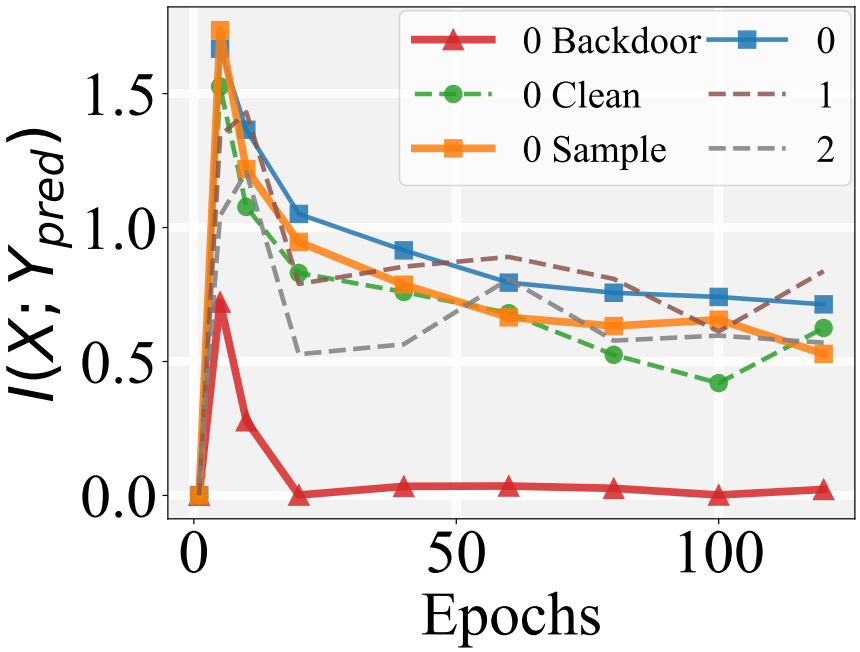

Figure R 5: Dynamics of MI $I(X;Y_{pred})$ of **different classes** under the Blend attack with the **same settings as in Figure 5** of the original paper(on the CIFAR-10 dataset using a ResNet-18 model with a 10% poisoning ratio and $\gamma = 0.4$). This inequality $I(X;Y_{pred}) \leq I(X;T)$ holds **true** at different stages of training.

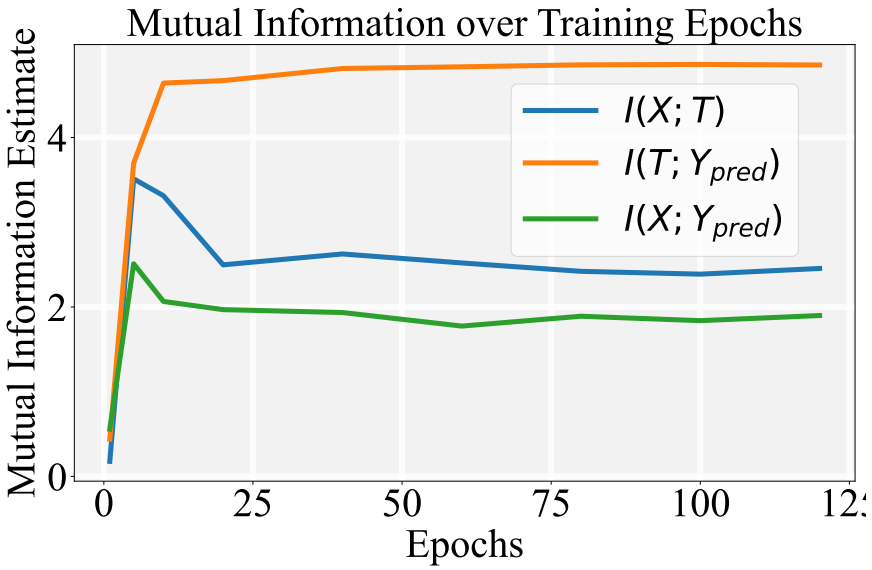

Figure R 6: Dynamics of MI under the Blend attack with the same settings as in Figure 5 of the original paper(on the CIFAR-10 dataset using a ResNet-18 model with a 10% poisoning ratio and $\gamma = 0.4$). Importantly, the X, T, Y_{pred} in this figure represent the **entire dataset** instead of a specific class in Fig.R5. This inequality $I(X;Y_{pred}) \leq I(X;T)$ holds **true** at different stages of training and $I(X;Y_{pred}) > 1.9\,nats$.