Towards Quantifying Incompatibilities in Evaluation Metrics for Feature Attributions: Appendix

Anonymous submission

Theoretical Proofs

Proposition 1 (Quadratic Representation)**.** *If $\mathcal{Q}$ is a generalized L^2 metric, then it has the quadratic representation*

$$\mathcal{Q}(\phi) = \phi^\top A\phi - 2b^\top \phi + c,$$

where $A = \mathbb{E}[\gamma_1^T\gamma_1] \in \mathbb{R}^{d\times d}$ is positive semi-definite ($A \succeq 0$), $b = \mathbb{E}[\gamma_1^T\gamma_2] \in \mathbb{R}^d$, and $c = \mathbb{E}[\|\gamma_2\|_2^2] \in \mathbb{R}$ is a constant.

Proof. Expanding the squared norm in Definition **??**:

$$\begin{aligned}\mathcal{Q}(\phi) &= \mathbb{E}\left[\|\gamma_1\phi - \gamma_2\|_2^2\right]\\ &= \mathbb{E}\left[(\gamma_1\phi-\gamma_2)^T(\gamma_1\phi-\gamma_2)\right]\\ &= \mathbb{E}\left[\phi^T\gamma_1^T\gamma_1\phi - 2\phi^T\gamma_1^T\gamma_2 + \gamma_2^T\gamma_2\right]\\ &= \phi^T\mathbb{E}[\gamma_1^T\gamma_1]\phi - 2\phi^T\mathbb{E}[\gamma_1^T\gamma_2] + \mathbb{E}[\gamma_2^T\gamma_2]\end{aligned}$$

Setting $A = \mathbb{E}[\gamma_1^T\gamma_1]$, $b = \mathbb{E}[\gamma_1^T\gamma_2]$, and $c = \mathbb{E}[\gamma_2^T\gamma_2]$ yields the result. Positive semi-definiteness of A follows from $A = \mathbb{E}[\gamma_1^T\gamma_1]$ being an expectation of Gram matrices. □

Proposition 2 (Optimal Metric Value)**.** *For any generalized L^2 metric $\mathcal{Q}$ with quadratic form $(\phi^TA\phi - 2b^T\phi + c)$, the minimal achievable value is:*

$$m(\mathcal{Q}) = \min_{\phi\in\mathbb{R}^d} \mathcal{Q}(\phi) = c - b^TA^\dagger b$$

where $A^\dagger$ denotes the Moore-Penrose pseudoinverse.

Proof. Taking the gradient with respect to ϕ and setting to zero: $2A\phi - 2b = 0$, yielding $\phi^* = A^\dagger b$ (using the pseudoinverse to handle potential rank deficiency). Substituting back:

$$\begin{aligned}m(\mathcal{Q}) &= (A^\dagger b)^TA(A^\dagger b) - 2b^TA^\dagger b + c\\ &= b^TA^\dagger AA^\dagger b - 2b^TA^\dagger b + c\\ &= b^TA^\dagger b - 2b^TA^\dagger b + c = c - b^TA^\dagger b\end{aligned}$$

where we used the property $AA^\dagger A = A$ for the pseudoinverse. □

Theorem 1 (Decomposition of Incompatibility)**.** *Let $\{(\lambda_k, v_k)\}_{k=1}^r$ be the generalized eigenpairs of (A_1, A_2) with A_2 invertible, where $r = rank(A_2)$. Define $\beta_{i,k} = v_k^Tb_i$ for $i \in \{1, 2\}$. Then the incompatibility index decomposes as:*

$$\mathcal{I}_{1,2}(\mathcal{Q}_1, \mathcal{Q}_2) = \sum_{k=1}^r \frac{(\beta_{1,k} - \lambda_k\beta_{2,k})^2}{\lambda_k(1+\lambda_k)}$$

Proof. By Proposition 2, we have:

$$\begin{aligned}m_1 + m_2 &= c_1 - b_1^TA_1^\dagger b_1 + c_2 - b_2^TA_2^\dagger b_2\\ m_{1,2} &= (c_1+c_2) - (b_1+b_2)^T(A_1+A_2)^\dagger(b_1+b_2)\end{aligned}$$

Thus:

$$\mathcal{I}_{1,2} = (b_1+b_2)^T(A_1+A_2)^\dagger(b_1+b_2) - b_1^TA_1^\dagger b_1 - b_2^TA_2^\dagger b_2$$

Assuming A_2 is invertible, the generalized eigenvectors $\{v_k\}$ form a basis. We can express $b_i = \sum_k \beta_{i,k}v_k$. Using the property that $A_1v_k = \lambda_kA_2v_k$ and orthogonality conditions, we obtain:

$$\begin{aligned}b_1^TA_1^\dagger b_1 &= \sum_k \frac{\beta_{1,k}^2}{\lambda_k}v_k^TA_2v_k\\ b_2^TA_2^\dagger b_2 &= \sum_k \beta_{2,k}^2 v_k^TA_2v_k\end{aligned}$$

As well as

$$(b_1+b_2)^T(A_1+A_2)^\dagger(b_1+b_2) = \sum_k \frac{(\beta_{1,k}+\beta_{2,k})^2}{1+\lambda_k}v_k^TA_2v_k$$

Substituting and simplifying yields the desired decomposition. □