Importantly, it is crucial to maintain a gap between the bias in human-generated text and that in machine-generated text. This ensures that biased language remains more likely to originate from human-generated text than from LLMs. By doing so, effective detection and separation of the two sources can be achieved, enabling us to fully harness the potential of LLMs without compromising their integrity. With careful consideration and responsible use, LLMs can make a positive impact on our society, helping us to communicate more effectively and promoting fairness and inclusivity in language use.

B Detailed Proofs

B.1 Revisiting Le Cam's Lemma and the Existence of the Optimal Detector

We first restate Le Cam's lemma and its proof, which appears in Le Cam (2012) and many lecture notes such as (Wasserman, 2013).

Lemma 1 (Le Cam's Lemma)**.** *Let $\mathcal{S}$ be an arbitrary set. For any two distributions m and h on $\mathcal{S}$, we have*

$$\inf_{\Psi} \left\{ \mathbb{P}_{s\sim m}[\Psi(s) \neq 1] + \mathbb{P}_{s\sim h}[\Psi(s) \neq 0] \right\} = 1 - \mathsf{TV}(m, h), \tag{18}$$

where the infimum is taken over all detectors (measurable maps) $\Psi : \mathcal{S} \to \{1, 0\}$. Particularly, the detector with the acceptance region $A^* := \{s : m(s) \geq h(s)\}$, defined as

$$\Psi^*(s) := \begin{cases} 1 & s \in A^* \\ 0 & s \in \mathcal{S}\backslash A^*, \end{cases}$$

achieves the infimum. We note that Ψ^* is the likelihood ratio-based detector.

Proof. For notation simplicity, we use m and h to denote both the probability measure and the probability density of the machine-generated and human-generated text, respectively, with the specific meaning discernible from the context. For any detector $\Psi : \mathcal{S} \to \{1, 0\}$, denote A as its acceptance region, where $\Psi(s) = 1$ for $s \in A$, and $\Psi(s) = 0$ for $s \in \mathcal{S}\backslash A$. Then we have

$$\begin{aligned} \mathbb{P}_{s\sim m}[\Psi(s) \neq 1] + \mathbb{P}_{s\sim h}[\Psi(s) \neq 0] &= m(\mathcal{S}\backslash A) + h(A) \\ &= 1 - (m(A) - h(A)). \end{aligned} \tag{19}$$

Taking the infimum over all acceptance regions on both sides in (19) yields

$$\begin{aligned} \inf_{\Psi} \left\{ \mathbb{P}_{s\sim m}[\Psi(s) \neq 1] + \mathbb{P}_{s\sim h}[\Psi(s) \neq 0] \right\} &= \inf_{\Psi} \left\{ 1 - (m(A) - h(A)) \right\} \\ &= 1 - \sup_{\Psi} \left\{ (m(A) - h(A)) \right\} \\ &= 1 - \mathsf{TV}(m, h). \end{aligned}$$

Next, we proceed to show that the ration-based detector $\Psi^*(s)$, defined in the statement of Lemma 1, achieves the infimum. We first note that the acceptance region $A^* := \{s : m(s) \geq h(s)\}$ is a measurable set that is included in the collection of all acceptance regions since $\mathbb{I}_{\{m(s)\geq h(s)\}}$ is a measurable function. Therefore,

$$\mathbb{P}_{s\sim m}[\Psi^*(s) \neq 1] + \mathbb{P}_{s\sim h}[\Psi^*(s) \neq 0] \geq \inf_{\Psi} \left\{ \mathbb{P}_{s\sim m}[\Psi(s) \neq 1] + \mathbb{P}_{s\sim h}[\Psi(s) \neq 0] \right\}. \tag{20}$$

On the other hand, for any measurable set B, we have $B\backslash A^* = \{s \in B : m(s) < h(s)\}$ and $A^*\backslash B = \{s \notin B : m(s) \geq h(s)\}$ by the definition of A^*. Therefore, by the sigma-additivity of measure, we have

$$1 - (m(A^*) - h(A^*)) = 1 - (m(A^* \cap B) - h(A^* \cap B)) - (m(A^*\backslash B) - h(A^*\backslash B)). \tag{21}$$

In the right-hand side of (21), we note that $m(A^*\backslash B) - h(A^*\backslash B) \geq 0$ because our detector is likelihood-ratio-based. This implies we can upper bound the right-hand side in (21) by dropping the negative term as follows

$$1 - (m(A^*) - h(A^*)) \leq 1 - (m(A^* \cap B) - h(A^* \cap B)). \tag{22}$$

Further from the definition of the ratio-based detector, we note that $m(B\backslash A^*) - h(B\backslash A^*) < 0$. This implies $-(m(B\backslash A^*) - h(B\backslash A^*)) > 0$ and we can upper bound the right hand side of (22) by adding just the positive number $-(m(B\backslash A^*) - h(B\backslash A^*))$ as follows,

$$1 - (m(A^*) - h(A^*)) \leq 1 - (m(A^* \cap B) - h(A^* \cap B)) - (m(B\backslash A^*) - h(B\backslash A^*)). \tag{23}$$

From the sigma-additivity of measure, we can write

$$1 - (m(A^*) - h(A^*)) \leq 1 - (m(B) - h(B)). \tag{24}$$

Since the inequality in (24) holds for any measurable set B, we can write

$$\mathbb{P}_{s\sim m}[\Psi^*(s) \neq 1] + \mathbb{P}_{s\sim h}[\Psi^*(s) \neq 0] \leq \inf_{\Psi} \left\{\mathbb{P}_{s\sim m}[\Psi(s) \neq 1] + \mathbb{P}_{s\sim h}[\Psi(s) \neq 0]\right\}. \tag{25}$$

Hence, from the lower bound in (20) and upper bound in (25), we conclude that $\Psi^*(s)$ achieves the infimum, which completes the proof. □

The Le Cam's lemma directly applies to our detector D with threshold γ by noting that any detector can be implemented via a detector with a threshold. Indeed, define $D_\gamma : \mathcal{S} \to \{1, 0\}$ via

$$D_\gamma(s) := \begin{cases} 1 & D(s) \geq \gamma \\ 0 & D(s) < \gamma, \end{cases}$$

then it holds that $\{\Psi : \mathcal{S} \to \{1, 0\}\} \subseteq \{D_\gamma : \mathcal{S} \to \{1, 0\}, D : \mathcal{S} \to \mathbb{R}, \gamma \in \mathbb{R}\}$ because for any Ψ, we can choose D to be exactly the same as Ψ (since $\{1, 0\} \in \mathbb{R}$) and set $\gamma = 0.5$.

In fact, the detector Ψ^* is exactly the likelihood-ratio-based detector which, by the Neyman-Pearson lemma (Cover, 1999, Chapter 11), is optimal in this (simple-vs.-simple) hypothesis test setting.

Relationship to the tightness analysis in Sadasivan et al. (2023). The authors of Sadasivan et al. (2023) provide a tightness analysis for their AUROC upper bound. The main part of the proof is to show the tightness of Equation 18. Specifically, for any given human-generated text distribution h, they construct a machine-generated text distribution m and a detector D with some threshold γ, and show that the detector with the threshold achieves the equality in Equation 18. We note that their constructed detector with the threshold is exactly the likelihood-ratio-based detector. Moreover, a key difference between our result and theirs is that we show that the tightness can be achieved for any given distribution of m and h while they construct a specific m given h. While their specific construction of the machine distribution gives many insights into the problem, it is not necessary for achieving the tightness. This difference also implies that we can be more optimistic about the problem since the classifier achieving the tightness exists for any machine-generated distributions.

B.2 Proof of Theorem 1

The first part of the proof follows from the standard application of Chernoff's bounds (Vadhan, 1999, Appendix A). From the statement of Theorem 1, we note that the AUROC of the best possible detector is given by

$$\mathsf{AUROC} = \frac{1}{2} + \mathsf{TV}(m^{\otimes n}, h^{\otimes n}) - \frac{\mathsf{TV}(m^{\otimes n}, h^{\otimes n})^2}{2}. \tag{26}$$

Let us start in a hard detection setting where $m(s)$ and $h(s)$ are really close and we know that $\mathsf{TV}(m, h) = \delta$ where $\delta > 0$ is small. From the definition of TV distance, we know that there exists some set $A \in \mathcal{S}$ such that given the samples $s^m \sim m(s)$ and $s^h \sim h(s)$ it holds

$$\mathbb{P}(s^m \in A) - \mathbb{P}(s^h \in A) = \delta. \tag{27}$$

Let us 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 $\{s_i\}_{i=1}^n$ 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

$$\begin{aligned}
\mathbb{P}\left(\text{at least } \left(p + \frac{\delta}{2}\right) n \text{ samples of } h \text{ are in } A\right) &\leq \exp^{\frac{-n\delta^2}{2}} \\
\mathbb{P}\left(\text{at most } \left(p + \frac{\delta}{2}\right) n \text{ samples of } m \text{ are in } A\right) &\leq \exp^{\frac{-n\delta^2}{2}}.
\end{aligned} \tag{28}$$

Now, let us denote the set of $n-$tuples by A' which contains more than $\left(p + \frac{\delta}{2}\right) n$ samples of A. Therefore, we can bound

$$\begin{aligned}
\mathsf{TV}(m^{\otimes n}, h^{\otimes n}) \geq& \mathbb{P}(\{s_i^m\}_{i=1}^n \in A') - \mathbb{P}(\{s_i^h\}_{i=1}^n \in A') \\
\geq& (1 - \exp^{\frac{-n\delta^2}{2}}) - \exp^{\frac{-n\delta^2}{2}} \\
=& 1 - 2\exp^{\frac{-n\delta^2}{2}}.
\end{aligned} \tag{29}$$

The TV norm lower bound in (29) tells us the minimum value of $\mathsf{TV}(m^{\otimes n}, h^{\otimes n})$ for given n and δ. Therefore, if we need to obtain the AUROC of the best possible detector to be equal to, or higher than say $\epsilon \in [0.5, 1]$, which means we want

$$\frac{1}{2} + \mathsf{TV}(m^{\otimes n}, h^{\otimes n}) - \frac{\mathsf{TV}(m^{\otimes n}, h^{\otimes n})^2}{2} \geq \epsilon. \tag{30}$$

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

$$\frac{1}{2} + (1 - 2\exp^{\frac{-n\delta^2}{2}}) - \frac{(1 - 2\exp^{\frac{-n\delta^2}{2}})^2}{2} \geq \epsilon. \tag{31}$$

After expanding the squares, we get

$$\frac{1}{2} + (1 - 2\exp^{\frac{-n\delta^2}{2}}) - \frac{1}{2} - 2\exp^{-n\delta^2} + 2\exp^{\frac{-n\delta^2}{2}} \geq \epsilon. \tag{32}$$