Notes on Detection - AWGN Case

Multimedia Data Security
(original author: Giulia Boato <giulia.boato@unitn.it>)
Matteo Franzil <matteo.franzil+github@gmail.com>

December 11, 2021

$$H_0 : (a_0) \quad f'_i = x_i = f_i + n_i \tag{1}$$
$$(b_0) \quad f'_i = x_i + \gamma v_i (v \neq w) \tag{2}$$
$$H_1 : \quad f'_i = x_i + \gamma w_i = f_i + \gamma w_i + n_i \tag{3}$$

We can make a_0, b_0 coincide if $v = null$. Our likelihood ratio is the following:

$$l(f') = \frac{p(f'|w)}{\int_{\mathbb{R}^n} p(f'|v)p(v)\,dv} \qquad (v \neq w) \tag{4}$$
$$= \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{\frac{(-f'_i - \mu_x - \gamma w_i)^2}{2\sigma_x^2}}}{\prod_{i=1}^{n} \int_{\mathbb{R}^n} \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{\frac{(-f'_i - \mu_x - \gamma w_i)^2}{2\sigma_x^2}} p(v_i)\,dv_i} \tag{5}$$

We can set γw_i to 0 since its so much smaller than $2\sigma_x^2$. We can just therefore take into consideration case a_0.

$$= \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{\frac{(-f'_i - \mu_x - \gamma w_i)^2}{2\sigma_x^2}}}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_x^2}} e^{\frac{(-f'_i - \mu_x)^2}{2\sigma_x^2}}} \tag{6}$$
$$= \frac{\prod_{i=1}^{n} e^{\frac{(-f'_i - \mu_x - \gamma w_i)^2}{2\sigma_x^2}}}{\prod_{i=1}^{n} e^{\frac{(-f'_i - \mu_x)^2}{2\sigma_x^2}}} \tag{7}$$

We can transform to the logarithmic domain, which allows us to convert products into sums.

$$= \sum \frac{1}{2\sigma_x^2} \left[(f'_i - \mu_x)^2 - (f'_i - \mu_x - \gamma w_i)^2 \right] \tag{8}$$
$$= \frac{1}{2\sigma_x^2} \left[\sum 2\gamma f'_i w_i - \sum 2\mu_x \gamma w_i - \sum \gamma^2 w_i^2 \right] \tag{9}$$

Finally, we can simplify some terms - indeed, the ones without features are not interesting and can be dropped out:

$$= \frac{1}{2\sigma_x^2}\left[\sum 2\gamma f_i' w_i\right] \tag{10}$$

$$\Leftrightarrow \rho = \frac{1}{n}\sum f_i' w_i = \frac{f'w}{n} \tag{11}$$

In order to decide whether a mark is present or not in a photo, the detector needs only to look at the correlation between the to-be-searched watermark and the host feature vector extractor from A', and compare it against a detection threshold T_p:

$$\int_T^{\infty} p(\rho|H_0)\, d\rho = \overline{\rho_f} \tag{12}$$

Here, we applied the Neyman-Pearson criterion in order to get the target value $\overline{P_f}$. Indeed, here ρ is a solid projection of f' over w.

When computing the false detection probability for setting the threshold at the detector, the watermark signal is known. Thus we have to average over all possible host assets. When evaluation the performance of the whole watermarking systems, we also have to average over all possible watermarks.

Watermark samples are zero-mean i.i.d. random variables. Our distribution of ρ, no matter if the features are watermarked or not - i.e., both for attack noise and host features - are both gaussian. The means and variances will change, though, and this is what we're going to estimate:

$$p(\rho|H_0)\tilde{(}\mu_{\rho|H_0}, \sigma^2_{\rho|H_0}) \tag{13}$$

$$p(\rho|H_1)\tilde{(}\mu_{\rho|H_1}, \sigma^2_{\rho|H_1}) \tag{14}$$

Our threshold is therefore set in the median point between the two gaussians. on the other hand, if the false alarm rate has to be fixed, so we can just $\int_T^{+\infty} p(\rho|H_0)\, dx = 10^{-2}$ (be sure that is H_0!!!)

Now let's assume our threshold is really what we wanted. Let's calculate the misdetection rate: it's just the other tail:

$$\int_{-\infty}^{T} p(\rho|H_1)\, dx = k, \qquad k = 10^{-2} \tag{15}$$

This is the one that Bayes would have set to $1 - P_d$.

Assume we're not in a bayesian case instead (so the threshold is not the mean). if we have a fixed alarm rate but the misdetection rate is too high, we can fix it by increasing the mean of the H_1 distribution, moving it to the right.

$$\mu_{\rho|H_0} = E[\rho|H_0] \xrightarrow{f_i'=x_i} \frac{1}{n}E[\sum x_i w_i] = \frac{1}{n}\sum E[x_i]w_i = \mu_x \sum \frac{w_i}{n} = 0 \tag{16}$$

The average of w is 0, so the mean is 0. this allows us to fix $\rho|H_0$ on 0. We can now change the variance:

$$\sigma^2_{\rho|H_0} = var(\frac{1}{n}\sum x_i w_i) = \frac{1}{n^2}\sigma_x^2 \sum w_i^2 = \frac{\sigma_x^2\sigma_w^2}{n} \tag{17}$$

Where the last step can be done if we assume that $n \gg$

$$\frac{1}{n}\sum w_i^2 = \frac{\|w\|^2}{n} \approx E[w^2] = \sigma_w^2 \tag{18}$$

Shifting to the right means:

$$\mu_{rho|H_1} = \gamma\sigma_w^2 \tag{19}$$

and therefore moving to a $\gamma' > \gamma$: the stronger the watermark, the less misdetection (but it will be more visible), or we can just select a different watermark with a different σ^2. On the other hand, to get the variance:

$$\sigma^2_{\rho|H_0} = \frac{1}{n}\sigma_x^2\sigma_w^2 \tag{20}$$

By also playing with the variance we can tamper the impact (longer watermark means less variance and thinner gaussian)

In the end, what we do need for the detector performance is:

$$\overline{P_f} = \frac{1}{2}\operatorname{erfc}\left(\sqrt{\frac{(T-\mu_{\rho|H_0})^2}{2\sigma^2_{\rho|H_0}}}\right) = \frac{1}{2}\operatorname{erfc}\left(\sqrt{\frac{T^2}{2\sigma^2_{\rho|H_0}}}\right) \tag{21}$$

$$1 - P_d = \frac{1}{2}\operatorname{erfc}\left(\sqrt{\frac{(\mu_{\rho|H_1}-T)^2}{2\sigma^2_{\rho|H_1}}}\right) \tag{22}$$

□