Proceedings of Machine Learning Research xx:1–1, 2020 ACML 2020

Dual Learning: Theoretical Study and an Algorithmic Extension (Supplementary Document)

Appendix A. Derivation

We leverage the negative logarithmic probability to measure the differences between the original $x^{(2)}$ and the reconstructed one. Let $\mathcal{R}(x^{(2)})$ denote the event that after passing the loop $S_2 \to S_k \to S_1 \to S_2$, $x^{(2)}$ is reconstructed to $x^{(2)}$. We have that

$$\begin{aligned}
&\ln \Pr(\mathcal{R}(x^{(2)})) = \sum_{x^{(k)}\in S_k}\sum_{x^{(1)}\in S_1} \ln \Pr(x^{(2)}, x^{(1)}, x^{(k)} | \\
&\text{starting from } x^{(2)}, \text{applied by } \theta_{2k}, \theta_{k1}, \theta_{12} \text{ sequentially}) \\
=& \sum_{x^{(k)}\in S_k}\sum_{x^{(1)}\in S_1} \ln \Big(\Pr(x^{(1)}, x^{(k)} | x^{(2)}; \theta_{2k}, \theta_{k1}) \cdot \Pr(x^{(2)} | x^{(1)}; \theta_{12}) \Big) \quad (1)\\
\geq& \sum_{x^{(k)}\in S_k}\sum_{x^{(1)}\in S_1} \Pr(x^{(1)}, x^{(k)} | x^{(2)}; \theta_{2k}, \theta_{k1}) \cdot \ln \Pr(x^{(2)} | x^{(1)}; \theta_{12}) \\
=& \sum_{x^{(k)}\in S_k}\sum_{x^{(1)}\in S_1} \Pr(x^{(k)} | x^{(2)}; \theta_{2k}, \theta_{k1}) \Pr(x^{(1)} | x^{(k)}, x^{(2)}; \theta_{2k}, \theta_{k1}) \cdot \ln \Pr(x^{(2)} | x^{(1)}; \theta_{12}) \\
=& \sum_{x^{(k)}\in S_k}\sum_{x^{(1)}\in S_1} \Pr(x^{(k)} | x^{(2)}; \theta_{2k}) \Pr(x^{(1)} | x^{(k)}; \theta_{k1}) \cdot \ln \Pr(x^{(2)} | x^{(1)}; \theta_{12}) \quad (2)\\
=& \mathbb{E}_{x^{(k)}\sim \Pr(\cdot | x^{(2)}; \theta_{2k})} \mathbb{E}_{x^{(1)}\sim \Pr(\cdot | x^{(k)}; \theta_{k1})} \ln \Pr(x^{(2)} | x^{(1)}; \theta_{12}). \quad (3)
\end{aligned}$$

In Eqn.(1), the first $\Pr$ represents the jointly probability that $x^{(2)}$ can be translated into $x^{(k)}$ with θ_{2k}, and the the obtained $x^{(k)}$ can be translated into $x^{(1)}$ with θ_{k1}; the second $\Pr$ represents the probability that given $x^{(1)}$, it can be translated back to $x^{(2)}$ with θ_{12}.