Appendix: Proofs

Theorem 2

The proof proceeds by induction. The initial parameters $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\Sigma_1 = aI$ can be trivially written in the desired form. For the induction step we first substitute (16) in (8) and get,

$$\boldsymbol{\mu}_{i+1} = \boldsymbol{\mu}_i + \alpha_i y_i \Sigma_i \boldsymbol{x}_i = \sum_{p=1}^{i-1} \left(\nu_p^{(i)} + \alpha_i y_i \sum_{q=1}^{i-1} \pi_{p,q}^{(i)} \boldsymbol{x}_q^\top \boldsymbol{x}_i \right) \boldsymbol{x}_p + a \boldsymbol{x}_i \ ,$$

which is of the desired form with

$$\nu_i^{(i+1)} = a \qquad \text{and} \qquad \nu_p^{(i+1)} = \nu_p^{(i)} + \alpha_i y_i \sum_{q}^{i-1} \pi_{p,q}^{(i)} \boldsymbol{x}_q^\top \boldsymbol{x}_i \ \text{ for } p < i \ . \tag{20}$$

A similar elementary calculation can be done for the covariance to obtain

$$\pi_{p,q}^{(i+1)} = -\beta_i \sum_{r,s} \pi_{p,r}^{(i)} \pi_{s,q}^{(i)} \boldsymbol{x}_r^\top \boldsymbol{x}_s + \pi_{p,q}^{(i)} \ , \ \pi_{p,i}^{(i+1)} = \pi_{i,p}^{(i+1)} = -\beta_i a \sum_{p,r=1}^{i-1} \pi_{p,r}^{(i)} \left(\boldsymbol{x}_r^\top \boldsymbol{x}_i \right) \ , \ \pi_{i,i}^{(i+1)} = -\beta_i a^2 \ , \tag{21}$$

for $p = 1 \ldots i-1$, where

$$\beta_i = (\alpha_i \phi) \,/\, \left(\sqrt{\boldsymbol{x}_i^\top \Sigma_{i+1} \boldsymbol{x}_i} + (\boldsymbol{x}_i^\top \Sigma_i \boldsymbol{x}_i) \alpha_i \phi \right) = (\alpha_i \phi) \,/\, (\sqrt{u_i} + v_i \alpha_i \phi) \ . \tag{22}$$

Finally, we show that the coefficients $\{\nu_p^{(i)}\}$ and $\{\pi_{p,q}^{(i)}\}$ depend on the data only through inner products. From (11) we have that both m_i and v_i can be written only using inner products. From (14), α_i can also be written as a function of inner products, which in turn, together with (12) implies that u_i can be written that way. Therefore, β_i can also be written as a function of inner products. Finally, using (20) and (21) we conclude that $\{\nu_p^{(i)}\}$ and $\{\pi_{p,q}^{(i)}\}$ depend on the data only through inner products.

Theorem 4

We prove the theorem in four steps. First, we define a notion of confidence loss. Second, we prove an auxilary lemma which relates the update to an update of a Euclidean projection. Third, we use the auxilary lemma to bound the cumulative confidence loss on a run of the algorithm. Finally, we prove the theorem using this bound and additional properties of the confidence loss.

Confidence Loss

Before analyzing the algorithm we define our *confidence loss* family of smooth convex loss functions. Given an input example $(\boldsymbol{x}, y)$ and a model $(\boldsymbol{\mu}, \Sigma)$, the confidence loss will be a function of the parameters m, v of the induced margin Gaussian $m = y\,(\boldsymbol{\mu} \cdot \boldsymbol{x})$ and $v = \boldsymbol{\mu}^\top \Sigma \boldsymbol{\mu}$. In our model, m plays a role similar to the geometric margin in standard margin-based analyses. However, the scale of m is not fixed, as it depends on the variance v: the magnitude of margin random variable M is large if the variance is large. We thus define our loss function to be a function of the margin m normalized by the standard deviation:

$$\bar{m} = \frac{m}{\sqrt{v}} \ .$$

By analogy with hinge-loss-based losses, the confidence loss is given by a family of functions f_ϕ parameterized by $\phi \geq 0$ that bound the 0-1 loss as follows:

$$\ell_\phi\left(\bar{m}\right) = \begin{cases} 0 & \bar{m} \geq \phi \\ f_\phi(\bar{m}) & \bar{m} < \phi \ , \end{cases} \tag{23}$$

where $f_\phi(x)$ is a monotonically decreasing function that satisfies $f_\phi(\phi) = 0$. For reasons that will become clear in what follows, we use the following f_ϕ in our analysis:

$$f_\phi(\bar{m}) = \frac{\left(-\bar{m}\psi + \sqrt{\bar{m}^2 \frac{\phi^4}{4} + \phi^2 \xi} \right)^2}{\phi^2 \xi} \ , \tag{24}$$

where ψ and ξ are defined above (13). The following lemma summarizes its main properties:

Lemma 5 *The function f_ϕ defined in* (24) *satisfies the following:*

1. $f_\phi(\phi) = 0$ *and* $f_\phi(0) = 1$.
2. $f_\phi(x)$ *is convex and decreasing for* $x \leq \phi$.
3. *If* $x \leq 0$ *then* $f_\phi(x) \geq 1$.
4. $f_\phi(x) \approx x^2 \frac{1+\phi^2}{\phi^2}$ *for* $x \ll -2\sqrt{\frac{1+\phi^2}{\phi^2}}$.
5. $f_\phi(x) \approx A\phi^2$ *for* $x \lesssim \phi$ *for some* $A > 0$.
6. $\ell_\phi\left(\frac{m_i}{\sqrt{v_i}}\right) = \frac{1+\phi^2}{\phi^2}\,\alpha_i^2 v_i$ *(Eqns. (14,11)).*

Proof: The first property can easily be verified via substitution. For the second property, we note that $f_\phi(x)$ is proportional to $g_\phi^2(x)$ for,

$$g_\phi(x) = -x\psi + \sqrt{x^2\frac{\phi^4}{4} + \phi^2\xi}\,,$$

and show that $g_\phi(x) \geq 0$ for $x \leq \phi$. Clearly it is correct if $x \leq 0$. We thus assume that $0 \leq x \leq \phi$ and get

$$\begin{aligned}
g_\phi(x) \geq 0 \quad &\Leftrightarrow \quad \sqrt{x^2\frac{\phi^4}{4} + \phi^2\xi} \geq x\psi \\
&\Leftrightarrow \quad x^2\frac{\phi^4}{4} + \phi^2\xi \geq x^2\psi^2 \\
&\Leftrightarrow \quad \phi^2(1+\phi^2) \geq x^2\left(\left(1+\frac{\phi^2}{2}\right)^2 - \frac{\phi^4}{4}\right) \\
&\Leftrightarrow \quad \phi^2(1+\phi^2) \geq x^2\left(1+\phi^2\right) \\
&\Leftrightarrow \quad \phi^2 \geq x^2\,,
\end{aligned}$$

which verifies the property of $g_\phi(x)$. We now analyze $g_\phi^2(x)$. Its first and second derivatives are

$$\begin{aligned}
\frac{d(g_\phi^2(x))}{dx} &= 2g_\phi(x)g_\phi'(x) \\
\frac{d^2(g_\phi^2(x))}{dx} &= 2\left(g_\phi'(x)\right)^2 + 2g_\phi(x)g_\phi''(x).
\end{aligned}$$

Since $g_\phi(x) \geq 0$ then $g_\phi^2(x)$ is decreasing and convex iff $g_\phi(x)$ is decreasing and convex.

We thus analyze $g_\phi(x)$ for $x \leq \phi$. Its first derivative is

$$g_\phi'(x) = -\psi + \frac{x\frac{\phi^4}{4}}{\sqrt{x^2\frac{\phi^4}{4} + \phi^2\xi}}\,.$$

It can be easily verified that $g_\phi'(\phi) < 0$. We compute its second derivative (omitting the constant of $\phi^4/4$):

$$\begin{aligned}
g_\phi''(x) &= \frac{1}{\sqrt{x^2\frac{\phi^4}{4} + \phi^2\xi}} - x\left(x^2\frac{\phi^4}{4} + \phi^2\xi\right)^{-\frac{3}{2}}\left(x\frac{\phi^4}{4}\right) \\
&= \frac{x^2\frac{\phi^4}{4} + \phi^2\xi - x^2\frac{\phi^4}{4}}{\left(x^2\frac{\phi^4}{4} + \phi^2\xi\right)^{\frac{3}{2}}} \\
&= \frac{\phi^2\xi}{\left(x^2\frac{\phi^4}{4} + \phi^2\xi\right)^{\frac{3}{2}}} \geq 0\,.
\end{aligned}$$

We thus established that $g_\phi(x)$ is strictly convex in the range, and since its first derivative is negative at $x = \phi$, it is also negative for $x \leq \phi$, which concludes the proof of property 2. Property 3 follows directly from the first two properties.

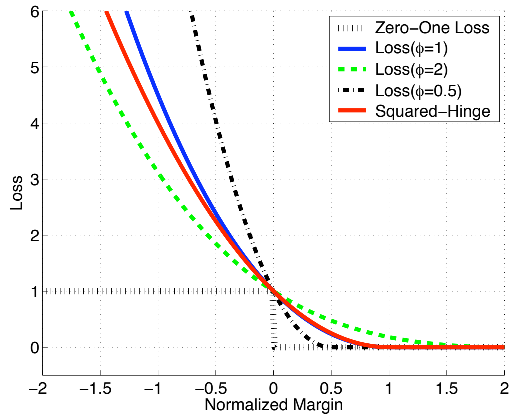

Figure 4: Squared hinge loss, $0-1$ loss, and $\ell_\phi(\cdot)$ for various values of ϕ as functions of the (normalized) margin.

For property 4, if $x \ll -2\sqrt{\frac{1+\phi^2}{\phi^2}}$ then $\sqrt{x^2\frac{\phi^4}{4}+\phi^2\xi} \approx \sqrt{x^2\frac{\phi^4}{4}} = -x\frac{\phi^2}{2}$. In this case

$$f_\phi(x) \approx \frac{\left(-x\psi - x\frac{\phi^2}{2}\right)^2}{\phi^2\xi} = x^2\frac{1+\phi^2}{\phi^2} \ .$$

For property 5 we show that the first derivative of $f_\phi(x)$ vanishes at $x=\phi$, indeed,

$$f'_\phi(\phi) = 2g_\phi(\phi)g'_\phi(\phi) = 0 \ ,$$

since $g_\phi(\phi)=0$. Thus, the first two coefficients of the Taylor expansion of $f_\phi(x)$ at $x=\phi$ vanish. The third coefficient is non-negative due the convexity of $f_\phi(x)$ at $x=\phi$.

Finally, property 6 follows directly from the definitions of $\ell_\phi(\bar{m})$, α_i and v. ■

From the first three properties we see that the confidence loss upper-bounds the 0-1 loss. Furthermore, from properties 4 and 5 we see that $\ell_\phi(x)$ is quadratic both for $x \ll 0$ and for x in the region where $\ell_\phi(x)$ is close to zero. In this respect, $\ell_\phi(x)$ behaves similarly to the squared hinge loss $\max\{1-x,0\}^2$. (Note that in the analysis of the PA algorithms [4], the squared optimal value of the Lagrange multiplier α_i^2 is proportional to the squared hinge loss. Interestingly, this also holds in our case for a much more complicated form for α_i.) Graphs of $\ell_\phi(\cdot)$ for various values of ϕ are given in Fig. 4, together with the squared hinge loss and the 0-1 loss. From the figure we see a trade-off in the value of ϕ: larger ϕ yields a tighter bound on the 0-1 loss for $\bar{m} \leq 0$, while smaller ϕ yields a tighter bound for $\bar{m} \geq 0$. This property shows up also in parameterized versions of the hinge loss [20]. The confidence loss is carefully designed to support the analysis of the next section.

It is worth recalling here the tight connection in this work between the algebraic notion of margin and the margin parameter ϕ on the one hand and the probabilistic notion of confidence and the confidence parameter η on the other. We achieve this by linking the margin parameter and the confidence parameter through the cumulative function of the normal distribution $\eta = \Phi(\phi)$.

Auxilary Lemma

Lemma 6 *Fix an iteration i and assume that $\boldsymbol{\mu}_i$,Σ_i and u_i (defined in (11)) are constants. Then the following two vectors are equal :*

- *The vector $\boldsymbol{\mu}_{i+1}$ defined in (9)*
- *The solution $\tilde{\boldsymbol{\mu}}_{i+1}$ of the following projection problem:*

$$\tilde{\boldsymbol{\mu}}_{i+1} = \arg\min_{\boldsymbol{\mu}} \frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_i)^\top \Sigma_i^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_i)^\top \quad (25)$$

$$\text{s.t.} \quad y_i(\boldsymbol{\mu}\cdot\boldsymbol{x}_i) \geq \phi\sqrt{u_i} \quad (26)$$

Bounding the Confidence Loss

The following lemma gives an upper bound on the cumulative confidence loss on a run of the algorithm:

Lemma 7 *Let $(\boldsymbol{x}_1, \boldsymbol{y}_1) \ldots (\boldsymbol{x}_n, \boldsymbol{y}_n)$ be an input sequence for the algorithm of Fig. 1, initialized with $(\mathbf{0}, I)$, with $\boldsymbol{x}_i \in \mathbb{R}^d$ and $\boldsymbol{y}_i \in \{-1, +1\}$. Assume there exist $\boldsymbol{\mu}^*$ and Σ^* such that for all i for which the algorithm made an update ($\alpha_i > 0$),*

$$\boldsymbol{\mu}^{*\top} \boldsymbol{x}_i y_i \geq \boldsymbol{\mu}_{i+1}^\top \boldsymbol{x}_i y_i \quad \text{and} \quad \boldsymbol{x}_i^\top \Sigma^* \boldsymbol{x}_i \leq \boldsymbol{x}_i^\top \Sigma_{i+1} \boldsymbol{x}_i \quad . \tag{27}$$

Let $\zeta_i = \alpha_i \phi / \sqrt{u_i}$. Then, the following bound holds:

$$\sum_i \ell_\phi \left(\frac{m_i}{\sqrt{v_i}} \right) \quad \leq \quad \frac{1+\phi^2}{\phi^2} \left(2D_{KL} \left(\mathcal{N}\left(\boldsymbol{\mu}^*, \Sigma^*\right) \| \mathcal{N}\left(\boldsymbol{\mu}_1, \Sigma_1\right) \right) + \boldsymbol{\mu}^{*\top} \left(\sum_i \zeta_i \boldsymbol{x}_i \boldsymbol{x}_i^\top \right) \boldsymbol{\mu}^* \right) \ . \tag{28}$$

Proof: From (9), we obtain

$$\Sigma_{i+1}^{-1} = \Sigma_i^{-1} + \zeta_i \boldsymbol{x}_i \boldsymbol{x}_i^\top \ . \tag{29}$$

Let

$$\Delta_i = 2\mathrm{D}_{\mathrm{KL}} \left(\mathcal{N}\left(\boldsymbol{\mu}^*, \Sigma^*\right) \| \mathcal{N}\left(\boldsymbol{\mu}_i, \Sigma_i\right) \right) - 2\mathrm{D}_{\mathrm{KL}} \left(\mathcal{N}\left(\boldsymbol{\mu}^*, \Sigma^*\right) \| \mathcal{N}\left(\boldsymbol{\mu}_{i+1}, \Sigma_{i+1}\right) \right) \ .$$

We bound $\sum_i \Delta_i$ from above and below, starting with the upper bound. Using the fact that the sum is telescopic, and substituting in the initial values $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\Sigma_1 = I$, we obtain

$$\begin{aligned} \sum_i \Delta_i &= 2\mathrm{D}_{\mathrm{KL}} \left(\mathcal{N}\left(\boldsymbol{\mu}^*, \Sigma^*\right) \| \mathcal{N}\left(\boldsymbol{\mu}_1, \Sigma_1\right) \right) - 2\mathrm{D}_{\mathrm{KL}} \left(\mathcal{N}\left(\boldsymbol{\mu}^*, \Sigma^*\right) \| \mathcal{N}\left(\boldsymbol{\mu}_{n+1}, \Sigma_{n+1}\right) \right) \\ &\leq 2\mathrm{D}_{\mathrm{KL}} \left(\mathcal{N}\left(\boldsymbol{\mu}^*, \Sigma^*\right) \| \mathcal{N}\left(\boldsymbol{\mu}_1, \Sigma_1\right) \right) \ . \end{aligned} \tag{30}$$

We now give a lower bound for Δ_i. Writing explicitly the definition of the Kullback-Leibler divergence we get,

$$\begin{aligned} \Delta_i \quad &= \quad \log\left(\frac{\det \Sigma_i}{\det \Sigma^*}\right) + \mathrm{Tr}\left(\Sigma_i^{-1}\Sigma^*\right) + (\boldsymbol{\mu}_i - \boldsymbol{\mu}^*)^\top \Sigma_i^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}^*) - d \\ &\quad - \left\{ \log\left(\frac{\det \Sigma_{i+1}}{\det \Sigma^*}\right) + \mathrm{Tr}\left(\Sigma_{i+1}^{-1}\Sigma^*\right) + (\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}^*)^\top \Sigma_{i+1}^{-1} (\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}^*) - d \right\} \\ &= \quad \log\left(\frac{\det \Sigma_i}{\det \Sigma_{i+1}}\right) + \mathrm{Tr}\left[\left(\Sigma_i^{-1} - \Sigma_{i+1}^{-1}\right)\Sigma^*\right] \\ &\quad + (\boldsymbol{\mu}_i - \boldsymbol{\mu}^*)^\top \Sigma_i^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}^*) (\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}^*)^\top \Sigma_{i+1}^{-1} (\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}^*) \ . \end{aligned} \tag{31}$$

Substituting (29) we get,

$$\begin{aligned} \Delta_i \quad &= \quad \log\left(\frac{\det \Sigma_i}{\det \Sigma_{i+1}}\right) + \mathrm{Tr}\left[\left(\Sigma_i^{-1} - \Sigma_i^{-1} - \zeta_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)\Sigma^*\right] \\ &\quad + (\boldsymbol{\mu}_i - \boldsymbol{\mu}^*)^\top \Sigma_i^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}^*) (\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}^*)^\top \left(\Sigma_i^{-1} + \zeta_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right) (\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}^*) \\ &= \quad \log\left(\frac{\det \Sigma_i}{\det \Sigma_{i+1}}\right) - \zeta_i \left(\boldsymbol{x}_i^\top \Sigma^* \boldsymbol{x}_i^\top\right) && (32) \\ &\quad + (\boldsymbol{\mu}_i - \boldsymbol{\mu}^*)^\top \Sigma_i^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}^*) - (\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}^*)^\top \Sigma_i^{-1} (\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}^*) && (33) \\ &\quad - \zeta_i \left((\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}^*) \cdot \boldsymbol{x}_i\right)^2 \ . && (34) \end{aligned}$$

We develop separately (32),(33),(34); starting with (32). We apply Lemma D.1 of [3], to obtain

$$\frac{\det \Sigma_{i+1}}{\det \Sigma_i} = \frac{\det \Sigma_i^{-1}}{\det \Sigma_{i+1}^{-1}} = 1 - \zeta_i \boldsymbol{x}_i^\top \Sigma_{i+1} \boldsymbol{x}_i \ .$$

Substituting in (32),

$$(32) \quad = \quad -\log\left(\frac{\det \Sigma_{i+1}}{\det \Sigma_i}\right) - \zeta_i \left(\boldsymbol{x}_i^\top \Sigma^* \boldsymbol{x}_i^\top\right) = -\log\left(1 - \zeta_i \boldsymbol{x}_i^\top \Sigma_{i+1} \boldsymbol{x}_i\right) - \zeta_i \left(\boldsymbol{x}_i^\top \Sigma^* \boldsymbol{x}_i^\top\right) \ .$$

From convexity, $-\log(1-x) \geq x$ and thus,

$$(32) \geq \zeta_i \left(\boldsymbol{x}_i^\top \Sigma_{i+1} \boldsymbol{x}_i\right) - \zeta_i \left(\boldsymbol{x}_i^\top \Sigma^* \boldsymbol{x}_i^\top\right) \geq 0 \ , \tag{35}$$

where the last inequality follows directly from the right set of conditions in (27).

Using Theorem 2.4.1 of [1] and Lemma 6 we develop (33) and obtain the following lower bound,

$$\begin{aligned}(33) \quad &= \quad (\boldsymbol{\mu}_i - \boldsymbol{\mu}^*)^\top \Sigma_i^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}^*) - (\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}^*)^\top \Sigma_i^{-1}(\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}^*) \\ &\geq \quad (\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}_i)^\top \Sigma_i^{-1}(\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}_i) \;.\end{aligned}$$

Substituting the value of (9) we get,

$$(33) \geq \alpha_i^2 \boldsymbol{x}_i \Sigma_i \Sigma_i^{-1} \Sigma_i \boldsymbol{x}_i = \alpha_i^2 \boldsymbol{x}_i \Sigma_i \boldsymbol{x}_i = \alpha_i^2 v_i \;. \tag{36}$$

Finally, we further develop (34)

$$(34) \quad = \quad -\zeta_i \left(\left(y_i(\boldsymbol{\mu}_{i+1} \cdot \boldsymbol{x}_i)\right)^2 - 2y_i(\boldsymbol{\mu}_{i+1} \cdot \boldsymbol{x}_i) y_i(\boldsymbol{\mu}^* \cdot \boldsymbol{x}_i) + \left(y_i(\boldsymbol{\mu}^* \cdot \boldsymbol{x}_i)\right)^2 \right) \;.$$

As noted above, in case of an update, the KKT conditions that the constraint (3) is equality after the update, that is

$$y_i(\boldsymbol{x}_i \cdot \boldsymbol{\mu}_{i+1}) = \phi \sqrt{\boldsymbol{x}_i^\top \Sigma_{i+1} \boldsymbol{x}_i} > 0 \;,$$

and from the left set of conditions in (27) we have,

$$y_i(\boldsymbol{\mu}^* \cdot \boldsymbol{x}_i) \geq y_i(\boldsymbol{\mu}_{i+1} \cdot \boldsymbol{x}_i) > 0 \;.$$

Combining the above three equations we get,

$$\begin{aligned}(34) \quad &\geq \quad -\zeta_i \left(\left(y_i(\boldsymbol{\mu}_{i+1} \cdot \boldsymbol{x}_i)\right)^2 - 2\left(y_i(\boldsymbol{\mu}_{i+1} \cdot \boldsymbol{x}_i)\right)^2 + \left(y_i(\boldsymbol{\mu}^* \cdot \boldsymbol{x}_i)\right)^2 \right) \\ &= \quad -\zeta_i \left(-\left(y_i(\boldsymbol{\mu}_{i+1} \cdot \boldsymbol{x}_i)\right)^2 + \left(y_i(\boldsymbol{\mu}^* \cdot \boldsymbol{x}_i)\right)^2 \right) \\ &\geq \quad -\zeta_i \left(\boldsymbol{\mu}^* \cdot \boldsymbol{x}_i\right)^2 \;.\end{aligned} \tag{37}$$

Substituting (35), (36) and (37) in (32), (33) and (34) we get a lower bound,

$$\Delta_i \geq 0 + \alpha_i^2 v_i - \zeta_i \left(\boldsymbol{\mu}^* \cdot \boldsymbol{x}_i\right)^2 = \alpha_i^2 v_i - \boldsymbol{\mu}^{*\top} \left(\zeta_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right) \boldsymbol{\mu}^* \;. \tag{38}$$

Combining (38) together with (30) and property 6 of Lemma 5 yields the desired bound. ■

Finishing The Proof

Given the assumptions of the theorems we have Lemma 7. By property 3 of Lemma 5, term i on the left-hand-side of (28) upper-bounds the $0-1$ loss of example i. We now develop the RHS of (28) by substituting $\boldsymbol{\mu}_1 = \mathbf{0}$,$\Sigma_1 = I$,

$$\begin{aligned}&2\mathrm{D}_{\mathrm{KL}} \left(\mathcal{N}\left(\boldsymbol{\mu}^*, \Sigma^*\right) \| \mathcal{N}\left(\boldsymbol{\mu}_1, \Sigma_1\right)\right) + \boldsymbol{\mu}^{*\top} \left(\sum_i \zeta_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right) \boldsymbol{\mu}^* \\ = \quad &\log\left(\frac{\det \Sigma_1}{\det \Sigma^*}\right) + \mathrm{Tr}\left(\Sigma_1^{-1}\Sigma^*\right) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}^*)^\top \Sigma_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}^*) - d \\ &+\boldsymbol{\mu}^{*\top} \left(\sum_i \zeta_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right) \boldsymbol{\mu}^* \\ = \quad &\log\left(\frac{\det I}{\det \Sigma^*}\right) + \mathrm{Tr}\left(I^{-1}\Sigma^*\right) + (\boldsymbol{\mu}^*)^\top \left(\Sigma_1^{-1} + \sum_i \zeta_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right) \boldsymbol{\mu}^* - d \\ = \quad &-\log\det\Sigma^* + \mathrm{Tr}\left(\Sigma^*\right) + \boldsymbol{\mu}^{*\top}\Sigma_{n+1}^{-1}\boldsymbol{\mu}^* - d \;,\end{aligned}$$

where the last equality follows (29) . ■