A PROOFS FOR l_2 NORM

In this section, we present the full proofs for Theorem. 4.1. The main tool for our proofs is the Neyman-Pearson lemma for two variables, which we establish in Appendix A.1. Based on this lemma, we obtain the certified result in Appendix A.2. Finally, we provide the details of the linear transformation used for certification in Appendix A.3.

A.1 Neyman-Pearson for Two Variables

Lemma A.1 (Neyman-Pearson for two variables). *Let X_1 and Y_1 be random variables in $\mathbb{R}^d$ with densities μ_{X_1} and μ_{Y_1}. Then, let X_2 and Y_2 be random variables in $\mathbb{R}^d$ with densities μ_{X_2} and μ_{Y_2}. Let $h : \mathbb{R}^d \times \mathbb{R}^d \to \{0, 1\}$ be any deterministic or random function with an input pair. Then:*
1. If $\mathcal{S}_1 \times \mathcal{S}_2 = \left\{z_1 \in \mathbb{R}^d, z_2 \in \mathbb{R}^d : \frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \leq t\right\}$ for some $t > 0$ and $P(h(X_1, X_2) = 1) \geq P((X_1, X_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$, then $P(h(Y_1, Y_2) = 1) \geq P((Y_1, Y_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$.
2. If $\mathcal{S}_1 \times \mathcal{S}_2 = \left\{z_1 \in \mathbb{R}^d, z_2 \in \mathbb{R}^d : \frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \geq t\right\}$ for some $t > 0$ and $P(h(X_1, X_2) = 1) \leq P((X_1, X_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$, then $P(h(Y_1, Y_2) = 1) \leq P((Y_1, Y_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$.

Proof. We denote the complement of $\mathcal{S}_1 \times \mathcal{S}_2$ as $\mathcal{S}^c$.

$$
\begin{aligned}
&P(h(Y_1, Y_2) = 1) - P((Y_1, Y_2) \in \mathcal{S}_1 \times \mathcal{S}_2) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} h(1 \mid z_1, z_2)\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)dz_1dz_2 - \int\int_{\mathcal{S}_1\times\mathcal{S}_2} \mu_{Y_1}(z_1)\mu_{Y_2}(z_2)dz_1dz_2 \\
&= \left[\int\int_{\mathcal{S}_1\times\mathcal{S}_2} h(1 \mid z_1, z_2)\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)dz_1dz_2 + \int\int_{\mathcal{S}^c} h(1 \mid z_1, z_2)\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)dz_1dz_2\right] \\
&- \left[\int\int_{\mathcal{S}_1\times\mathcal{S}_2} h(1 \mid z_1, z_2)\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)dz_1dz_2 + \int\int_{\mathcal{S}_1\times\mathcal{S}_2} h(0 \mid z_1, z_2)\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)dz_1dz_2\right] \\
&= \int\int_{\mathcal{S}^c} h(1 \mid z_1, z_2)\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)dz_1dz_2 - \int\int_{\mathcal{S}_1\times\mathcal{S}_2} h(0 \mid z_1, z_2)\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)dz_1dz_2 \\
&\geq t\left[\int\int_{\mathcal{S}^c} h(1 \mid z_1, z_2)\mu_{X_1}(z_1)\mu_{X_2}(z_2)dz_1dz_2 - \int\int_{\mathcal{S}_1\times\mathcal{S}_2} h(0 \mid z_1, z_2)\mu_{X_1}(z_1)\mu_{X_2}(z_2)dz_1dz_2\right] \\
&= t[\int\int_{\mathcal{S}^c} h(1 \mid z_1, z_2)\mu_{X_1}(z_1)\mu_{X_2}(z_2)dz_1dz_2 + \int\int_{\mathcal{S}_1\times\mathcal{S}_2} h(1 \mid z_1, z_2)\mu_{X_1}(z_1)\mu_{X_2}(z_2)dz_1dz_2 \\
&- \int\int_{\mathcal{S}_1\times\mathcal{S}_2} h(1 \mid z_1, z_2)\mu_{X_1}(z_1)\mu_{X_2}(z_2)dz_1dz_2 - \int\int_{\mathcal{S}_1\times\mathcal{S}_2} h(0 \mid z_1, z_2)\mu_{X_1}(z_1)\mu_{X_2}(z_2)dz_1dz_2] \\
&= t\left[\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} h(1 \mid z_1, z_2)\mu_{X_1}(z_1)\mu_{X_2}(z_2)dz_1dz_2 - \int\int_{\mathcal{S}_1\times\mathcal{S}_2} \mu_{X_1}(z_1)\mu_{X_2}(z_2)dz_1dz_2\right] \\
&= t[P(h(X_1, X_2) = 1) - P((X_1, X_2) \in \mathcal{S}_1 \times \mathcal{S}_2)] \\
&\geq 0
\end{aligned}
$$

□

Next, we prove Lemma A.2, which is a special case of Lemma A.1 and states the Neyman-Pearson lemma for two joint Gaussian noise variables.

Lemma A.2 (Neyman-Pearson for Two Joint Gaussian Noise). *Let $X_1 \sim \mathcal{N}(x_1, \Sigma_1)$, $X_2 \sim \mathcal{N}(x_2, \Sigma_2)$ and $Y_1 \sim \mathcal{N}(x_1 + \delta_1, \Sigma_1)$, $Y_2 \sim \mathcal{N}(x_2 + \delta_2, \Sigma_2)$. Let $h : \mathbb{R}^d \times \mathbb{R}^d \to \{0, 1\}$ be any deterministic or random function. Then:*
1. If $\mathcal{S}_1 \times \mathcal{S}_2 = \left\{z_1 \in \mathbb{R}^d, z_2 \in \mathbb{R}^d : \delta_1^\top \Sigma_1^{-1} z_1 + \delta_2^\top \Sigma_2^{-1} z_2 \leq \beta\right\}$ for some β and $P(h(X_1, X_2) = 1) \geq P((X_1, X_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$, then $P(h(Y_1, Y_2) = 1) \geq P((Y_1, Y_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$.
2. If $\mathcal{S}_1 \times \mathcal{S}_2 = \left\{z_1 \in \mathbb{R}^d, z_2 \in \mathbb{R}^d : \delta_1^\top \Sigma_1^{-1} z_1 + \delta_2^\top \Sigma_2^{-1} z_2 \geq \beta\right\}$ for some β and $P(h(X_1, X_2) = 1) \leq P((X_1, X_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$, then $P(h(Y_1, Y_2) = 1) \leq P((Y_1, Y_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$.

Proof. This lemma is the special case of Neyman-Pearson for two variables when X_1, X_2, Y_1, and Y_2 are joint Gaussian noises. It suffices to simply show that for any β, there is some $t > 0$ for which:

$$
\begin{aligned}
\left\{z_1, z_2 : \delta_1^\top \Sigma_1^{-1} z_1 + \delta_2^\top \Sigma_2^{-1} z_2 \leq \beta\right\} &= \left\{z_1, z_2 : \frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \leq t\right\}, \\
\left\{z_1, z_2 : \delta_1^\top \Sigma_1^{-1} z_1 + \delta_2^\top \Sigma_2^{-1} z_2 \geq \beta\right\} &= \left\{z_1, z_2 : \frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \geq t\right\}.
\end{aligned} \tag{18}
$$

For ease of representation, we use $M_1 \in \mathbb{R}^{d\times d}$ (with element $m_{1_{ij}}$) instead of Σ_1^{-1} and $M_2 \in \mathbb{R}^{d\times d}$ (with element $m_{2_{ij}}$) instead of Σ_2^{-1}. The likelihood ratio for this choice of X_1, X_2, Y_1 and Y_2 turns out to be:

$$
\begin{aligned}
&\frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)}\\
&= \frac{\exp\left(-\frac{1}{2}(z_1-(x_1+\delta_1))^\top\Sigma_1^{-1}(z_1-(x_1+\delta_1))\right)}{\exp\left(-\frac{1}{2}(z_1-x_1)^\top\Sigma_1^{-1}(z_1-x_1)\right)} \times \frac{\exp\left(-\frac{1}{2}(z_2-(x_2+\delta_2))^\top\Sigma_2^{-1}(z_2-(x_2+\delta_2))\right)}{\exp\left(-\frac{1}{2}(z_2-x_2)^\top\Sigma_2^{-1}(z_2-x_2)\right)}\\
&= \frac{\exp\left(-\frac{1}{2}\sum_i^d\sum_j^d\left(z_{1_i}-(x_{1_i}+\delta_{1_i})\right)m_{1_{ij}}\left(z_{1_j}-\left(x_{1_j}+\delta_{1_j}\right)\right)\right)}{\exp\left(-\frac{1}{2}\sum_i^d\sum_j^d\left(z_{1_i}-x_{1_i}\right)m_{1_{ij}}\left(z_{1_j}-x_{1_j}\right)\right)}\\
&\times \frac{\exp\left(-\frac{1}{2}\sum_i^d\sum_j^d\left(z_{2_i}-(x_{2_i}+\delta_{2_i})\right)m_{2_{ij}}\left(z_{2_j}-\left(x_{2_j}+\delta_{2_j}\right)\right)\right)}{\exp\left(-\frac{1}{2}\sum_i^d\sum_j^d\left(z_{2_i}-x_{2_i}\right)m_{2_{ij}}\left(z_{2_j}-x_{2_j}\right)\right)}\\
&= \exp\left(\delta_1^\top\Sigma_1^{-1}z_1-\delta_1^\top\Sigma_1^{-1}x_1-\frac{1}{2}\delta_1^\top\Sigma_1^{-1}\delta_1\right)\times\exp\left(\delta_2^\top\Sigma_2^{-1}z_2-\delta_2^\top\Sigma_2^{-1}x_2-\frac{1}{2}\delta_2^\top\Sigma_2^{-1}\delta_2\right)\\
&= \exp\left(\delta_1^\top\Sigma_1^{-1}z_1+\delta_2^\top\Sigma_2^{-1}z_2-\delta_1^\top\Sigma_1^{-1}x_1-\frac{1}{2}\delta_1^\top\Sigma_1^{-1}\delta_1-\delta_2^\top\Sigma_2^{-1}x_2-\frac{1}{2}\delta_2^\top\Sigma_2^{-1}\delta_2\right)\\
&= \exp\left(\delta_1^\top\Sigma_1^{-1}z_1+\delta_2^\top\Sigma_2^{-1}z_2+b\right)\le t,
\end{aligned}
$$

where b is a constant, specifically $b=-\delta_1^\top\Sigma_1^{-1}x_1-\frac{1}{2}\delta_1^\top\Sigma_1^{-1}\delta_1-\delta_2^\top\Sigma_2^{-1}x_2-\frac{1}{2}\delta_2^\top\Sigma_2^{-1}\delta_2$. Therefore given any β, we may take $t=\exp(\beta+b)$ and obtain this correlation:

$$
\begin{aligned}
\delta_1^\top\Sigma_1^{-1}z_1+\delta_2^\top\Sigma_2^{-1}z_2\le\beta &\Longleftrightarrow \exp\left(\delta_1^\top\Sigma_1^{-1}z_1+\delta_2^\top\Sigma_2^{-1}z_2+b\right)\le t,\\
\delta_1^\top\Sigma_1^{-1}z_1+\delta_2^\top\Sigma_2^{-1}z_2\ge\beta &\Longleftrightarrow \exp\left(\delta_1^\top\Sigma_1^{-1}z_1+\delta_2^\top\Sigma_2^{-1}z_2+b\right)\ge t.
\end{aligned}
\tag{19}
$$

□

A.2 Proof of the Certified Robustness

This subsection presents the logic for proving robustness guarantees and derives the certified spaces for these guarantees in Eq. 9.

To show that $g_0\left(\mathbf{c}^1,\mathbf{c}^2,\mathbf{z}^1+\delta_1,\mathbf{z}^2+\delta_2\right)=1$, it follows from the definition of g_0 that we need to show that:

$$P(f\left(\mathbf{c}^1,\mathbf{c}^2,\mathbf{z}^1+\varepsilon_1+\delta_1,\mathbf{z}^2+\varepsilon_2+\delta_2\right)\in\mathcal{X}')\ge P(f\left(\mathbf{c}^1,\mathbf{c}^2,\mathbf{z}^1+\varepsilon_1+\delta_1,\mathbf{z}^2+\varepsilon_2+\delta_2\right)\notin\mathcal{X}').$$

We define two random variables:

$$
\begin{aligned}
I &:= \left(\mathbf{c}^1,\mathbf{c}^2,\mathbf{z}^1+\varepsilon_1,\mathbf{z}^2+\varepsilon_2\right)=\left(\mathbf{c}^1,\mathbf{c}^2,\mathcal{N}\left(\mathbf{z}^1,\Sigma_1\right),\mathcal{N}\left(\mathbf{z}^2,\Sigma_2\right)\right)\\
O &:= \left(\mathbf{c}^1,\mathbf{c}^2,\mathbf{z}^1+\varepsilon_1+\delta_1,\mathbf{z}^2+\varepsilon_2+\delta_2\right)=\left(\mathbf{c}^1,\mathbf{c}^2,\mathcal{N}\left(\mathbf{z}^1+\delta_1,\Sigma_1\right),\mathcal{N}\left(\mathbf{z}^2+\delta_2,\Sigma_2\right)\right).
\end{aligned}
$$

We know that:

$$P(f(I)\in\mathcal{X}')\ge\underline{p}. \tag{20}$$

Our goal is to show that

$$P(f(O)\in\mathcal{X}')>P(f(O)\notin\mathcal{X}'). \tag{21}$$

According to lemma A.2, we can define the half-spaces:

$$
\begin{aligned}
\mathcal{A} &= \left\{z_1,z_2:\delta_1^\top\Sigma_1^{-1}(z_1-\mathbf{z}^1)+\delta_2^\top\Sigma_2^{-1}(z_2-\mathbf{z}^2)\le\|\delta_1^\top\Sigma_1^{-1}\mathbf{B}_1+\delta_2^\top\Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right\},\\
\mathcal{B} &= \left\{z_1,z_2:\delta_1^\top\Sigma_1^{-1}(z_1-\mathbf{z}^1)+\delta_2^\top\Sigma_2^{-1}(z_2-\mathbf{z}^2)\ge\|\delta_1^\top\Sigma_1^{-1}\mathbf{B}_1+\delta_2^\top\Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right\}.
\end{aligned}
$$

Claim 1 shows that $P(I\in\mathcal{A})=\underline{p}$, therefore we can obtain $P(f(I)\in\mathcal{X}')\ge P(I\in\mathcal{A})$. Hence we may apply Lemma A.2 to conclude:

$$P(f(O)\in\mathcal{X}')\ge P(O\in\mathcal{A}). \tag{22}$$

Similarly, we obtain $P(f(I)\notin\mathcal{X}')\le P(I\in\mathcal{B})$. Hence we may apply Lemma A.2 to conclude:

$$P(f(O)\notin\mathcal{X}')\le P(O\in\mathcal{B}). \tag{23}$$

Combining Eq. 22 and 23, we can obtain the conditions of Eq. 21:

$$P(f(O)\in\mathcal{X}')\ge P(O\in\mathcal{A})>P(O\in\mathcal{B})\ge P(f(O)\notin\mathcal{X}'). \tag{24}$$

According to Claim 3 and Claim 4, we can obtain $P(O \in \mathcal{A})$ and $P(O \in \mathcal{B})$ as:

$$\begin{aligned} P(O \in \mathcal{A}) &= \Phi\left(\Phi^{-1}\left(\underline{p}\right) - \frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|}\right), \\ P(O \in \mathcal{B}) &= \Phi\left(-\Phi^{-1}\left(\underline{p}\right) + \frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|}\right). \end{aligned} \tag{25}$$

Finally, we obtain that $P(O \in \mathcal{A}) > P(O \in \mathcal{B})$ if and only if:

$$\frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|} < \Phi^{-1}\left(\underline{p}\right).$$

A.3 Linear Transformation and Derivation

This subsection begins with Lemma A.3, which is the main tool for deriving all claims. Then, we present the proof process of claims, which is applied in Sec. A.2.

Lemma A.3 (Joint Gaussian Distribution). *If there is a random matrix $X \sim \mathcal{N}(\mu, \Sigma)$, where $\mu \in \mathbb{R}^n$ is the mean matrix. A positive semi-definite real symmetric matrix $\Sigma \in \mathbb{S}_{++}^{n\times n}$ is the covariance matrix of X. There is a full rank matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$, which makes $X = \mathbf{B}Z + \mu$, $Z \sim \mathcal{N}(\mathbf{0}, I)$ and $\mathbf{B}^\top\mathbf{B} = \Sigma$.*

We obtain four claims based on linear transformation:

Claim 1. $P(I \in \mathcal{A}) = \underline{p}$

Proof. Recall that $\mathcal{A} = \left\{z_1, z_2 : \delta_1^\top \Sigma_1^{-1}(z_1 - \mathbf{z}^1) + \delta_2^\top \Sigma_2^{-1}(z_2 - \mathbf{z}^2) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right\}$, according to lemma A.3, we can obtain:

$$\begin{aligned} P(I \in \mathcal{A}) &= P\left(\delta_1^\top \Sigma_1^{-1}(\mathcal{N}\left(\mathbf{z}^1, \Sigma_1\right) - \mathbf{z}^1) + \delta_2^\top \Sigma_2^{-1}(\mathcal{N}\left(\mathbf{z}^2, \Sigma_2\right) - \mathbf{z}^2) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\ &= P\left(\delta_1^\top \Sigma_1^{-1}\mathcal{N}\left(0, \Sigma_1\right) + \delta_2^\top \Sigma_2^{-1}\mathcal{N}\left(0, \Sigma_2\right) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\ &= P\left(\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1\mathcal{N}\left(0, I\right) + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\mathcal{N}\left(0, I\right) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\ &= P\left(\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\mathcal{N}\left(0, 1\right) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\ &= \Phi\left(\Phi^{-1}\left(\underline{p}\right)\right) \\ &= \underline{p}. \end{aligned}$$

□

Claim 2. $P(I \in \mathcal{B}) = 1 - \underline{p}$

Proof. Recall that $\mathcal{B} = \left\{z_1, z_2 : \delta_1^\top \Sigma_1^{-1}(z_1 - \mathbf{z}^1) + \delta_2^\top \Sigma_2^{-1}(z_2 - \mathbf{z}^2) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right\}$, according to lemma A.3, we can obtain:

$$\begin{aligned} P(I \in \mathcal{B}) &= P\left(\delta_1^\top \Sigma_1^{-1}(\mathcal{N}\left(\mathbf{z}^1, \Sigma_1\right) - \mathbf{z}^1) + \delta_2^\top \Sigma_2^{-1}(\mathcal{N}\left(\mathbf{z}^2, \Sigma_2\right) - \mathbf{z}^2) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\ &= P\left(\delta_1^\top \Sigma_1^{-1}\mathcal{N}\left(0, \Sigma_1\right) + \delta_2^\top \Sigma_2^{-1}\mathcal{N}\left(0, \Sigma_2\right) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\ &= P\left(\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1\mathcal{N}\left(0, I\right) + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\mathcal{N}\left(0, I\right) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\ &= P\left(\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\mathcal{N}\left(0, 1\right) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right) \\ &= 1 - \Phi\left(\Phi^{-1}\left(\underline{p}\right)\right) \\ &= 1 - \underline{p}. \end{aligned}$$

□

Claim 3. $P(O \in \mathcal{A}) = \Phi\left(\Phi^{-1}\left(\underline{p}\right) - \frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|}\right)$

Proof. Recall that $\mathcal{A} = \left\{z_1, z_2 : \delta_1^\top \Sigma_1^{-1}(z_1 - \mathbf{z}^1) + \delta_2^\top \Sigma_2^{-1}(z_2 - \mathbf{z}^2) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right\}$ and $O \sim \left(\mathbf{c}^1, \mathbf{c}^2, \mathcal{N}\left(\mathbf{z}^1 + \delta_1, \Sigma_1\right), \mathcal{N}\left(\mathbf{z}^2 + \delta_2, \Sigma_2\right)\right)$, according to lemma A.3, we can obtain:

$$
\begin{aligned}
&P(O \in \mathcal{A})\\
&= P\left(\delta_1^\top \Sigma_1^{-1}(\mathcal{N}\left(\mathbf{z}^1 + \delta_1, \Sigma_1\right) - \mathbf{z}^1) + \delta_2^\top \Sigma_2^{-1}(\mathcal{N}\left(\mathbf{z}^2 + \delta_2, \Sigma_2\right) - \mathbf{z}^2) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right)\\
&= P\left(\delta_1^\top \Sigma_1^{-1}\mathcal{N}\left(\delta_1, \Sigma_1\right) + \delta_2^\top \Sigma_2^{-1}\mathcal{N}\left(\delta_2, \Sigma_2\right) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right)\\
&= P\left(\delta_1^\top \Sigma_1^{-1}(\mathbf{B}_1\mathcal{N}\left(0, I\right) + \delta_1) + \delta_2^\top \Sigma_2^{-1}(\mathbf{B}_2\mathcal{N}\left(0, I\right) + \delta_2) \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right)\\
&= P\left(\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\mathcal{N}\left(0, 1\right) + \delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2 \le \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right)\\
&= P\left(\mathcal{N}\left(0, 1\right) \le \Phi^{-1}\left(\underline{p}\right) - \frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|}\right)\\
&= \Phi\left(\Phi^{-1}\left(\underline{p}\right) - \frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|}\right).
\end{aligned}
$$

□

Claim 4. $P(O \in \mathcal{B}) = \Phi\left(-\Phi^{-1}\left(\underline{p}\right) + \frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|}\right)$

Proof. Recall that $\mathcal{B} = \left\{z_1, z_2 : \delta_1^\top \Sigma_1^{-1}(z_1 - \mathbf{z}^1) + \delta_2^\top \Sigma_2^{-1}(z_2 - \mathbf{z}^2) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right\}$ and $O \sim \left(\mathbf{c}^1, \mathbf{c}^2, \mathcal{N}\left(\mathbf{z}^1 + \delta_1, \Sigma_1\right), \mathcal{N}\left(\mathbf{z}^2 + \delta_2, \Sigma_2\right)\right)$, according to lemma A.3, we can obtain:

$$
\begin{aligned}
&P(O \in \mathcal{B})\\
&= P\left(\delta_1^\top \Sigma_1^{-1}(\mathcal{N}\left(\mathbf{z}^1 + \delta_1, \Sigma_1\right) - \mathbf{z}^1) + \delta_2^\top \Sigma_2^{-1}(\mathcal{N}\left(\mathbf{z}^2 + \delta_2, \Sigma_2\right) - \mathbf{z}^2) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right)\\
&= P\left(\delta_1^\top \Sigma_1^{-1}\mathcal{N}\left(\delta_1, \Sigma_1\right) + \delta_2^\top \Sigma_2^{-1}\mathcal{N}\left(\delta_2, \Sigma_2\right) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right)\\
&= P\left(\delta_1^\top \Sigma_1^{-1}(\mathbf{B}_1\mathcal{N}\left(0, I\right) + \delta_1) + \delta_2^\top \Sigma_2^{-1}(\mathbf{B}_2\mathcal{N}\left(0, I\right) + \delta_2) \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right)\\
&= P\left(\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\mathcal{N}\left(0, 1\right) + \delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2 \ge \|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|\Phi^{-1}\left(\underline{p}\right)\right)\\
&= P\left(\mathcal{N}\left(0, 1\right) \ge \Phi^{-1}\left(\underline{p}\right) - \frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|}\right)\\
&= \Phi\left(-\Phi^{-1}\left(\underline{p}\right) + \frac{\delta_1^\top \Sigma_1^{-1}\delta_1 + \delta_2^\top \Sigma_2^{-1}\delta_2}{\|\delta_1^\top \Sigma_1^{-1}\mathbf{B}_1 + \delta_2^\top \Sigma_2^{-1}\mathbf{B}_2\|}\right).
\end{aligned}
$$

□

B PROOF FOR l_1 NORM

In this section, we present the full proofs for the robustness guarantee for l_1 norm. The main tool for our proofs is the Neyman-Pearson lemma for two variables, which we establish in Lemma A.1. Next, we prove Lemma B.1, which is a special case of Lemma A.1 and states the Neyman-Pearson lemma for two Laplace noise variables. Based on this lemma, we obtain the certified result in Appendix B.1.

Lemma B.1 (Neyman-Pearson for Two Laplace Noise). *Let $X_1 \sim x_1 + \mathcal{L}\left(\lambda_1\right)$, $X_2 \sim x_2 + \mathcal{L}\left(\lambda_2\right)$ and $Y_1 \sim x_1 + \mathcal{L}\left(\lambda_1\right) + \delta_1$, $Y_2 \sim x_2 + \mathcal{L}\left(\lambda_2\right) + \delta_2$. Let $h : \mathbb{R}^d \times \mathbb{R}^d \to \{0, 1\}$ be any deterministic or random function. Then:*

1. If $\mathcal{S}_1 \times \mathcal{S}_2 = \left\{z_1 \in \mathbb{R}^d, z_2 \in \mathbb{R}^d : \frac{1}{\lambda_1}(\|\mathbf{z_1} - \boldsymbol{\delta_1}\|_1 - \|\mathbf{z_1}\|_1) + \frac{1}{\lambda_2}(\|\mathbf{z_2} - \boldsymbol{\delta_2}\|_1 - \|\mathbf{z_2}\|_1)\right) \ge \beta\}$ for some β and $P(h(X_1, X_2) = 1) \ge P((X_1, X_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$, then $P(h(Y_1, Y_2) = 1) \ge P((Y_1, Y_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$.

2. If $\mathcal{S}_1 \times \mathcal{S}_2 = \left\{z_1 \in \mathbb{R}^d, z_2 \in \mathbb{R}^d : \frac{1}{\lambda_1}(\|\mathbf{z_1} - \boldsymbol{\delta_1}\|_1 - \|\mathbf{z_1}\|_1) + \frac{1}{\lambda_2}(\|\mathbf{z_2} - \boldsymbol{\delta_2}\|_1 - \|\mathbf{z_2}\|_1)\right) \le \beta\}$ for some β and $P(h(X_1, X_2) = 1) \le P((X_1, X_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$, then $P(h(Y_1, Y_2) = 1) \le P((Y_1, Y_2) \in \mathcal{S}_1 \times \mathcal{S}_2)$.

Proof. This lemma is the special case of Neyman-Pearson for two variables when X_1, X_2, Y_1, and Y_2 are Laplace noises. It suffices to simply show that for any β, there is some $t > 0$ for which:

$$\begin{aligned}\left\{z_1, z_2 : \frac{1}{\lambda_1}(\|\boldsymbol{z_1} - \boldsymbol{\delta_1}\|_1 - \|\boldsymbol{z_1}\|_1) + \frac{1}{\lambda_2}(\|\boldsymbol{z_2} - \boldsymbol{\delta_2}\|_1 - \|\boldsymbol{z_2}\|_1) \geq \beta\right\} = \left\{z_1, z_2 : \frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \leq t\right\}, \\ \left\{z_1, z_2 : \frac{1}{\lambda_1}(\|\boldsymbol{z_1} - \boldsymbol{\delta_1}\|_1 - \|\boldsymbol{z_1}\|_1) + \frac{1}{\lambda_2}(\|\boldsymbol{z_2} - \boldsymbol{\delta_2}\|_1 - \|\boldsymbol{z_2}\|_1) \leq \beta\right\} = \left\{z_1, z_2 : \frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \geq t\right\}.\end{aligned} \tag{26}$$

$$\begin{aligned}&\frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \\ &= \frac{\exp\left(-\frac{1}{\lambda_1}\|\boldsymbol{z_1} - \boldsymbol{\delta_1}\|_1\right)\exp\left(-\frac{1}{\lambda_2}\|\boldsymbol{z_2} - \boldsymbol{\delta_2}\|_1\right)}{\exp\left(-\frac{1}{\lambda_1}\|\boldsymbol{z_1}\|_1\right)\exp\left(-\frac{1}{\lambda_2}\|\boldsymbol{z_2}\|_1\right)} \\ &= \exp\left(-\frac{1}{\lambda_1}(\|\boldsymbol{z_1} - \boldsymbol{\delta_1}\|_1 - \|\boldsymbol{z_1}\|_1) - \frac{1}{\lambda_2}(\|\boldsymbol{z_2} - \boldsymbol{\delta_2}\|_1 - \|\boldsymbol{z_2}\|_1)\right)\end{aligned}$$

By choosing $\beta = -\log(t)$, we can derive that

$$\begin{aligned}\frac{1}{\lambda_1}(\|\boldsymbol{z_1} - \boldsymbol{\delta_1}\|_1 - \|\boldsymbol{z_1}\|_1) + \frac{1}{\lambda_2}(\|\boldsymbol{z_2} - \boldsymbol{\delta_2}\|_1 - \|\boldsymbol{z_2}\|_1) \geq \beta \iff \frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \leq t, \\ \frac{1}{\lambda_1}(\|\boldsymbol{z_1} - \boldsymbol{\delta_1}\|_1 - \|\boldsymbol{z_1}\|_1) + \frac{1}{\lambda_2}(\|\boldsymbol{z_2} - \boldsymbol{\delta_2}\|_1 - \|\boldsymbol{z_2}\|_1) \leq \beta \iff \frac{\mu_{Y_1}(z_1)\mu_{Y_2}(z_2)}{\mu_{X_1}(z_1)\mu_{X_2}(z_2)} \geq t.\end{aligned}$$

□

B.1 Proof of the Certified Robustness for l_1 norm

Theorem B.2 (ℓ_1 norm certified space for visual GM). *Let f be a matching function, f_0 and g_0 be defined as in Eq. 6 and Eq. 7, $\varepsilon_1 \sim \mathcal{L}(\lambda_1)$, $\varepsilon_2 \sim \mathcal{L}(\lambda_2)$. Suppose $\underline{p} \in (\frac{1}{2}, 1]$ satisfy:*

$$\begin{aligned}P(f_0\left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1, \mathbf{z}^2 + \varepsilon_2\right) = 1) = \\ P(f(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1, \mathbf{z}^2 + \varepsilon_2) \in \mathcal{X}') = p \geq \underline{p}.\end{aligned} \tag{27}$$

Then we obtain the ℓ_1 norm certified space for the perturbation pair (δ_1, δ_2):

$$\frac{\|\delta_1\|_1}{\lambda_1} + \frac{\|\delta_2\|_1}{\lambda_2} \leq -\log\left[2\left(1 - \underline{p}\right)\right], \tag{28}$$

which guarantees $g_0\left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \delta_1, \mathbf{z}^2 + \delta_2\right) = 1$.

To show that $g_0\left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \delta_1, \mathbf{z}^2 + \delta_2\right) = 1$, it follows from the definition of g_0 that we need to show that:

$$P(f\left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1 + \delta_1, \mathbf{z}^2 + \varepsilon_2 + \delta_2\right) \in \mathcal{X}') \geq P(f\left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1 + \delta_1, \mathbf{z}^2 + \varepsilon_2 + \delta_2\right) \notin \mathcal{X}').$$

We define two random variables:

$$\begin{aligned}I &:= \left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1, \mathbf{z}^2 + \varepsilon_2\right) \\ O &:= \left(\mathbf{c}^1, \mathbf{c}^2, \mathbf{z}^1 + \varepsilon_1 + \delta_1, \mathbf{z}^2 + \varepsilon_2 + \delta_2\right).\end{aligned}$$

We know that:

$$P(f(I) \in \mathcal{X}') \geq \underline{p}. \tag{29}$$

Our goal is to show that

$$P(f(O) \in \mathcal{X}') > P(f(O) \notin \mathcal{X}'). \tag{30}$$

Denote $T(\mathbf{z}^1, \mathbf{z}^2) = \frac{1}{\lambda_1}(\|\mathbf{z}^1 - \boldsymbol{\delta_1}\|_1 - \|\mathbf{z}^1\|_1) + \frac{1}{\lambda_2}(\|\mathbf{z}^2 - \boldsymbol{\delta_2}\|_1 - \|\mathbf{z}^2\|_1)$. Use Triangle Inequality we can derive a bound for $T(\mathbf{z}^1, \mathbf{z}^2)$:

$$-\frac{\|\delta_1\|_1}{\lambda_1} - \frac{\|\delta_2\|_1}{\lambda_2} \leq T(\mathbf{z}^1, \mathbf{z}^2) \leq \frac{\|\delta_1\|_1}{\lambda_1} + \frac{\|\delta_2\|_1}{\lambda_2}. \tag{31}$$

Pick β' such that there exists $B' \subseteq \{\boldsymbol{z_1}, \boldsymbol{z_2} : T(\boldsymbol{z_1}, \boldsymbol{z_2}) = \beta'\}$, and

$$P\left(I \in \left\{\boldsymbol{z_1}, \boldsymbol{z_2} : T(\boldsymbol{z_1}, \boldsymbol{z_2}) < \beta'\right\} \cup B'\right) = 1 - \underline{p} = P(f(I) \notin \mathcal{X}')). \tag{32}$$

Define

$$S := \left\{\boldsymbol{z_1}, \boldsymbol{z_2} : T(\boldsymbol{z_1}, \boldsymbol{z_2}) < \beta'\right\} \cup B', \tag{33}$$

so we also have $P(X \notin S) = p = P(f(I) \notin \mathcal{X}')$. Plug into Lemma B.1, we can get

$$\begin{aligned} P(Y \notin S) &\leq P(f(O) \in \mathcal{X}'), \\ P(Y \in S) &\geq P(f(O) \notin \mathcal{X}'). \end{aligned} \tag{34}$$

Then we can obtain

$$\begin{aligned} \mathbb{P}(Y \in S) &= \int \int_S [2\lambda_1]^{-d} [2\lambda_2]^{-d} \exp\left(-\frac{\|\mathbf{z}^1 - \boldsymbol{\delta}_1\|_1}{\lambda_1}\right) \exp\left(-\frac{\|\mathbf{z}^2 - \boldsymbol{\delta}_2\|_1}{\lambda_2}\right) \mathrm{d}\mathbf{z}^1 \mathrm{d}\mathbf{z}^2 \\ &= \int \int_S [2\lambda_1]^{-d} [2\lambda_2]^{-d} \exp\left(-\frac{\|\mathbf{z}^1\|_1}{\lambda_1}\right) \left(-\frac{\|\mathbf{z}^2\|_1}{\lambda_2}\right) \exp\left(-T(\mathbf{z}^1, \mathbf{z}^2)\right) \mathrm{d}\mathbf{z}^1 \mathrm{d}\mathbf{z}^2 \\ &\leq \exp\left(\frac{\|\delta_1\|_1}{\lambda_1} + \frac{\|\delta_2\|_1}{\lambda_2}\right) \int \int_S [2\lambda_1]^{-d} [2\lambda_2]^{-d} \exp\left(-\frac{\|\mathbf{z}^1\|_1}{\lambda_1}\right) \left(-\frac{\|\mathbf{z}^2\|_1}{\lambda_2}\right) \mathrm{d}\mathbf{z}^1 \mathrm{d}\mathbf{z}^2 \\ &= \exp\left(\frac{\|\delta_1\|_1}{\lambda_1} + \frac{\|\delta_2\|_1}{\lambda_2}\right) (1 - \underline{p}). \end{aligned} \tag{35}$$

Thus, if $\frac{\|\delta_1\|_1}{\lambda_1} + \frac{\|\delta_2\|_1}{\lambda_2} \leq -\log\left[2\left(1 - \underline{p}\right)\right]$, it holds that

$$\begin{aligned} P(Y \in S) &\leq \exp\left(\frac{\|\delta_1\|_1}{\lambda_1} + \frac{\|\delta_2\|_1}{\lambda_2}\right) (1 - \underline{p}) \\ &\exp\left(-\log\left[2\left(1 - \underline{p}\right)\right]\right) (1 - \underline{p}) \\ &= \frac{1}{2}. \end{aligned} \tag{36}$$