

After rearranging the terms, we get

$$\frac{1-\epsilon}{2} \geq \exp^{-n\delta^2}. \quad (33)$$

Taking log on both sides and rearranging terms yields

$$n \geq \frac{1}{\delta^2} \log \left(\frac{2}{1-\epsilon} \right). \quad (34)$$

Hence proved.

B.3 Proof of Theorem 2

Before starting the analysis, let us restate the following bound from (Dhurandhar, 2013) for quick reference.

Lemma 2 (Upper Bound for Non-iid scenario). *Let n be the number of samples drawn sequentially from $\mathbb{P}(S_1, S_2 \cdots S_n) = \prod_{j=1}^L \tau_j$, where τ_j are independent subsets consisting of c_j dependent sequences $(s_1, s_2 \cdots s_{c_j})$ such that $\sum_{j=1}^L c_j = n$. Under dependence structure in (16), for any $\delta > \frac{\sum_{j=1}^L (c_j - 1)\rho_j}{n}$, it holds that*

$$\mathbb{P}(|\bar{S} - \mathbb{E}[\bar{S}]| \geq \delta) \leq 2 \exp \frac{-2(n\delta - \sum_{j=1}^L (c_j - 1)\rho_j)^2}{n}, \quad (35)$$

where $\bar{S} = \frac{1}{n} \sum_{i=1}^n s_i$ and $\mathbb{E}[S_i | S_{i-1} = s_{i-1}, \dots, S_1 = s_1] = \frac{\rho}{i-1} \sum_{k=1}^{i-1} s_k + (1-\rho)\mathbb{E}[S_i]$.

Lemma B.3 provided upper bounds for non-iid scenarios, with an exponential bound in sample size n along with an additional dependence on the strength of association ρ_j and the size of the dependent sequence c_j . It is important to note that when we have $\rho = 0$, it exactly boils down to the standard Chernoff bound.

Now, we move to do the sample complexity analysis for the non-iid setting. Similar to the proof for the iid case, we define $\mathbb{P}(s^h \in A) = p$ which implies that $\mathbb{P}(s^m \in A) = p + \delta$. Let us now collect n samples sequentially $\{s_i\}_{i=1}^n$ from $m(s)$, we know that the probability of any sample s_i in A is given by $p + \delta$. Hence, on average $(p + \delta)n$ number of samples will be in A . In a similar manner, if we have n samples from $h(s)$, pn will be in A on average. Therefore, we can utilize the Chernoff bound to write

$$\begin{aligned} \mathbb{P}\left(\text{at least } \left(p + \frac{\delta}{2}\right)n \text{ samples of } h \text{ are in } A\right) &\leq 2e^{-\frac{-2\left(n\frac{\delta}{2} - \sum_{j=1}^L (c_j - 1)\rho_j\right)^2}{n}} \\ \mathbb{P}\left(\text{at most } \left(p + \frac{\delta}{2}\right)n \text{ samples of } m \text{ are in } A\right) &\leq 2e^{-\frac{-2\left(n\frac{\delta}{2} - \sum_{j=1}^L (c_j - 1)\rho_j\right)^2}{n}}, \end{aligned} \quad (36)$$

where for simplicity of notations let's consider $\beta = -\frac{2\left(n\frac{\delta}{2} - \sum_{j=1}^L (c_j - 1)\rho_j\right)^2}{n}$. Now, let us denote the set of n tuples by A' which contains more than $(p + \frac{\delta}{2})n$ samples of A . Therefore, we can bound

$$\begin{aligned} \text{TV}(m^{\otimes n}, h^{\otimes n}) &\geq \mathbb{P}(\{s_i^m\}_{i=1}^n \in A') - \mathbb{P}(\{s_i^h\}_{i=1}^n \in A') \\ &= 1 - 4 \exp^\beta. \end{aligned} \quad (37)$$

The TV norm lower bound in (37) tells us the minimum value of $\text{TV}(m^{\otimes n}, h^{\otimes n})$ for given n and δ . Therefore, to obtain the AUROC of the best possible detector to be equal to, or higher than say $\epsilon \in [0.5, 1]$, it should hold that

$$\frac{1}{2} + \text{TV}(m^{\otimes n}, h^{\otimes n}) - \frac{\text{TV}(m^{\otimes n}, h^{\otimes n})^2}{2} \geq \epsilon. \quad (38)$$

Since the left-hand side in (38) is the monotonically increasing function of $\text{TV}(m^{\otimes n}, h^{\otimes n})$, it holds from the minimum value in (29) that

$$\frac{1}{2} + (1 - 4 \exp^\beta) - \frac{(1 - 4 \exp^\beta)^2}{2} \geq \epsilon. \quad (39)$$

After expanding the squares, we get

$$\frac{1}{2} + 1 - 4 \exp^\beta - \frac{1}{2} - 8 \exp^{2\beta} + 4 \exp^\beta \geq \epsilon, \quad (40)$$

which implies

$$\frac{1 - \epsilon}{8} \geq \exp^{2\beta} = \exp^{-\frac{4(n\frac{\delta}{2} - \sum_{j=1}^L (c_j - 1)\rho_j)}{n}}, \quad (41)$$

where substitute the value of β and taking logarithm on both sides, we get

$$\begin{aligned} \log\left(\frac{8}{1 - \epsilon}\right) &\leq \frac{4}{n} \left(n\frac{\delta}{2} - \sum_{j=1}^L (c_j - 1)\rho_j \right)^2 \\ &= \frac{4}{n} \left(n\frac{\delta}{2} - \sum_{j=1}^L (c_j - 1)\rho_j \right)^2 \\ &= n\delta^2 - 4\delta \left(\sum_{j=1}^L (c_j - 1)\rho_j \right) + \frac{4}{n} \left(\sum_{j=1}^L (c_j - 1)\rho_j \right)^2. \end{aligned} \quad (42)$$

Let's denote $\alpha = \sum_{j=1}^L (c_j - 1)\rho_j$ and $\gamma(\epsilon) = \log\left(\frac{8}{1 - \epsilon}\right)$, for simplicity of calculations. The quadratic inequality from the above equation boils down to solving

$$\delta^2 n^2 - n(4\alpha\delta + \gamma(\epsilon)) + 4\alpha^2 \geq 0, \quad (43)$$

which is in the form of a standard quadratic equation and the corresponding solution is given by

$$\begin{aligned} n &\geq \frac{\gamma(\epsilon)}{2\delta^2} + 2\frac{\alpha}{\delta} + \frac{1}{2\delta^2} \sqrt{(4\alpha\delta + \gamma(\epsilon))^2 - 16\alpha^2\delta^2} \\ &= \frac{\gamma(\epsilon)}{2\delta^2} + 2\frac{\alpha}{\delta} + \frac{1}{2\delta^2} \sqrt{\gamma(\epsilon)^2 + 8\alpha\delta\gamma(\epsilon)} \\ &= \frac{\gamma(\epsilon)}{2\delta^2} + \frac{2}{\delta} \sum_{j=1}^L (c_j - 1)\rho_j + \frac{1}{2\delta^2} \sqrt{(\gamma(\epsilon))^2 + 8 \left(\sum_{j=1}^L (c_j - 1)\rho_j \right) \delta\gamma(\epsilon)}. \end{aligned} \quad (44)$$

Now, we further expand upon the expression as

$$\begin{aligned} n &\geq \frac{1}{2\delta^2}\gamma(\epsilon) + \frac{2}{\delta}\sum_{j=1}^L(c_j - 1)\rho_j + \frac{1}{\sqrt{2\delta^2}}\sqrt{\frac{1}{2}\left((\gamma(\epsilon))^2 + 8\left(\sum_{j=1}^L(c_j - 1)\rho_j\right)\delta\gamma(\epsilon)\right)} \\ &\geq \frac{1}{2\delta^2}\gamma(\epsilon) + \frac{2}{\delta}\sum_{j=1}^L(c_j - 1)\rho_j + \frac{1}{2\sqrt{2\delta^2}}\gamma(\epsilon) + \frac{1}{\sqrt{2\delta^2}}\sqrt{2\left(\sum_{j=1}^L(c_j - 1)\rho_j\right)\delta\gamma(\epsilon)}, \end{aligned} \quad (45)$$

where, the first-term results from multiplying and dividing by a constant factor 2, and the second term is an application of Jensen's inequality for convex functions. Using the order notation, we obtain

$$n = \Omega\left(\frac{1}{\delta^2}\log\left(\frac{1}{1-\epsilon}\right) + \frac{1}{\delta}\sum_{j=1}^L(c_j - 1)\rho_j + \sqrt{\frac{1}{\delta^3}\log\left(\frac{1}{1-\epsilon}\right)\left(\sum_{j=1}^L(c_j - 1)\rho_j\right)}\right). \quad (46)$$

C Additional Figures of Experimental Results

C.1 Additional Experimental Details

IMDb Dataset Experiments. To validate our claims on the possibilities of detection, we run experiments on the IMDb dataset (Maas et al., 2011), which is a widely-used benchmark dataset in the field of natural language processing. The dataset consists of 50,000 movie reviews from the internet movie database that have been labeled as positive or negative based on their sentiment. The goal is to classify the reviews accordingly based on their text content. The experiments are done to validate our hypothesis on a more general class of language tasks including classification and detection. We specifically focus on the representation space of the inputs for both the human and machine distributions and try to validate our hypothesis by comparing the input space of words to the input space of a group of sentences. The objective is to analyze the variations in performance of the detector when detecting at word-level versus paragraphs. Hence, there are two scenarios to consider. The first is where we're given a word and we have to determine whether it came from positive or negative class. The second, and more practical case, is where we're given a paragraph i.e a group of sentences and we have to detect whether it came from positive or negative class.

So, we first compute the total variation distance between the positive and negative classes at the word level. This is done by computing the divergence between the distribution over the space of words between the two classes. Figure 7(a) shows that the best possible AUROC achieved by the detector is 0.585 at the word level. From these results, it seems almost impossible to distinguish the two classes. However when we perform the detection at a paragraph level using a real detector (standard ML models, including random forest, logistic regression, and a vanilla multi-layer perception), we see a remarkable improvement in the detection performance. As shown in Figure 7(a), all the real detectors achieve a train AUROC of greater than 0.85 (≥ 0.93 for random forest and MLP), and a test AUROC of greater than 0.8, which surpasses the upper-bounds of the best detector at a word level, validating our theory and intuition. This impressive performance is fully aligned with our claims and provides evidence that designing a detector with high performance for AI-generated text is always possible even for general NLP detection tasks.