Optimal Log-Likelihood Tests for Distinguishing Generative Models under Relative Entropy Constraints

Adam Vinestock
Reichman University
Herzliya, Israel
adam.vinestock@post.runi.ac.il

Alon Kipnis
Reichman University
Herzliya, Israel
alon.kipnis@runi.ac.il

Abstract*—Suppose that a sequence of tokens is generated by one of two sequential generative models (information sources) G_0 or G_1, and the goal is to determine which one. When the models themselves are inaccessible but past generations are available, a common machine-learning approach is to construct a detector based on the log-likelihood of the observed sequence under a third, "open-source" language model P. We show that, under mild regularity assumptions, the signal-to-noise ratio (SNR) of such a detector is proportional to the difference in relative entropies $D(\mathrm{G}_1\|\mathrm{P}) - D(\mathrm{G}_0\|\ \mathrm{P})$. We then study the problem of maximizing this difference over a class of probability models subject to a relative-entropy ball constraint. This formulation captures settings in which P is known a priori to be closer to one of the sources, for instance due to shared architecture or training methodology. We characterize the structure of the optimal P, derive the resulting SNR, and analyze its sensitivity to the constraint radius. Finally, we perform extensive numerical experiments using real text and modern large language models. The results support the theoretical predictions and reveal a somewhat counterintuitive phenomenon: in scenarios where G_1 is human and G_0 is a language model, the choice $\mathrm{P} = \mathrm{G}_0$ can be advantageous. We explain this effect by modeling the language model as a mixture of human sources, and show that $\mathrm{P} = \mathrm{G}_0$ is optimal under a local minimax analysis.

I. Introduction

A. Problem Setup

Let G_0 and G_1 be two probability distributions over a finite alphabet $\mathcal{X}$. We seek a probability distribution P so that the statistic $L(X;\mathrm{P}) := \log 1/\mathrm{P}(X)$ has optimal properties for testing the binary hypotheses

$$H_0 : X \sim \mathrm{G}_0 \text{ versus } H_1 : X \sim \mathrm{G}_1. \tag{1}$$

Specifically, we wish to minimize the error in (1) under a choice of P subject to a constraint on the relative entropy (Kullback-Leibler divergence) from H_0:

$$\mathrm{P} \in \mathcal{P}_\epsilon(\mathrm{G}_0) := \{\mathrm{P}' \,:\, D(\mathrm{G}_0\|\mathrm{P}') \le \epsilon\} \tag{2}$$

This problem appears different from common binary hypothesis testing setups arising frequently in information theory such as [1, Ch. 11], [2], [3], [4]. It is motivated by the popularity of the log-likelihood/perplexity test for identifying the source of text generated by a language model, as we explain next.

The test statistic $L(X;\mathrm{P})$ takes different names depending on the context: logarithmic loss [5], log perplexity [6], negative log-likelihood, and cross-entropy of P under the empirical distribution [7]. We will use the term logarithmic loss due to its prevalence in information theory.

B. Motivation: Language Model Authorship Detection

Consider the problem of deciding whether a sequence of tokens $t_{1:n} = (t_1, \ldots, t_n)$ such as a sentence or a document was written (generated) by a prescribed language model G_0, or not. In many cases, the statistician can only sample from G_0 without access to likelihood evaluations; consider OpenAI's family of language models as a typical example [8]. In [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], the test statistic for this detection problem is chosen to be $L(t_{1:n};\mathrm{P})$, where P is an open-source language model whose token probabilities can be evaluated. Henceforth, we denote P as the detector model. The problem (1) corresponds to the case where the alternative author is another generative model G_1. The constraint (2) arises when the detector model is known to be relatively close to the generative model G_0, a situation that can arise due to similarities in language model architectures and training processes. For example, [18] used the detector models P = GPT2 and P = Phi2 to decide whether a given sentence was written by G_0=GPT3.5 or by a human. These detectors and other open-source models are more similar to GPT3.5 than to human in many aspects.

The problem of minimizing the risk in (1) under the constraint (2) closely matches this situation.

C. Contributions

We analyze the test's Type II error subject to a Type I constraint. Specifically, given a significance level $\alpha \in (0,1)$, and with the understanding that our test rejects for large values of $L(X;\mathrm{P})$, we define the test's power by

$$1 - \beta(\mathrm{P},\alpha) := \Pr\left(L(X;\mathrm{P}) \ge \tau_\alpha \mid X \sim \mathrm{G}_1\right),$$

where τ_α is determined by

$$\tau_\alpha := \arg\min_\tau \left\{\tau \,:\, \Pr\left(L(X;\mathrm{P}) \ge \tau \mid X \sim \mathrm{G}_0\right) \le \alpha\right\}. \tag{3}$$

In Section II below, we tighten the authorship detection motivation by proving that, under some assumptions on the finite-sample distribution of $L(t_{1:n};\mathrm{P})$, the power of discriminating H_0 from H_1 in (1) depends on

$$\mathcal{H}(\mathrm{G}_1) - \mathcal{H}(\mathrm{G}_0) + \Delta(\mathrm{G}_1, \mathrm{G}_0; \mathrm{P}), \tag{4}$$

where $\mathcal{H}$ denotes entropy and

$$\Delta(\mathrm{G}_1, \mathrm{G}_0; \mathrm{P}) := D(\mathrm{G}_1 \| \mathrm{P}) - D(\mathrm{G}_0 \| \mathrm{P}). \tag{5}$$

Next, we consider P that maximizes $\Delta(\mathrm{G}_1, \mathrm{G}_0; \mathrm{P})$ under the ball constraint (2). This maximizer P^* is obtained by moving from G_0 linearly in a direction opposite to G_1. In particular, the solution is not a geometric scaling that often arises in optimal hypothesis testing situations [1, ch. 11]. We derive the behavior of $\Delta(\mathrm{G}_1, \mathrm{G}_0; \mathrm{P}^*)$ in G_0 and G_1 for small ϵ, an analysis that can help measure the optimality of a candidate P in practice.

Our results show in particular that the choice $\mathrm{P} = \mathrm{G}_0$ that is motivated by empirical evaluations, is in general not optimal. We show that optimality of $\mathrm{P} = \mathrm{G}_0$, however, arises in a local minimax analysis when G_0 is an isotropic convex combination of several models from the alternative.

Finally, we conduct extensive empirical evaluation using several language models and real and generated text data. These evaluations validate our theoretical assumptions for the finite-sample distribution of $L(t_{1:n};\mathrm{P})$ used in deriving our main results.

D. Unconstrained Optimal Detector

As background to the problem, we comment briefly on optimal detection when the constraint (2) is removed.

Any strictly monotone function of the log-likelihood ratio statistic $\Lambda(x) := \log(\mathrm{G}_1(x)/\mathrm{G}_0(x))$ is known to provide a test of minimal risk for (1) [19]. Therefore, when P is not subject to (2) and provided G_0 is absolutely continuous with respect to G_1, we may take P to be a scaled version of $\mathrm{G}_0(x)/\mathrm{G}_1(x)$, which leads to

$$L(x;\mathrm{P}) = \log \Lambda(x) - c,$$

where c is the normalizing constant. Therefore, this choice of P is optimal and yields a test of minimal risk across all tests for (1). The same conclusion holds when there exists a monotone function of $1/\Lambda(x)$ whose scaled and normalized version diverges by ϵ or less from G_0 in relative entropy. Therefore, the results in this paper are of interest when this situation does not hold, as is the typical case in high-dimensional distributions like language modeling applications motivating this study.

E. Paper Structure

The rest of this paper is organized as follows. In Section II we connect the power of a test based on $L(\cdot;\mathrm{P})$ to $\Delta(\mathrm{P})$. In Section III, we derive our main theoretical results. In Section IV we report empirical results. Concluding remarks are in Section V. The proofs are provided in a longer version of this paper, available at [20].

II. Asymptotic and finite-sample distributions

A. Asymptotic Negative Log-Likelihood

Let P_a be a language model. Sampling a sentence $t_{1:n} = (t_1, \ldots, t_n)$ from P_a is achieved by conditioning the current token probability on previous tokens and an initial context. Namely,

$$t_i \sim \mathrm{P}_a(\cdot \mid t_0, t_{1:i-1})\mathrm{P}_a(t_0), \quad i = 1, \ldots, n, \tag{6}$$

for some initial state t_0 that can represent the initial context. Suppose that we evaluate $L(t_{1:n};\mathrm{P}_b)$ for a second language model P_b. Under some conditions on the laws $(\mathrm{P}_a, \mathrm{P}_b)$, the limit of $L(t_{1:n};\mathrm{P})/n$ as $n \to \infty$ exists almost surely and obeys

$$\lim_{n\to\infty} \frac{L(t_{1:n};\mathrm{P}_b)}{n} = \mathcal{H}(\mathrm{P}_b;\mathrm{P}_a) = \mathcal{H}(\mathrm{P}_a) + \bar{\mathrm{D}}(\mathrm{P}_a \| \mathrm{P}_b), \tag{7}$$

where $\bar{\mathrm{D}}(\mathrm{P}_a \| \mathrm{P}_b)$ is the relative entropy rate of P_b to P_a [21, Ch. 7]. The term $\mathcal{H}(\mathrm{P}_b;\mathrm{P}_a)$ is denoted as the cross-entropy rate of P_b under the law P_a.

B. Finite-Sample Distribution

Relation (7) suggests the following finite-sample distribution assumption for $X^{(n)} := t_{1:n} \sim \mathrm{G}$:

$$L(X^{(n)};\mathrm{P})/n \overset{D}{=} \mathcal{H}(\mathrm{G}) + D(\mathrm{G} \| \mathrm{G}_0) + \sigma Z. \tag{8}$$

Here $\overset{D}{=}$ represents equality in distribution, Z is a zero-mean unit-variance random variable, and $\sigma > 0$ vanishes as $n \to \infty$. The following assumptions posit that the distribution of Z is unimodal and is unaffected by P.

Let $\mathcal{F}$ denote some scale-location family of unimodal continuous distributions and let $X^{(n)} := t_{1:n}$ be a random sequence. Consider the following assumptions.

(A1) Under $H_0 : X^{(n)} \sim \mathrm{G}_0$, $L(X^{(n)};\mathrm{P})$ is a member of $\mathcal{F}$ with mean $\mathcal{H}(\mathrm{P};\mathrm{G}_0)$ and scale $\sigma_0^{(n)}$.

(A2) Under $H_1 : X^{(n)} \sim \mathrm{G}_1$, $L(X^{(n)};\mathrm{P})/n$ is a member of $\mathcal{F}$ with mean $\mathcal{H}(\mathrm{P};\mathrm{G}_1)$ and scale $\sigma_1^{(n)}$.

(A3) The asymptotic scales σ_0 and σ_1 are independent of P.

We have the following result.

Theorem 1. *Let G_0 and G_1 be two stationary ergodic sources. Suppose that $\mathcal{P}$ is a set of stationary ergodic sources such that for any $\mathrm{P} \in \mathcal{P}$, the relative entropy rates $D(\mathrm{G}_1\|\mathrm{P})$ and $D(\mathrm{G}_0\|\mathrm{P})$ exist and are finite. Assume A1-A3. Suppose that*

$$\mathrm{P}^* \in \arg\max_{\mathrm{P}\in\mathcal{P}} \Delta(\mathrm{G}_1, \mathrm{G}_0; \mathrm{P}),$$

and $X \sim \mathrm{G}_1$. For any prescribed level α and $\mathrm{P} \in \mathcal{P}$,

$$\beta(\mathrm{P}^*, \alpha) \le \beta(\mathrm{P}, \alpha).$$

Namely, the smallest asymptotic Type II error, and thus the maximal power, is obtained when P maximizes $\Delta(\mathrm{G}_1, \mathrm{G}_0; \mathrm{P})$.

III. Optimal Detector

Following Theorem 1, we are now interested in maximizing $\Delta(\mathrm{G}_1, \mathrm{G}_0; \mathrm{P})$ over a set of available discriminating models $\mathcal{P}$ satisfying the relative entropy ball constraint (2).

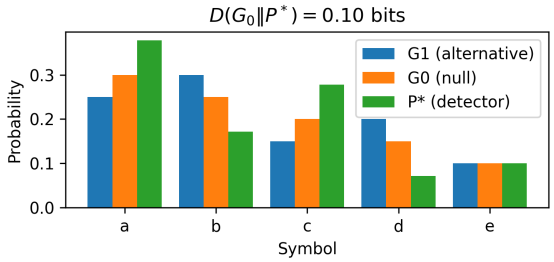

Fig. 1. Two probability distributions over 5 symbols and the detector distributions P^* maximizing the test's power.

A. Optimal Detector against a Simple Alternative

Theorem 2. *Let* G_0 *and* G_1 *be two probability models over an alphabet* $\mathcal{X}$ *such that* $\mathrm{G}_1 \ll \mathrm{G}_0$. *For* $\epsilon > 0$, *let*

$$\mathrm{P}^* \in \arg\max_{\mathrm{P}\in\mathcal{P}_\epsilon} \Delta(\mathrm{G}_1, \mathrm{G}_0; \mathrm{P}).$$

There exist $\gamma > 0$ *such that*

$$\mathrm{P}^*(x) = \mathrm{G}_0(x) - \gamma\left[\mathrm{G}_1(x) - \mathrm{G}_0(x)\right], \tag{9}$$

where γ *is determined by*

$$D(\mathrm{G}_0\|\mathrm{P}^*) = \epsilon. \tag{10}$$

Additionally,

$$\Delta(\mathrm{G}_1, \mathrm{G}_0; \mathrm{P}^*) = D(\mathrm{G}_1\|\mathrm{G}_0) + \frac{1}{\gamma}\left(D(\mathrm{P}^*\|\mathrm{G}_0) + \epsilon\right). \tag{11}$$

Remark 1. *The condition* $D(\mathrm{G}_0\|\mathrm{P}) \le \epsilon$ *also implies* $\mathrm{G}_0(x) \ge \gamma\left[\mathrm{G}_1(x) - \mathrm{G}_0(x)\right]$, *thus* $\mathrm{P}^*(x)$ *is well-defined. Indeed, if* $\mathrm{G}_0(x) - \gamma\left[\mathrm{G}_1(x) - \mathrm{G}_0(x)\right] \to 0$ *on the support of* G_0, *then* $D(\mathrm{G}_0\|\mathrm{P}) \to \infty$.

Theorem 2 shows that the optimal detector effectively subtracts mass from tokens in which G_1 is dominant, so as to increase the score $-\log \mathrm{P}(x)$ for such tokens. Interestingly, this subtraction is linear, rather than geometric as often arises in optimization (c.f. [1, Ch. 11]). Figure 1 illustrates an example of P^* for two distributions over 5 elements.

B. Local Optimal Detector

The following results provides the small-ϵ expansion of the optimal $\Delta(\mathrm{G}_1, \mathrm{G}_0; \mathrm{P})$ under the constraint (2). This expansion is provided in terms of the chi-squared divergence $D_{\chi^2}(\mathrm{G}_1\|\mathrm{G}_0)$ [22].

Theorem 3. *Let* G_0 *and* G_1 *be probability distributions on a countable alphabet* $\mathcal{X}$ *such that* $\mathrm{G}_1 \ll \mathrm{G}_0$ *and set*

$$D_{\chi^2}(\mathrm{G}_1\|\mathrm{G}_0) := \sum_{x\in\mathcal{X}} \frac{(\mathrm{G}_1(x) - \mathrm{G}_0(x))^2}{\mathrm{G}_0(x)} < \infty.$$

As $\epsilon \to 0$,

$$\Delta(\mathrm{G}_1, \mathrm{G}_0; P^*) = D(\mathrm{G}_1\|\mathrm{G}_0) + \sqrt{2\epsilon D_{\chi^2}(\mathrm{G}_1\|\mathrm{G}_0)} + O(\epsilon),$$

and

$$\mathrm{P}^* = \mathrm{G}_0(x)\left[1 + \sqrt{\frac{2\epsilon(1+o(1))}{D_{\chi^2}(\mathrm{G}_1\|\mathrm{G}_0)}}\left(1 - \frac{\mathrm{G}_1(x)}{\mathrm{G}_0(x)}\right)\right]$$

Theorem 3 shows that for small ϵ, the optimal detector P^* is driven by the likelihood ratio (LR) of the two hypotheses, discounting tokens with large LR values. Additionally, it follows that allowing P to deviate from G_0 by $D(\mathrm{G}_0\|\mathrm{P}) \le \epsilon$ increases $\Delta(\mathrm{G}_1, \mathrm{G}_0; \mathrm{P})$ by $\approx \sqrt{2\epsilon D_{\chi^2}(\mathrm{G}_1\|\mathrm{G}_0)}$. This quantifies the sub-optimality of using $\mathrm{P} = \mathrm{G}_0$.

C. Locally Minimax Optimal Detector

We now explore the optimal solution under a composite alternative, because in many cases we can think about the alternative as a mixture of sources such as humans with different writing style.

Let $\mathcal{G}_1$ be a family of alternative distributions with $G \ll \mathrm{G}_0$ for all $G \in \mathcal{G}_1$. We define the maximin power at a prescribed test level $\alpha > 0$ as

$$\sup_{\mathrm{P}\in\mathcal{P}_\epsilon} \inf_{G\in\mathcal{G}_1} \Pr\left(L(X;\mathrm{P}) \ge \tau_\alpha \mid X \sim G\right),$$

where α is determined as in (3). It is immediate to deduce from the arguments of Theorem 1 that the maximin power is controlled by the analogous maximin objective $\Delta(G, \mathrm{G}_0; \mathrm{P})$. Therefore, if G_0 is a fixed null distribution we seek

$$\Delta^*(\mathcal{G}_1, \mathrm{G}_0) = \sup_{\mathrm{P}\in\mathcal{P}_\epsilon} \inf_{G\in\mathcal{G}_1} \Delta(G, \mathrm{G}_0; \mathrm{P}). \tag{12}$$

The following theorem provides $\Delta^*(\mathcal{G}_1, \mathrm{G}_0)$ and a least-favorable prior π^* on $\mathcal{G}_1$ for small values of ϵ.

Theorem 4. *Consider the minimization* (12). *There exists a distribution* π^* *on* $\mathcal{G}_1$ *such that, as* $\epsilon \to 0$,

$$\Delta^*(\mathcal{G}_1, \mathrm{G}_0) = \mathbb{E}_{G\sim\pi^*}\left[D(G\|\mathrm{G}_0)\right] + \sqrt{2\epsilon D_{\chi^2}(\bar{G}_{\pi^*}\|\mathrm{G}_0)} + o(1),$$

where

$$\bar{G}_{\pi^*}(x) = \mathbb{E}_{G\sim\pi^*}\left[G(x)\right], \tag{13}$$

and π^* *is the minimizer of*

$$\inf_{\pi}\left[\mathbb{E}_{G\sim\pi}\left[D(G\|\mathrm{G}_0)\right] + \sqrt{2\epsilon D_{\chi^2}(\bar{G}_\pi\|\mathrm{G}_0)}\right]. \tag{14}$$

Theorem 4 says that, locally as $\epsilon \to 0$, the maximin Δ is determined by a least-favorable convex combination π^*. The situation described in Theorem 4, is analogous to the well-known equivalence between minimax inference and Bayesian inference [23]. If $\mathcal{G}_1$ is convex, then convexity of the relative entropy implies that π^* is a point mass at $\arg\min_{G\in\mathcal{G}_1} D(G\|\mathrm{G}_0)$. The more interesting case arises when $\mathcal{G}_1$ is non-convex, in which case π^* is supported on elements of $\mathcal{G}_1$ closest to G_0. In what follows, we argue that for a non-convex $\mathcal{G}_1$ that isotropically surrounds G_0, the choice $\mathrm{P} = \mathrm{G}_0$ is optimal.

D. Optimality of $\mathrm{P} = \mathrm{G}_0$ under isotropic alternative set

Let Π_1 be the set of priors π on $\mathcal{G}_1$. For a given prior $\pi \in \Pi_1$, define

$$\bar{D}_\pi := \mathbb{E}_{G\sim\pi}\left[D(G\|\mathrm{G}_0)\right],$$
$$r_\pi := \sup G \in \mathrm{supp}(\pi)\big|D(G\|\mathrm{G}_0) - \bar{D}_\pi\big|.$$

We think about r_π as a geometric anisotropy index of the prior π around G_0. Indeed, $r_\pi = 0$ if and only if all elements of $\mathrm{supp}(\pi)$ are equidistant from G_0 in relative entropy.

The following result establishes that if G_0 lies in the convex hull of $\mathcal{G}_1$ (represented by a mixture $\pi^\#$), then the sub-optimality of the choice $\mathrm{P} = \mathrm{G}_0$ can be bounded uniformly using the anisotropy index $r_{\pi^\#}$ of the least-favorable prior $\pi^\#$ on $\mathcal{G}_1$.

Theorem 5. *Let $\pi^\# \in \Pi_1$ such that $\mathrm{G}_0 = \bar{G}_{\pi^\#}$. For any alternative $\mathrm{G}_1 \in \mathcal{G}_1$ we have*

$$\Delta(\mathrm{G}_1, \mathrm{G}_0; \mathrm{G}_0) = D(\mathrm{G}_1\|\mathrm{G}_0) \geq \Delta^*(\mathcal{G}_1, \mathrm{G}_0) - r_{\pi^\#} + o(1).$$

We conclude:

Corollary 5.1. *If there exists $\pi^\#$ such that $\mathrm{G}_0 = \bar{G}_{\pi^\#}$ with $r_{\pi^\#} = 0$, then*

$$\inf_{G\in\mathcal{G}_1} \Delta(G, \mathrm{G}_0; \mathrm{G}_0) = \inf_{G\in\mathcal{G}_1} D(G\|\mathrm{G}_0) = \Delta^*(\mathcal{G}_1, \mathrm{G}_0) + o(1),$$

as $\epsilon \to 0$. Namely, the detector $\mathrm{P} = \mathrm{G}_0$ is maximin optimal.

IV. Empirical Results

We evaluate the theoretical predictions on sentence-level authorship detection. The dataset consists of text from three domains (Wikipedia articles, news articles, and scientific abstracts) written by five authors: four LLMs (Llama-3.1-8B, Falcon-7B, GPT-3.5, DeepSeek-R1) and humans. All authors wrote about the same topics within each domain, ensuring that detection reflects authorship rather than content differences. In total, we analyze approximately 4,500 documents comprising over 270,000 sentences (across all authors) after restricting to lengths of 10–40 words to ensure sufficient tokens for reliable negative log-likelihood estimation while maintaining adequate sample sizes.

For each sentence X, we compute $L(X;\mathrm{P})/n$ without preceding context, as P varies across several open-source detector models, some of which coincide with the LLM authors, enabling self-detection analysis ($\mathrm{P} = \mathrm{G}_0$). The cross-entropy difference (4) is estimated as $\hat{\mu}_1(\mathrm{P}) - \hat{\mu}_0(\mathrm{P})$, where $\hat{\mu}_i(\mathrm{P})$ denotes the sample mean of $L(X;\mathrm{P})$ over sentences from author G_i. For each pairwise author comparison, detection performance is measured by the area under the receiver operating characteristic curve (AUC).

A. Variance Insensitivity

Assumption A3 posits that the variance σ^2 of $L(X;\mathrm{P})$ is not affected by the detector model P. Figure 2 displays the empirical standard deviation $\sigma(L(X;\mathrm{P}))$ for each detector, stratified by sentence length. The max-to-min σ ratio across detectors ranges from $1.2\times$ to $1.5\times$ depending on the dataset, supporting the validity of A3.

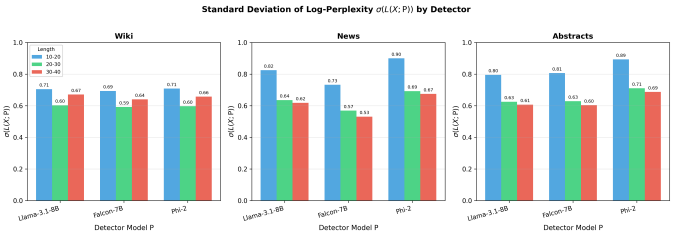

Fig. 2. **Weak dependence of the variance of the log-likelihood statistic $L(X;\mathrm{P})$ in P.** Standard deviation $\sigma(L(X;\mathrm{P}))$ of sentence-level negative log-likelihood across three detector models (Llama-3.1-8B, Falcon-7B, Phi-2), stratified by sentence length (10–20, 20–30, 30–40 words). The max-to-min σ ratio across detectors is 1.2–1.5$\times$ per dataset, indicating that variance does not strongly depend on the choice of P, supporting Assumption A3.

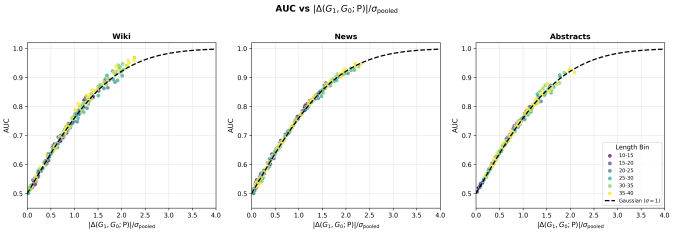

Fig. 3. **Empirical AUC matches Gaussian prediction after normalization.** AUC versus the normalized cross-entropy difference estimator $|\hat{\mu}_1(\mathrm{P}) - \hat{\mu}_0(\mathrm{P})|/\sigma_{\mathrm{pooled}}$ for pairwise author comparisons across three domains (sentences with 10–40 words). After normalization by the pooled standard deviation, points from all length bins collapse onto the theoretical Gaussian prediction $\mathrm{AUC} = \Phi(x/\sqrt{2})$ (dashed line, $\sigma = 1$), achieving Pearson $r = 0.99$. This supports the modeling assumption that the log-likelihood statistic $L(X;\mathrm{P})$ is unimodal and variance-stable, as posited in Assumptions A1–A2 of Theorem 1.

B. AUC versus Δ

Theorem 1 predicts that the power of a test based on the statistic $L(X;\mathrm{P})$ is governed by $\Delta(G_1, G_0; \mathrm{P})$. Under the Gaussian model (8), AUC relates to the normalized separation as $\mathrm{AUC} = \Phi(|\Delta|/(\sigma\sqrt{2}))$. Figure 3 plots empirical AUC against $|\hat{\Delta}|/\hat{\sigma}_{\mathrm{pooled}}$ for all pairwise comparisons, where $\hat{\sigma}_{\mathrm{pooled}}$ is the pooled standard deviation of each pair. The tight fit to the theoretical curve confirms that the Gaussian approximation in Theorem 1 accurately describes the separation.

C. Human versus LLM Detection

Corollary 5.1 suggests that when the alternative $\mathcal{G}_1$ isotropically surrounds G_0, the choice $\mathrm{P} = \mathrm{G}_0$ is maximin optimal. This situation may arise in human versus LLM detection under the following model: human text is a mixture of diverse sources centered around the LLM. Figure 4 reports AUC for distinguishing human text from each LLM author using each detector. The results show that the matched detector ($\mathrm{P} = G_0$, where G_0 is the LLM) consistently achieves the best or second-best AUC, providing empirical support for the relevance of Corollary 5.1 in this setting. Table I further quantifies this advantage: detecting human text against an LLM using $\mathrm{P} = G_0$ yields mean AUC of 0.83, compared to 0.63 for LLM-versus-LLM detection—a gap of $+0.21$.

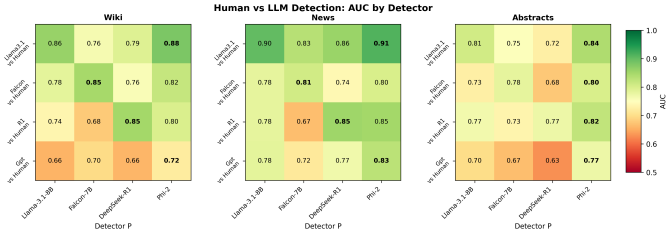

Fig. 4. **AUC for detecting human versus LLM-generated text.** Each cell shows the AUC for distinguishing human text from a specific LLM author (rows) using detector P (columns), for sentences of 10–40 words. Bold values indicate the best detector for each row. Self-detection ($P = G_0$) achieves best AUC in 4/9 cases and second-best in the remaining 5/9 (excluding GPT row, which has no corresponding detector).

TABLE I

AUC when detector matches source ($P = G_0$)**.** EACH ROW SHOWS MEAN AUC $\pm$ SAMPLE STD FOR TWO SCENARIOS: (1) DETECTING HUMAN TEXT AGAINST AN LLM USING THE LLM'S OWN MODEL AS DETECTOR, AND (2) DISTINGUISHING TWO LLMS USING ONE OF THEM AS DETECTOR. Δ IS THE DIFFERENCE BETWEEN COLUMNS. BASED ON THREE MATCHED DETECTOR–AUTHOR PAIRS (LLAMA-3.1-8B, FALCON-7B, DEEPSEEK-R1) WITH n=3 AND n=6 COMPARISONS, RESPECTIVELY. SENTENCES: 10–40 WORDS.

Dataset	Human vs. G_0	G_0 vs. G_1	Δ
Wiki	0.853 ± 0.006	0.659 ± 0.065	$+0.194$
News	0.854 ± 0.047	0.626 ± 0.094	$+0.229$
Abstracts	0.787 ± 0.018	0.592 ± 0.013	$+0.195$
Mean	**0.831**	**0.626**	**+0.206**

V. CONCLUSIONS

We considered a binary hypothesis test, and analyzed the structure of a test based on the negative log-likelihood (aka. log perplexity) with respect to a third probability model P. After assuming that the test statistic belongs to a scale-location family with variance independent of P, we derived the power-maximizing P under a relative entropy ball constraint, both under a simple and a minimax composite setup. We empirically validated the relevance of the theoretical results to authorship detection of language models.

Our evaluations propose the following model for large language models in the context of authorship discrimination: such models are isotropic convex combinations of various human text sources.

REFERENCES

[1] T. Cover and J. A. Thomas, "Elements of information theory," 2006.

[2] Y. Li and V. Y. Tan, "Second-order asymptotics of sequential hypothesis testing," *IEEE Transactions on Information Theory*, vol. 66, no. 11, pp. 7222–7230, 2020.

[3] M. Bell and Y. Kochman, "On universality and training in binary hypothesis testing," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 3824–3846, 2021.

[4] R. AHLSWEDE and I. CSISZAR, "Hypothesis testing with communication constraints," *IEEE transactions on information theory*, vol. 32, no. 4, pp. 533–542, 1986.

[5] N. Merhav and M. Feder, "Universal prediction," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.

[6] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2023, third edition draft.

[7] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[8] OpenAI, "GPT-4 technical report," 2023.

[9] S. Gehrmann, H. Strobelt, and A. M. Rush, "GLTR: Statistical detection and visualization of generated text," *arXiv preprint arXiv:1906.04043*, 2019.

[10] C. Vasilatos, M. Alam, T. Rahwan, Y. Zaki, and M. Maniatakos, "Howkgpt: Investigating the detection of chatgpt-generated university student homework through context-aware perplexity analysis," *arXiv preprint arXiv:2305.18226*, 2023.

[11] Y. Tian, H. Chen, X. Wang, Z. Bai, Q. ZHANG, R. Li, C. Xu, and Y. Wang, "Multiscale positive-unlabeled detection of ai-generated texts," in *The Twelfth International Conference on Learning Representations*, 2023.

[12] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, and T. Goldstein, "Spotting llms with binoculars: Zero-shot detection of machine-generated text," in *International Conference on Machine Learning*. PMLR, 2024, pp. 17 519–17 537.

[13] Y. Xu, Y. Wang, Y. Bi, H. Cao, Z. Lin, Y. Zhao, and F. Wu, "Training-free LLM-generated text detection by mining token probability sequences," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=vo4AHjowKi

[14] W. Huang, A. Murakami, and J. Grieve, "Attributing authorship via the perplexity of authorial language models," *PloS one*, vol. 20, no. 7, p. e0327081, 2025.

[15] K. Taguchi, Y. Gu, and K. Sakurai, "The impact of prompts on zero-shot detection of ai-generated text," in *CEUR Workshop Proceedings*, vol. 3856. CEUR-WS, 2024.

[16] J. Xu, H. Zhang, Y. Yang, L. Yang, Z. Cheng, J. Lyu, B. Liu, X. Zhou, A. Bacchelli, Y. K. Chiam *et al.*, "One size does not fit all: Investigating efficacy of perplexity in detecting llm-generated code," *ACM Transactions on Software Engineering and Methodology*, 2024.

[17] S. Chakraborty, A. Bedi, S. Zhu, B. An, D. Manocha, and F. Huang, "Position: On the possibilities of AI-generated text detection," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., vol. 235. PMLR, 21–27 Jul 2024, pp. 6093–6115. [Online]. Available: https://proceedings.mlr.press/v235/chakraborty24a.html

[18] I. Kashtan and A. Kipnis, "An information-theoretic approach for detecting edits in ai-generated text," *Harvard Data Science Review*, no. Special Issue 5, 2024.

[19] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[20] A. Vinestock and A. Kipnis, "Optimal log-likelihood tests for distinguishing generative models under relative entropy constraints," 2026, includes proofs. [Online]. Available: https://cs.idc.ac.il/~kipnis/

[21] R. M. Gray, *Entropy and information theory*. Springer Science & Business Media, 2011.

[22] Y. Polyanskiy and Y. Wu, *Information theory: From coding to learning*. Cambridge university press, 2025.

[23] J. O. Berger, *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[24] Y. Ingster and I. A. Suslina, *Nonparametric goodness-of-fit testing under Gaussian models*. Springer Science & Business Media, 2012, vol. 169.

[25] K. Fan, "Minimax theorems," *Proceedings of the National Academy of Sciences*, vol. 39, no. 1, pp. 42–47, 1953.

APPENDIX

This Appendix contains the proofs of Theorems 1, 2, 3, 4, 5, and Corollary 5.1.

Proof of Theorem 1

Let $F_0 \in \mathcal{F}$ have zero mean and unit variance. Denote by Z the random variable with distribution $\Pr(Z \le x) = F_0(x)$. Denote by $\bar{F}_0(x) = 1 - F_0(x)$ the corresponding survival function. By A1, we have

$$\begin{aligned} p(x) &:= \Pr_{X \sim \mathrm{G}_0}\left[L(X;\mathrm{P}) \ge x\right] \\ &= \Pr_{X \sim \mathrm{G}_0}\left[\frac{L(X;\mathrm{P}) - \mathcal{H}(\mathrm{P};\mathrm{G}_0)}{\sigma_0} \ge \frac{x - \mathcal{H}(\mathrm{P};\mathrm{G}_0)}{\sigma_0}\right] \\ &= \bar{F}_0\left(\frac{x - \mathcal{H}(\mathrm{P};\mathrm{G}_0)}{\sigma_0}\right). \end{aligned}$$

For a prescribed significance level $\alpha \in (0,1)$, the test rejects if $p(X) \le \alpha$. By A2, we have under $X \sim \mathrm{G}_1$ the equality in distribution:

$$L(X;\mathrm{P}) \overset{D}{=} \mathcal{H}(\mathrm{P};\mathrm{G}_1) + \sigma_1 Z,$$

Therefore, the test's power is given by

$$\begin{aligned} 1 - \beta &= \Pr_{X \sim \mathrm{G}_1}\left[p(X) \le \alpha\right] \\ &= \Pr\left(\bar{F}_0\left(\frac{\mathcal{H}(\mathrm{P};\mathrm{G}_1) - \mathcal{H}(\mathrm{P};\mathrm{G}_0)}{\sigma_0} + \frac{\sigma_1}{\sigma_0} Z\right) \le \alpha\right). \end{aligned}$$

By the unimodality assumption, the last expression is a monotonic non-decreasing function of

$$\mathcal{H}(\mathrm{P};\mathrm{G}_1) - \mathcal{H}(\mathrm{P};\mathrm{G}_0) = \mathcal{H}(\mathrm{G}_1) - \mathcal{H}(\mathrm{G}_0) + \Delta(\mathrm{G}_1, \mathrm{G}_0; \mathrm{P}).$$

The claim follows because only the last term depends on P. □

Proof of Theorem 2

We need to maximize $\Delta(P; \mathrm{G}_1, \mathrm{G}_0)$ over P with constraints $D(\mathrm{G}_0 \| \mathrm{P}) \le \epsilon$ and $\sum_{x \in \mathcal{X}} \mathrm{P}(x) = 1$. The Lagrangian is

$$\mathcal{L}(\mathrm{P}, \lambda, \mu) := D(\mathrm{G}_1 \| \mathrm{P}) - D(\mathrm{G}_0 \| \mathrm{P}) + \lambda\left(\epsilon - D(\mathrm{G}_0 \| \mathrm{P})\right) + \mu\left(1 - \sum_{x \in \mathcal{X}} \mathrm{P}(x)\right).$$

Differentiating with respect to P gives

$$\frac{\partial \mathcal{L}}{\partial \mathrm{P}} = -\frac{\mathrm{G}_1}{\mathrm{P}} + (1+\lambda)\frac{\mathrm{G}_0}{\mathrm{P}} - \mu.$$

For any x with $\mathrm{P}(x) > 0$, the stationary point P^* satisfies

$$\mathrm{P}(x) = \frac{(1+\lambda)\mathrm{G}_0(x) - \mathrm{G}_1(x)}{\mu}. \tag{15}$$

Because P is a probability distribution, we must have $(1+\lambda)\mathrm{G}_0(x) \ge \mathrm{G}_1(x)$ and $\sum_{x \in \mathcal{X}} \mathrm{P}(x) = 1$. Both conditions are satisfied with $\mu = \lambda$ provided

$$1 + \lambda \ge \sup_{x \in \mathcal{X}} \frac{\mathrm{G}_1(x)}{\mathrm{G}_0(x)},$$

which in turn can be satisfied since $\mathcal{X}$ is finite and $\mathrm{G}_1 \ll \mathrm{G}_0$ implies that the likelihood ratio $\mathrm{G}_1/\mathrm{G}_0$ is bounded.

In the next lemma we show that the constraint $D(\mathrm{G}_0 \| \mathrm{P}) \le \epsilon$ is binding, thus condition (10) must hold. The proof of this lemma is provided at the end of the proof of Theorem 2.

Lemma 5.1. *Assume* $\mathrm{G}_1 \ne \mathrm{G}_0$ *and* $\varepsilon > 0$*. Consider*

$$\sup_{\mathrm{P}} \ \Delta(\mathrm{P}) := D(\mathrm{G}_1 \| \mathrm{P}) - D(\mathrm{G}_0 \| \mathrm{P}) \qquad \textit{s.t.} \qquad D(\mathrm{G}_0 \| \mathrm{P}) \le \varepsilon.$$

Then any maximizer $\mathrm{P}^\star$ *satisfies* $D(\mathrm{G}_0 \| \mathrm{P}^\star) = \varepsilon$*.*

As an intuition for the statement in Lemma 5.1, notice $\Delta(\mathrm{P})$ strictly increases when moving P away from G_0 in the directions that separate G_1 from G_0.

It is left to evaluate $\Delta^*(\mathrm{G}_1, \mathrm{G}_0)$. Substituting P^* to $\Delta(\mathrm{P}) := \Delta(\mathrm{G}_1, \mathrm{G}_0; \mathrm{P})$ and using that $\mathrm{G}_0(x) - \mathrm{G}_1(x) = \frac{1}{\gamma}(\mathrm{P}^*(x) - \mathrm{G}_0(x))$, we get

$$\begin{aligned}\Delta(\mathrm{P}^*) - \Delta(\mathrm{G}_0) &= \sum_x (\mathrm{G}_0(x) - \mathrm{G}_1(x)) \log \frac{\mathrm{P}^*(x)}{\mathrm{G}_0(x)} \\ &= \frac{1}{\gamma} \sum_x (\mathrm{P}^*(x) - \mathrm{G}_0(x)) \log \frac{\mathrm{P}^*(x)}{\mathrm{G}_0(x)} \\ &= \frac{1}{\gamma} \left[D(\mathrm{P}^* \| \mathrm{G}_0) + D(\mathrm{G}_0 \| \mathrm{P}^*) \right] \\ &= \frac{1}{\gamma} \left[D(\mathrm{P}^* \| \mathrm{G}_0) + \epsilon \right].\end{aligned}$$

Proof of Lemma 5.1: Suppose by contradiction that there exists an optimal $\mathrm{P}^\star$ with $D(\mathrm{G}_0 \| \mathrm{P}^\star) < \epsilon$. Since $\mathrm{G}_1 \neq \mathrm{G}_0$, there exists a measurable set A such that $\mathrm{G}_1(A) > \mathrm{G}_0(A)$. For $0 < \delta < 1$ define

$$\mathrm{P}_\delta := (1-\delta)\mathrm{P}^\star + \delta R, \qquad R := \mathrm{G}_0(\cdot \mid A^c).$$

Then P_δ is a valid distribution and, by continuity of $\mathrm{P} \to D(\mathrm{G}_0 \| \mathrm{P})$, we have $D(\mathrm{G}_0 \| \mathrm{P}_\delta) \leq \varepsilon$ for all sufficiently small $\delta > 0$. Since R puts zero mass on A, we have $\mathrm{P}_\delta(x) = (1-\delta)\mathrm{P}^\star(x)$ for $x \in A$, hence $\log \mathrm{P}_\delta(x) = \log \mathrm{P}^\star(x) + \log(1-\delta) < \log \mathrm{P}^\star(x)$ on A. Therefore,

$$\begin{aligned}\Delta(\mathrm{G}_1, \mathrm{G}_0; \mathrm{P}) &= \text{const} - \sum_{x \in \mathcal{X}} (\mathrm{G}_1(x) - \mathrm{G}_0(x)) \log \mathrm{P}(x) \\ &> \text{const} - \sum_{x \in \mathcal{X}} (\mathrm{G}_1(x) - \mathrm{G}_0(x)) \log \mathrm{P}^\star(x) \\ &= \Delta(\mathrm{G}_1, \mathrm{G}_0; \mathrm{P}^\star)\end{aligned}$$

contradicting optimality of $\mathrm{P}^\star$. □

A. Proof of Theorem 3

Set $V := D_{\chi^2}(\mathrm{G}_1 \| \mathrm{G}_0)$. For $\lambda > 0$ large enough so that $\mathrm{P}_\lambda(x) \geq 0$ for all x, define

$$\begin{aligned}r(x) &:= \frac{\mathrm{G}_1(x) - \mathrm{G}_0(x)}{\mathrm{G}_0(x)}, \\ \mathrm{P}_\lambda(x) &:= \mathrm{G}_0(x) - \frac{\mathrm{G}_1(x) - \mathrm{G}_0(x)}{\lambda} = \mathrm{G}_0(x)\Big(1 - \frac{r(x)}{\lambda}\Big),\end{aligned} \tag{16}$$

and $\epsilon(\lambda) := D(\mathrm{G}_0 \| P_\lambda)$. Notice that

$$\sum_{x \in \mathcal{X}} \mathrm{G}_0(x) r(x) = 0 \tag{17}$$

and

$$V = D_{\chi^2}(\mathrm{G}_1 \| \mathrm{G}_0) = \sum_{x \in \mathcal{X}} \mathrm{G}_0(x) r^2(x). \tag{18}$$

We work with the one-parameter family P_λ in (16) and derive expansions for $\epsilon(\lambda) = D(\mathrm{G}_0 \| \mathrm{P}_\lambda)$ and $\Delta(\mathrm{P}_\lambda)$ as $\lambda \to \infty$, then eliminate λ in favor of ϵ.

By (16),

$$\epsilon(\lambda) = D(\mathrm{G}_0 \| P_\lambda) = -\sum_x \mathrm{G}_0(x) \log\Big(1 - \frac{r(x)}{\lambda}\Big).$$

From $\log(1-u) = -u - \frac{u^2}{2} + O(u^3)$ as $u \to 0$, we get

$$-\log\Big(1 - \frac{r(x)}{\lambda}\Big) = \frac{r(x)}{\lambda} + \frac{r(x)^2}{2\lambda^2} + O\Big(\frac{1}{\lambda^3}\Big).$$

Using this, (17), and (18), leads to

$$\begin{aligned}\epsilon(\lambda) &= \frac{1}{\lambda} \sum_x \mathrm{G}_0(x) r(x) + \frac{1}{2\lambda^2} \sum_x \mathrm{G}_0(x) r^2(x) + O\Big(\frac{1}{\lambda^3}\Big) \\ &= \frac{1}{2\lambda^2} V + O\Big(\frac{1}{\lambda^3}\Big).\end{aligned} \tag{19}$$

We now expand $\Delta(P_\lambda)$ for small values of $1/\lambda$. Using $P_\lambda(x) = \mathrm{G}_0(x)\big(1 - r(x)/\lambda\big)$,

$$\begin{aligned}
\Delta(P_\lambda) &= D(\mathrm{G}_1\|\mathrm{P}) - D(\mathrm{G}_0\|\mathrm{P}) \\
&= \sum_x \mathrm{G}_1(x) \log \frac{\mathrm{G}_1(x)}{P_\lambda(x)} - \sum_x \mathrm{G}_0(x) \log \frac{\mathrm{G}_0(x)}{P_\lambda(x)} \\
&= \sum_x \mathrm{G}_1(x) \log \frac{\mathrm{G}_1(x)}{\mathrm{G}_0(x)} - \sum_x (\mathrm{G}_1(x) - \mathrm{G}_0(x)) \log \frac{P_\lambda(x)}{\mathrm{G}_0(x)} \\
&= D(\mathrm{G}_1\|\mathrm{G}_0) - \sum_x (\mathrm{G}_1(x) - \mathrm{G}_0(x)) \log \Big(1 - \frac{r(x)}{\lambda}\Big).
\end{aligned}$$

Using again $\log(1-u) = -u - \frac{u^2}{2} + O(u^3)$ with $u = r(x)/\lambda$, and noting that $\mathrm{G}_1(x) - \mathrm{G}_0(x) = \mathrm{G}_0(x)r(x)$, we get

$$\begin{aligned}
\Delta(P_\lambda) &= D(\mathrm{G}_1\|\mathrm{G}_0) + \sum_x \mathrm{G}_0(x) r(x) \left[\frac{r(x)}{\lambda} + \frac{r(x)^2}{2\lambda^2} + O\Big(\frac{1}{\lambda^3}\Big)\right] \\
&= D(\mathrm{G}_1\|\mathrm{G}_0) + \frac{1}{\lambda} \sum_x \mathrm{G}_0(x) r(x)^2 + O\Big(\frac{1}{\lambda^2}\Big) \\
&= D(\mathrm{G}_1\|\mathrm{G}_0) + \frac{V}{\lambda} + O\Big(\frac{1}{\lambda^2}\Big).
\end{aligned}$$

Thus,

$$\Delta(P_\lambda) = D(\mathrm{G}_1\|\mathrm{G}_0) + \frac{V}{\lambda} + O\Big(\frac{1}{\lambda^2}\Big). \tag{20}$$

Finally, from (19), we get $\epsilon(\lambda) = \frac{V}{2\lambda^2} + O(\lambda^{-3})$, hence

$$\frac{1}{\lambda} = \sqrt{\frac{2\epsilon}{V}} \; + \; O(\epsilon), \qquad \text{as } \epsilon \to 0.$$

Plugging this into (20) gives

$$\Delta(P_\epsilon) = D(\mathrm{G}_1\|\mathrm{G}_0) + V\left(\sqrt{\frac{2\epsilon}{V}} + O(\epsilon)\right) + O(\epsilon) = D(\mathrm{G}_1\|\mathrm{G}_0) + \sqrt{2V\epsilon} + O(\epsilon).$$

□

B. Proof of Theorem 4

Let Π_1 denote the set of priors π over $\mathcal{G}_1$. For a fixed P, by a standard identity (c.f. [24]),

$$\begin{aligned}
\inf_{G \in \mathcal{G}_1} \Delta(G, \mathrm{G}_0; \mathrm{P}) &= \inf_{\pi \in \Pi_1} \mathbb{E}_{G\sim\pi}\left[\Delta(G, \mathrm{G}_0; \mathrm{P})\right] \\
&= \inf_{\pi \in \Pi_1} \int_{\mathcal{G}_1} \Delta(G, \mathrm{G}_0; \mathrm{P}) d\pi(G).
\end{aligned}$$

Since the set of probability measures $\mathcal{P}_\epsilon$ is convex and compact, and since Π_1 is convex, we may apply the minimax theorem to swap the order of optimization (c.f. [25]). This leads to

$$\Delta^*(\mathcal{G}_1, \mathrm{P}) = \inf_{\pi \in \Pi_1} \sup_{\mathrm{P} \in \mathcal{P}_\epsilon} \mathbb{E}_{G\sim\pi}\left[\Delta(G, \mathrm{G}_0; \mathrm{P})\right] \tag{21}$$

We consider first the inner optimization in (21). Let $J(\mathrm{P}, \pi) := \mathbb{E}_{G\sim\pi}\left[\Delta(G, \mathrm{G}_0; \mathrm{P})\right]$. We have:

$$\begin{aligned}
J(\mathrm{P}, \pi) &= \int_{\mathcal{G}_1} \left[D(G\|\mathrm{P}) - D(\mathrm{G}_0\|\mathrm{P})\right] d\pi(G) \\
&= \int_{\mathcal{G}_1} \left[D(G\|\mathrm{G}_0) + \sum_{x\in\mathcal{X}} (G(x) - \mathrm{G}_0(x)) \log \frac{\mathrm{G}_0(x)}{\mathrm{P}(x)}\right] d\pi(G) \\
&= \mathbb{E}_{G\sim\pi}\left[D(G\|\mathrm{G}_0)\right] + \sum_{x\in\mathcal{X}} \big(\mathrm{G}_0(x) - \bar{G}_\pi(x)\big) \log \frac{\mathrm{G}_0(x)}{\mathrm{P}(x)},
\end{aligned} \tag{22}$$

where $\bar{G}_\pi(x) := \int G(x) d\pi(G)$.

We now solve for the optimal P subject to the constraint $D(\mathrm{G}_0\|\mathrm{P}) \le \epsilon$. Let $\mathrm{P}(x) = \mathrm{G}_0(x)(1+\delta(x))$ where $\mathbb{E}_{\mathrm{G}_0}[\delta] = 0$. By $\log(1+x) = x - x^2/2 + o(x^2)$,

$$D(\mathrm{G}_0\|\mathrm{P}) = \frac{1}{2}\sum_{x\in\mathcal{X}} \mathrm{G}_0(x)\delta(x)^2 = \frac{1}{2}\|\delta\|_{\mathrm{G}_0}^2 + o(\|\delta\|_{\mathrm{G}_0}^2).$$

Therefore, the constraint (2) implies

$$\|\delta\|_{\mathrm{G}_0}^2 \le 2\epsilon(1+o(1)).$$

Using the approximation $\log(\mathrm{G}_0/\mathrm{P}) = -\log(1+\delta) \approx -\delta$, the variational term in (22) becomes a linear functional of δ:

$$\begin{aligned}\sum_{x\in\mathcal{X}} (\mathrm{G}_0(x) - \bar{G}_\pi(x))(-\delta(x)) &= \sum_{x\in\mathcal{X}} \mathrm{G}_0(x)\left(1 - \frac{\bar{G}_\pi(x)}{\mathrm{G}_0(x)}\right)(-\delta(x)) \\ &= \langle \frac{\bar{G}_\pi}{\mathrm{G}_0} - 1, \delta\rangle_{\mathrm{G}_0}. \end{aligned} \tag{23}$$

By the Cauchy-Schwarz inequality, the supremum of this inner product subject to $\|\delta\|_{\mathrm{G}_0} \le \sqrt{2\epsilon}$ is achieved when δ aligns with the vector $v = \bar{G}_\pi/\mathrm{G}_0 - 1$. The maximum value is $\sqrt{2\epsilon}\,\|\bar{G}_\pi/\mathrm{G}_0 - 1\|_{\mathrm{G}_0}$. Recognizing that this norm corresponds to the Chi-square divergence, we have:

$$\sup_{\mathrm{P}\in\mathcal{P}_\epsilon} \sum_{x\in\mathcal{X}} (\mathrm{G}_0(x) - \bar{G}_\pi(x)) \log \frac{\mathrm{G}_0(x)}{\mathrm{P}(x)} = \sqrt{2\epsilon(1+o(1))\cdot D_{\chi^2}(\bar{G}_\pi\|\mathrm{G}_0)}. \tag{24}$$

Substituting this back into (21) yields,

$$\Delta^*(\mathcal{G}_1, \mathrm{P}) = \inf_{\pi\in\Pi_1} \left[\mathbb{E}_{G\sim\pi}[D(G\|\mathrm{G}_0)] + \sqrt{2(\epsilon+o(1))\cdot D_{\chi^2}(\bar{G}_\pi\|\mathrm{G}_0)}\right]. \tag{25}$$

□

Proof of Theorem 5: By Theorem 3 and following similar arguments as in the proof of Theorem 4, for any prior π on $\mathcal{G}_1$,

$$\begin{aligned}\sup_{\mathrm{P}\in\mathcal{P}_\epsilon} \inf_{G\in\mathcal{G}_1} \Delta(G, \mathrm{G}_0; \mathrm{P}) &\le \sup_{\mathrm{P}\in\mathcal{P}_\epsilon} \int_{\mathcal{G}} \Delta(G, \mathrm{G}_0; \mathrm{P})\pi(G) \\ &= \left[\bar{D}_\pi + \sqrt{2\epsilon D_{\chi^2}(\bar{G}_\pi\|\mathrm{G}_0)} + o(1)\right],\end{aligned}$$

as $\epsilon \to 0$. Thus, for $\pi = \pi^\#$ with $\mathrm{G}_0 = \bar{G}_{\pi^\#}$, the chi-squared term vanishes and we get

$$\Delta^*(\mathcal{G}_1, \mathrm{G}_0) \le \bar{D}_{\pi^\#} + o(1) \tag{26}$$

For a specific alternative $\mathrm{G}_1 \in \mathcal{G}_1$, we use (26) to write

$$\begin{aligned}D(\mathrm{G}_1\|\mathrm{G}_0) &= \bar{D}_{\pi^\#} + \left(D(\mathrm{G}_1\|\mathrm{G}_0) - \bar{D}_{\pi^\#}\right) \\ &\ge \Delta^*(\mathcal{G}_1, \mathrm{G}_0) + \left(D(\mathrm{G}_1\|\mathrm{G}_0) - \bar{D}_{\pi^\#}\right) + o(1) \\ &\ge \Delta^*(\mathcal{G}_1, \mathrm{G}_0) - \left|D(\mathrm{G}_1\|\mathrm{G}_0) - \bar{D}_{\pi^\#}\right| + o(1).\end{aligned}$$

It follows that

$$\begin{aligned}o(1) + D(\mathrm{G}_1\|\mathrm{G}_0)\pi^\#(G) &\ge \Delta^*(\mathcal{G}_1, \mathrm{G}_0) - \sup G \in \mathrm{supp}(\pi^\#) \left|D(\mathrm{G}_1\|\mathrm{G}_0) - \bar{D}_{\pi^\#}\right| \\ &= \Delta^*(\mathcal{G}_1, \mathrm{G}_0) - r_{\pi^\#}.\end{aligned}$$

□