

Detailed proof for deficiency reformulation

Loïc Fosse^{1,2},
Philippe Formont^{3,6,7,8},
Eric Aubinais^{3,5},
Pablo Piantanida^{3,4,6,7},

¹Orange Research, Lannion, France

²CNRS, LIS, Aix Marseille Université, France

³CNRS, Université Paris-Saclay

⁴CNRS, CentraleSupélec

⁵Laboratoire de mathématiques d'Orsay (LMO), France

⁶International Laboratory on Learning Systems (ILLS), Canada

⁷Mila - Quebec AI Institute, Canada

⁸ÉTS Montreal, Canada

In the document, we provide a detailed proof about deficiency reformulation. As stated in the main paper, the proof is a direct consequence of the following lemma.

Lemma 1. *Let T be a Bernoulli of parameter $\frac{1}{2}$. Let U, V, W be a random variables on the same space, such that $\mathbb{P}_{W|T=0} = \mathbb{P}_U$ and $\mathbb{P}_{W|T=1} = \mathbb{P}_V$, and there exists a measure λ such that $\mathbb{P}_U \ll \lambda$ and $\mathbb{P}_V \ll \lambda$, then,*

$$\begin{aligned} \|\mathbb{P}_V - \mathbb{P}_U\|_{TV} &= 1 - 2 \min_{\psi: \mathbf{W} \rightarrow \{0,1\}} \Pr(\psi(W) \neq T). \\ &= -1 + 2 \max_{\psi: \mathbf{W} \rightarrow \{0,1\}} \mathbb{E}_{\mathbb{P}_{T,W}} [\mathbb{1}(\psi(W) = T)] \\ &= -1 + \max_{\psi: \mathbf{W} \rightarrow \{0,1\}} (\mathbb{E} [\mathbb{1}(\psi(U) = 0)] + \mathbb{E} [\mathbb{1}(\psi(V) = 1)]) \end{aligned} \tag{1}$$

Proof. Some elements of the proof, can be found in [1, Theorem 2.2]. The main element of the proof being the identity

$$\begin{aligned} \min_{\psi: \mathbf{W} \rightarrow \{0,1\}} \Pr(\psi(W) \neq T) &= \frac{1}{2} \int \min \left(\frac{d\mathbb{P}_U}{d\lambda}, \frac{d\mathbb{P}_V}{d\lambda} \right) d\lambda \\ &= \frac{1}{2} - \frac{1}{4} \int \left| \frac{d\mathbb{P}_U}{d\lambda} - \frac{d\mathbb{P}_V}{d\lambda} \right| d\lambda = \frac{1 - \|\mathbb{P}_U - \mathbb{P}_V\|_{TV}}{2}, \end{aligned}$$

which can mainly be obtained by Scheffe's theorem [2], and the definition of the Bayesian error rate. \square

Additionally to Lemma 1, we introduce two auxiliary random variables: T a Bernoulli of parameter $\frac{1}{2}$ and $W \in \mathcal{W}$ such that for every $x \in \mathcal{X}$,

$$\begin{aligned} (W|T=0; X=x) &\sim \mathbb{P}_{U|X=x}, \\ (W|T=1; X=x) &\sim \mathbb{P}_{V|X=x}. \end{aligned} \quad (2)$$

Theorem 1. *Let $\mathcal{F} \triangleq \{f : \mathcal{W} \rightarrow \{0, 1\}\}$ be the set of binary functions that take as input values from \mathcal{W} and $f \in \mathcal{F}$. For any $K \in \mathcal{M}(\mathcal{V}|\mathcal{U})$, $x \in \mathcal{X}$, let $L(f, K, x)$ be defined as:*

$$\mathbb{E}_{\mathbb{P}_{T,W}} [\mathbb{E}_{Z \sim C_K(\cdot|W,T)} [\mathbb{1}(f(Z) = T)] \mid X = x], \quad (3)$$

where,

$$C_K(\cdot \mid W, T) = \begin{cases} K(\cdot \mid W) & \text{if } T = 0 \\ W & \text{if } T = 1 \end{cases} \quad (4)$$

Then we have,

$$\begin{aligned} \delta(\mathbb{P}_{U|X} \rightarrow \mathbb{P}_{V|X}) &= L^* \\ &\triangleq -1 + 2 \min_{K \in \mathcal{M}(\mathcal{V}|\mathcal{U})} \left(\max_{x \in \mathcal{X}} \left(\max_{f \in \mathcal{F}} L(f, K, x) \right) \right). \end{aligned} \quad (5)$$

Proof. One will agree that the proof of Theorem 1 relies only on the following,

$$\|K \circ \mathbb{P}_{U|X}(\cdot|x) - \mathbb{P}_{V|X}(\cdot|x)\|_{\text{TV}} = -1 + 2 \max_f L(f, K, x).$$

We will then only focus on the demonstration of this result. First, from Fubini-Tonelli and Markov composition operation definition, we have,

$$\begin{aligned} \mathbb{E}_{Z \sim K \circ \mathbb{P}_{U|X=x}} [\mathbb{1}(f(Z) = 0)] &= \int_z \mathbb{1}(f(z) = 0) K \circ \mathbb{P}_{U|X=x}(du) \\ &\stackrel{\text{F.T.}}{=} \int_u \left(\int_z \mathbb{1}(f(z) = 0) K(dz|u) \right) \mathbb{P}_{U|X=x}(du) \\ &= \mathbb{E}_{\mathbb{P}_U} [\mathbb{E}_{Z \sim K(\cdot|U)} [\mathbb{1}(f(Z) = 0)] \mid X = x]. \end{aligned} \quad (6)$$

Equivalently, we have,

$$\begin{aligned} \mathbb{E}_{Z \sim \mathbb{P}_{V|X=x}} [\mathbb{1}(f(Z) = 1)] &= \int_v \mathbb{1}(f(v) = 1) \mathbb{P}_{V|X=x}(dv) \\ &= \int_v \left(\int_z \mathbb{1}(f(z) = 1) S(dz|v) \right) \mathbb{P}_{V|X=x}(dv) \\ &= \mathbb{E}_{\mathbb{P}_V} [\mathbb{E}_{Z \sim S(\cdot|V)} [\mathbb{1}(f(Z) = 1)] \mid X = x], \end{aligned} \quad (7)$$

where $S(\cdot|\cdot)$ is a degenerated kernel, i.e. $S(\cdot|v)$ is a Dirac mass on v . Then, by taking W and T defined as in Eq. 2, and by defining $K(\cdot|W, T)$ as,

$$C_K(\cdot|W, T) = \begin{cases} K(\cdot|W) & \text{if } T = 0 \\ S(\cdot|W) = W & \text{if } T = 1 \end{cases} \quad (\text{degenerated kernel})$$

then by Eq. 1, we have that,

$$\|K \circ \mathbb{P}_{U|X}(\cdot|x) - \mathbb{P}_{V|X}(\cdot|x)\|_{\text{TV}} = -1 + \max_f \mathbb{E}_{Z \sim K \circ \mathbb{P}_{U|X=x}} [\mathbb{1}(f(Z) = 0)] + \mathbb{E}_{Z \sim \mathbb{P}_{V|X=x}} [\mathbb{1}(f(Z) = 1)]$$

Then from Eq. 6 and Eq. 7, we have,

$$\begin{aligned} \|K \circ \mathbb{P}_{U|X}(\cdot|x) - \mathbb{P}_{V|X}(\cdot|x)\|_{\text{TV}} &= -1 + 2 \max_f \mathbb{E}_{\mathbb{P}_{W,T}} [\mathbb{E}_{Z \sim C_K(\cdot|W,T)} [\mathbb{1}(f(Z) = T)] | X = x] \\ &= -1 + 2 \max_f L(f, K, x), \end{aligned}$$

which concludes the proof. \square

References

- [1] Alexandre B. Tsybakov, *Introduction to Nonparametric Estimation*, Springer New York, NY, 2009.
- [2] Henry Scheffé, “A Useful Convergence Theorem for Probability Distributions,” *The Annals of Mathematical Statistics*, 1947.