A. Derivations

A.1. Accuracy of the 0-1 attack

We note g_1 the binary random variable that indicates whether z_1 was classified correctly, and thus considered part of the training set by the 0-1 attack. The attack is accurate if $g_1 = 1$ on training images and $g_1 = 0$ on other images. This happens with probability

$$\begin{aligned} p_{\text{bayes}} &= \mathbb{P}(m_1 = g_1) \\ &= \mathbb{P}(g_1{=}1 \mid m_1{=}1)\mathbb{P}(m_1{=}1) + \mathbb{P}(g_1{=}0 \mid m_1{=}0)\mathbb{P}(m_1{=}0) \\ &= \lambda p_{\mathrm{train}} + (1-\lambda)(1-p_{\mathrm{test}}). \end{aligned} \tag{42}$$

A.2. Gaussian data

Estimation of average distribution. We assume without loss of generality that $\mu = 0$. θ is the mean of n Gaussian variables, centered on μ with covariance I. Thus, θ follows a Gaussian distribution, of variance $\frac{1}{n}I$.

$$\int_t e^{-\ell(z,t)} p(t) dt = \frac{1}{\sqrt{\det\left(\frac{2\pi}{n} I\right)}} \int_t e^{\frac{-\|z-t\|^2 - n\|t\|^2}{2}} dt \tag{43}$$

Denoting $\omega := \frac{z}{n+1}$, we have

$$n\|t\|^2 + \|z-t\|^2 = (n+1)\|t-\omega\|^2 + \frac{n}{n+1}\|z\|^2, \tag{44}$$

hence

$$\int_t e^{\frac{-\|z-t\|^2 - n\|t\|^2}{2}} dt = \sqrt{\det\left(\frac{2\pi}{n+1} I\right)} e^{-\frac{n\|z\|^2}{2(n+1)}}. \tag{45}$$

We have:

$$\log\left(\int_t e^{-\ell(z,t)} p(t) dt\right) = C - \frac{n}{2(n+1)}\|z\|^2 \tag{46}$$

A.3. Bound on variations of a sigmoid

We show that

$$\sigma(u) \leq \sigma(v) + |u-v|_+/4 \quad \forall u, v \in \mathbb{R}. \tag{47}$$

Since σ is increasing, the relation is obvious for $v > u$.

For $u > v$, we observe that

$$\sup_u |\sigma'(u)| = \sup_u \frac{e^{-u}}{(1+e^{-u})^2} = \frac{1}{4}. \tag{48}$$

Thus, σ is Lipschitz-continuous with constant $1/4$, which entails Equation (47).

A.4. Hessian approximations

We give here a rough justification of the approximation conducted in the MATT paragraph of Section 5.

Equation (37) writes:

$$\log\left(\frac{\mathbb{P}(\theta \mid m_1 = 1, z_1, \mathcal{T})}{\mathbb{P}(\theta \mid m_1 = 0, z_1, \mathcal{T})}\right) \tag{49}$$

$$\approx -(\theta - \theta_1^*)^T H (\theta - \theta_1^*) + (\theta - \theta_0^*)^T H (\theta - \theta_0^*). \tag{50}$$

This approximation holds up to the following quantity:

$$\delta = \underbrace{-\frac{1}{2}\log\left(\frac{\det(H_1)}{\det(H_0)}\right)}_{\delta_1} + \underbrace{(\theta_1^* - \theta_0^*)^T (H_1 - H_0)(\theta_1^* - \theta_0^*)}_{\delta_2}. \tag{51}$$

We reason qualitatively in orders of magnitude. $\theta_0^* - \theta_1^*$ has order of magnitude $1/n$, and $H_1 - H_0$ has order of magnitude 1, so δ_2 has order of magnitude $1/n^2$. As for δ_1, we observe that $H_0^{-1}(H_1 - H_0)$ has order of magnitude $1/n$ and therefore

$$\begin{aligned} \delta_1 &= -\frac{1}{2}\log\left(\frac{\det(H_1)}{\det(H_0)}\right) && (52) \\ &= -\frac{1}{2}\log\left(\det\left(I + H_0^{-1}(H_1 - H_0)\right)\right) && (53) \\ &\approx -\operatorname{Tr}(H_0^{-1}(H_1 - H_0)). && (54) \end{aligned}$$

Hence, δ_1 has order of magnitude $1/n$ as well. Since the main term in Equation (37) is in the order of $1/\sqrt{n}$, δ_1 and δ_2 can be safely neglected.