i = +1
\]
\[
0 \leq \xi_i
\]

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:

\[
\min_{x, z} \quad \|z - x\|^2
\s.t. \quad (z - q)^\top P^{-1}(z - q) = 1
\]
\[
w^\top x + b = 0
\]

We form the Lagrangian, using multiplier \( \eta \) for the ellipsoidal constraint and \( \gamma \) for the hyperplane.

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

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

\[
2(z - x) = \gamma w
\]
\[
2(z - x) + 2\eta P^{-1}(z - q) = 0
\]

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

\[
z = -\frac{\gamma}{2\eta} P w + q
\]

and using this in Equation (5) gives

\[
x = -\frac{\gamma}{2\eta} P w + q - \frac{\gamma}{2} w
\]

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

\[
\mathcal{L}(\eta, \gamma) = -\frac{\gamma^2}{4} \|w\|^2 + \eta \left( \frac{\gamma}{2\eta} P w \right)^\top P^{-1} \left( \frac{\gamma}{2\eta} P w \right)
\]
\[
- \eta + \gamma w^\top \left( -\frac{\gamma}{2\eta} P w + \frac{\gamma}{2} w \right) + \gamma b
\]
\[
= -\frac{\gamma^2}{4} \|w\|^2 - \frac{\gamma^2}{4\eta} w^\top P w - \eta + \gamma w^\top q + \gamma b
\]

We would like to maximize the dual with respect to \( \eta \) and \( \gamma \), and this point is achieved at the stationary points

\[
\frac{\partial \mathcal{L}}{\partial \gamma} = -\frac{\gamma}{2} \|w\|^2 - \frac{\gamma}{2\eta} w^\top P w + w^\top q + b = 0
\]

and

\[
\frac{\partial \mathcal{L}}{\partial \eta} = \frac{\gamma^2}{4\eta^2} w^\top P w - 1 = 0
\]

Equation (10) implies

\[
\eta = \pm \frac{\gamma}{2} \sqrt{w^\top P w}
\]

Substituting the expression for \( \eta \) (Equation (11)) into the stationary condition for \( \gamma \) (Equation (9)) gives

\[
-\frac{\gamma}{2} \|w\|^2 \pm \sqrt{w^\top P w} - w^\top q + b = 0
\]

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

\[
\frac{\gamma}{2} \|w\| = \frac{1}{\|w\|} \left( \pm \sqrt{w^\top P w} - w^\top q + b \right)
\]

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

\[
\max_{x, z} \quad \|z - x\|^2
\s.t. \quad (z - q)^\top P^{-1}(z - q) = 1
\]
\[
w^\top x + b = 0
\]

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 \( w \) and \( b \) is given respectively by

\[
\frac{\partial}{\partial w} \ell(x_i, y_i, w, b) = \begin{cases}
1 - y_i f(x_i) & \text{if } \Phi \\
y_i \frac{\partial}{\partial w} f(x_i) & \text{if } \Psi \\
0 & \text{if } \Omega
\end{cases}
\]

\[
\frac{\partial}{\partial b} \ell(x_i, y_i, w, b) = \begin{cases}
1 - y_i f(x_i) & \text{if } \Phi \\
y_i \frac{\partial}{\partial b} f(x_i) & \text{if } \Psi \\
0 & \text{if } \Omega
\end{cases}
\]
Supplementary Information: Ellipsoidal Multiple Instance Learning

Where
\[
\Phi \equiv 1 - \delta < y_i \cdot f(P_i; q_i) \leq 1
\]
\[
\Psi \equiv y_i \cdot f(P_i; q_i) \leq 1 - \delta
\]
\[
\Omega \equiv y_i \cdot f(P_i; q_i) > 1
\]

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

\[
\frac{\partial}{\partial w} f(q; P) = q + \frac{Pw}{\sqrt{w^\top P w}} \quad (15)
\]
\[
\frac{\partial}{\partial b} f(q; P) = 1 \quad (16)
\]

References

