Equivalence Testing of Parametric Models via the Effective Dimension

Marc Kaufmann [*1] **Rishit Chatterjee** [2] **Alexandre Dang** [3] **Mirco Giacobbe** [4 5] **Pascal Berrang** [4 5]

Abstract

A natural way to test whether two parameterized models behave equivalently is to query both on n i.i.d. inputs and check whether their predictions agree. What such a test certifies on the underlying input distribution is a statistical question, and classical answers via combinatorial complexity measures such as the VC dimension are loose for overparametrized families. We derive a theoretical guarantee for equivalence testing of parameterized probabilistic models using the notion of the Fisher-information–based effective dimension d_{eff} Abbas et al. (2021). The bound certifies expected predictive total-variation disagreement of $O(\varepsilon\sqrt{\ln n/n})$ on unseen data, where ε is a tunable parameter controlling the tradeoff between the error bound and the confidence level. We instantiate the result on binary logistic regression and derive sample-complexity bounds for testing model equivalence, describing the dichotomy between a *parameter-close* regime, where d_{eff} governs the rate, and a *parameter-distant* regime, where it does not.

1. Introduction

Machine-learning workflows routinely produce two models that should behave the same. A model owner may need to confirm that quantization, distillation, or a hardware migration has preserved the behavior of a reference checkpoint; a regulator or auditor may need to verify that a deployed system matches a model that was previously evaluated; a practitioner may want to confirm that two training runs with different random seeds have converged to the same input–output map. In each case, the goal is to certify that the two models make indistinguishable predictions on inputs drawn from a deployment distribution — not on every conceivable input, but on the ones that will actually be queried at test time. Even when both models are accessible, exhaustive comparison is infeasible whenever the input space is large, so the question is statistical rather than algorithmic: what we can certify, and at what confidence, from a finite sample.

The natural test is also the obvious one: draw n i.i.d. inputs from the deployment distribution, query both models, and accept if their predictions agree on every drawn input. The question is what such a test certifies. Classical answers go through uniform convergence over a hypothesis class with controlled combinatorial complexity, the VC dimension, or the raw parameter count, and yield bounds that scale with that complexity (Vapnik & Chervonenkis, 1971; Blumer et al., 1989; Bartlett et al., 2019). For modern overparameterized families, those bounds are loose to the point of being uninformative: the parameter count of a deep network is orders of magnitude larger than the number of test queries any practitioner would run, so the certificate they deliver is vacuous. What is missing is a complexity measure that adapts to the data and to the model's effective capacity rather than its raw size.

Contribution. We give a uniform-convergence bound for equivalence testing of parametric probabilistic models in which the complexity term is the *local effective dimension* d_{eff} of Abbas et al. (2021), a Fisher-information–based quantity that counts the parameter directions resolvable at sample size n, rather than the raw dimension d. Given a candidate θ_1^* and a reference θ_{ref} whose soft predictions agree on every one of n i.i.d. test inputs drawn from a deployment distribution $\mathcal{D}$, the bound certifies that the expected total-variation distance between their predictive distributions on a fresh input from $\mathcal{D}$ is $O(\varepsilon\sqrt{\ln n/n})$, where ε controls a tradeoff between the error rate and the confidence level. Specialized to sample complexity, the bound exhibits a dichotomy. When the candidate is genuinely close to the reference in parameter space, the sample complexity is $\widetilde{\Theta}(d_{\text{eff}}/\eta^2)$ at target error η, with d_{eff} replacing the parameter dimension that a VC-style analysis would deliver. When the candidate is parametrically distant, d_{eff} drops out and the rate reverts to the Lipschitz-driven $\widetilde{\Theta}(M_1^2\varepsilon^2/\eta^2)$. The dichotomy is

[1]Independent, Zürich, Switzerland [2]Thayer School of Engineering, Dartmouth College, Hanover, NH, USA [3]Independent, Rennes, France [4]School of Computer Science, University of Birmingham, UK [5]Zeroth Research, UK. Correspondence to: Marc Kaufmann <researchcorrespondence@proton.me>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

informative rather than incidental: d_{eff} is the right complexity notion precisely in the near-equivalence regime that motivates the test in the first place.

Hypothesis-testing reading. Equivalence testing inverts the usual default of null-hypothesis significance testing: the hypothesis of interest is that the two models are close, not that they differ. Writing $R(\theta) := \mathbb{E}_{x\sim\mathcal{D}}[\,\mathrm{TV}(p(\cdot \mid x;\theta), p(\cdot \mid x;\theta_{\text{ref}}))\,]$ for the expected predictive total-variation distance from the reference, this hypothesis is $H_1 : R(\theta_1^*) \le \eta$, and our guarantee is a one-sided certificate of H_1: conditional on zero disagreement on n i.i.d. inputs, H_1 holds at confidence $1-\delta$ provided n exceeds the threshold of Corollary 5. The structure is analogous to two one-sided tests (TOST) for distributional equivalence (Schuirmann, 1987; Wellek, 2010), but with two changes that matter for modern ML: the certificate is uniform over an ε-ball of candidate parameters around the reference, and its complexity term is d_{eff} rather than a worst-case combinatorial quantity.

Outline. Section 2 situates our result within the ecosystem of classical uniform-convergence bounds, traces back the roots of the notion of effective dimension, and give a concise introduction to the equivalence-testing literature. Section 3 fixes notation, recalls the local effective dimension, and defines the soft-disagreement loss. Section 4 states the uniform-convergence theorem and the equivalence-testing corollary. Section 5 derives the sample complexity and exhibits the dichotomy of parameter-close vs. parameter-distant regimes. Section 6 works out the setting of logistic regression, verifying all hypotheses and computing concrete constants. Section 7 discusses scope, looseness, and extensions. Full proofs and some comments on a possible martingale-difference extension to non-i.i.d. inputs are deferred to the appendix.

2. Related Work

Classical uniform convergence and equivalence. The standard route to a guarantee of the form (4) is uniform convergence over a hypothesis class with controlled combinatorial complexity. For binary classifiers, the VC dimension yields a bound of order $\sqrt{(d_{\text{VC}} + \log(1/\beta))/n}$ on the deviation between empirical and population disagreement (Vapnik & Chervonenkis, 1971; Blumer et al., 1989). Specialized to equivalence testing, observing zero empirical disagreement on n i.i.d. inputs from a class with VC dimension d_{VC} certifies population disagreement at most $O(\sqrt{(d_{\text{VC}} + \log(1/\beta))/n})$ (Blumer et al., 1989). For neural networks, $d_{\text{VC}} = \widetilde{\Theta}(W)$ in the parameter count W (Bartlett et al., 2019), which makes the bound loose for over-parametrized models that are nonetheless empirically well-behaved. Rademacher and PAC-Bayesian analyses (Bartlett & Mendelson, 2002; Neyshabur et al., 2018) replace the worst-case dimension with norm-based or distribution-dependent quantities, and have produced substantially tighter bounds for deep networks. Our result flows across the same conceptual vein — a uniform-convergence bound where the complexity term is data-aware — but the data dependence enters through the Fisher information rather than through margins or weight norms.

Effective dimension. The local effective dimension d_{eff} of Abbas et al. (2021) is a Fisher-information–based capacity measure introduced to quantify and compare the expressive power of quantum and classical parametrized families. The motivation in that work is twofold: providing a data-dependent generalization bound that scales with the rank structure of the Fisher information rather than the raw parameter count, and giving a spectral diagnostic for trainability that connects to the barren-plateau phenomenon in variational quantum circuits. The construction is built from the geometry of the model manifold under the Fisher metric and is closely related to earlier information-geometric capacity notions, including the MDL stochastic complexity of Rissanen (1996) and Myung et al.'s geometric complexity (Myung et al., 2000; Balasubramanian, 1997). In the limit, all of these concentrate on the same object: the volume of the model manifold under the Fisher metric, with d_{eff} effectively counting parameter directions whose Fisher eigenvalues exceed a resolution scale set by the sample size. We use d_{eff} as a black-box ingredient in our proof: the only property we need is the covering-number bound of Lemma 7 of Abbas et al. (2021), which is what allows us to replace the d-dimensional volume term in a standard covering argument by $\kappa^{d_{\text{eff}}/2}$. Our contribution bootstraps the soft-disagreement uniform convergence statement and derives from it sample complexity bounds for equivalence-testing.

Equivalence testing of distributions and models. Equivalence testing in statistics traditionally refers to the problem of certifying that two distributions are within a tolerance, rather than rejecting equality. The two one-sided tests (TOST) procedure is the standard parametric solution (Schuirmann, 1987) and the closeness-testing literature in theoretical computer science studies the analogous problem in the discrete-distribution setting with optimal sample complexity bounds depending on the support size and tolerance (Batu et al., 2013; Daskalakis et al., 2018b; Canonne, 2020). These problems take two distributions as the objects of interest; our problem takes two *conditional* distributions parametrized by θ and asks for a guarantee in expectation over the input distribution. The closest analog in the model-comparison literature is hypothesis testing of model fit (Lloyd & Ghahramani, 2015; Kim et al., 2021), which compares two models against data via two-sample tests on

their predictions; our setting is the degenerate one-sample case in which both models are evaluated on the same inputs and the question is whether they agree. In the property-testing literature, identity testing for parametric families with structural assumptions has been studied for graphical models (Daskalakis et al., 2018a) and for classes admitting efficient covers; effective-dimension–based bounds offer an analog for generic differentiable parametric families.

Behavioral and black-box model verification. A separate line of work asks similar equivalence questions from a verification standpoint, often without parametric assumptions. For programs and finite automata, equivalence is decidable under structural constraints (Hopcroft et al., 2006; Angluin, 1987; Chow, 1978; Isberner et al., 2014; Merten et al., 2009); for polynomials over finite fields, Schwartz–Zippel-style identity testing yields exponentially decaying soundness with the number of evaluations (Schwartz, 1980; Zippel, 1979; DeMillo & Lipton, 1978). For neural networks, while theoretical bounds for exact equivalence are intractable, existing approaches rely either on white-box parameter inspection or on coverage-driven testing without distributional guarantees. The present paper is complementary: we trade exact equivalence for a distributional guarantee — two models that agree on a sample are close in expected predictive TV distance — and quantify the sample size needed via d_{eff}.

3. Setup and Notation

We fix an input space $\mathcal{X}$, a finite output space $\mathcal{Y}$ with $|\mathcal{Y}| = K$, and a probability distribution $\mathcal{D}$ on $\mathcal{X}$ that is independent of the parameter θ. The parameter space is $\Theta = [-1, 1]^d$ and we work with a parameterized family $\mathcal{M}_\Theta = \{p(\cdot \mid \cdot; \theta) : \theta \in \Theta\}$ of conditional distributions on $\mathcal{Y}$ given $x \in \mathcal{X}$. Test inputs $x_1, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{D}$ are drawn from a held-out set or known input distribution; we discuss the i.i.d. assumption further in Section 7 and Appendix B.

Because $\mathcal{D}$ is independent of θ, the joint and conditional Fisher information matrices coincide, and we work with the conditional form

$$F(\theta) := \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y|x;\theta}\big[\nabla_\theta \log p(y \mid x; \theta)\, \nabla_\theta \log p(y \mid x; \theta)^\top\big]. \tag{1}$$

For $\bar\theta \in \Theta$ and $\varepsilon > 0$ we write $B_\varepsilon(\bar\theta) := \{\theta \in \Theta : \|\theta - \bar\theta\| \le \varepsilon\}$ for the closed Euclidean ball and V_ε for its Lebesgue volume.

Definition 1 (Normalized Fisher information)**.** The *normalized Fisher information* on $B_\varepsilon(\bar\theta)$ is

$$\bar F(\theta) \;:=\; \frac{d}{V_\varepsilon \int_{B_\varepsilon(\bar\theta)} \operatorname{tr} F(\theta')\, d\theta'}\, F(\theta),$$

so that $V_\varepsilon^{-1} \int_{B_\varepsilon(\bar\theta)} \operatorname{tr} \bar F(\theta)\, d\theta = d$.

Definition 2 (Local effective dimension)**.** With $\kappa_{n,\gamma} := \gamma n/(2\pi \ln n)$ and $\gamma \in (2\pi \ln n/n,\, 1]$, the local effective dimension on $B_\varepsilon(\bar\theta)$ is

$$d_{\text{eff}} \;:=\; \frac{2\,\ln\!\Big(\frac{1}{V_\varepsilon}\int_{B_\varepsilon(\bar\theta)} \sqrt{\det\big(I_d + \kappa_{n,\gamma}\bar F(\theta)\big)}\, d\theta\Big)}{\ln \kappa_{n,\gamma}}.$$

Approximating the integrand by its value at $\bar\theta$, which is tight when $\bar F$ varies slowly across the ball, gives $d_{\text{eff}} \approx (\ln \kappa_{n,\gamma})^{-1} \sum_{i=1}^d \ln(1 + \kappa_{n,\gamma}\bar\lambda_i(\bar\theta))$ where $\bar\lambda_i$ are the eigenvalues of $\bar F(\bar\theta)$. Eigenvalues much larger than $1/\kappa_{n,\gamma}$ contribute ≈ 1 each ("active" parameters), and eigenvalues much smaller contribute ≈ 0. Hence d_{eff} is a soft count of parameter directions that matter at the resolution scale $1/\kappa_{n,\gamma}$.

The role of γ. The hyperparameter γ controls the resolution scale $\kappa_{n,\gamma}$ at which the effective dimension is measured. Smaller γ means coarser resolution (fewer eigenvalues counted as "active"), hence smaller d_{eff} and better confidence, but a worse error bound (since $\kappa_{n,\gamma}$ is smaller). Larger γ means finer resolution, larger d_{eff}, and a tighter error bound at the cost of weaker confidence.

Soft prediction disagreement. Fix a reference $\theta_{\text{ref}} \in \Theta$. The *pointwise soft disagreement* between $M(\theta)$ and $M(\theta_{\text{ref}})$ at input x is the total-variation distance

$$h_\theta(x) \;:=\; \tfrac{1}{2}\sum_{y\in\mathcal{Y}} \big|p(y \mid x; \theta) - p(y \mid x; \theta_{\text{ref}})\big|. \tag{2}$$

The *population* and *empirical* disagreements are

$$R(\theta) := \mathbb{E}_{x\sim\mathcal{D}}[h_\theta(x)], \qquad R_n(\theta) := \frac{1}{n}\sum_{i=1}^n h_\theta(x_i). \tag{3}$$

A standard calculation (Lemma 10 in the appendix) shows $h_\theta(x) \in [0, 1]$, that h_θ is Lipschitz in θ inheriting the same constant as the model class, and that $h_\theta(x_1), \ldots, h_\theta(x_n)$ are i.i.d. for fixed θ.

4. Main Result

We state the uniform-convergence theorem for soft prediction disagreement and the equivalence-testing corollary it implies. Both follow from a covering argument paired with the Fisher-based covering-number bound of Abbas et al. (2021); the full proof is in Appendix D.

Theorem 3 (Uniform convergence for soft disagreement)**.** *Let $d, n \in \mathbb{N}$ with $n \ge 19$. Let $\Theta = [-1, 1]^d$, $\bar\theta \in \Theta$, $\varepsilon \in (1/\sqrt{n},\, 2\sqrt{d}]$, $\gamma \in (2\pi \ln n/n,\, 1]$, $M_1 > 0$, $\Lambda \ge 0$. Suppose*

1. the Model is Lipschitz: $\sup_x \frac{1}{2}\sum_y |p(y \mid x; \theta_1) - p(y \mid x; \theta_2)| \le M_1\|\theta_1 - \theta_2\|$ *for all* $\theta_1, \theta_2 \in \Theta$.

2. its Fisher information matrix has full rank: $\bar{F}(\theta) \succ 0$ *for every* $\theta \in B_\varepsilon(\bar{\theta})$.

3. Bounded log-Fisher operator: $\sup_{\theta \in B_\varepsilon(\bar{\theta})} \|\nabla_\theta \log \bar{F}(\theta)\|_{\mathrm{op}} \le \Lambda$.

Draw $x_1, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{D}$. *Then, for every* $\theta_1 \in B_\varepsilon(\bar{\theta})$, *with probability at least* $1-\beta$,

$$R(\theta_1) \;\le\; R_n(\theta_1) \;+\; \frac{4M_1\varepsilon}{\sqrt{\kappa_{n,\gamma}}}, \tag{4}$$

where $\kappa_{n,\gamma} = \gamma n/(2\pi \ln n)$ *and*

$$\beta = 2\, c_d\, (1+\varepsilon\Lambda)^d\, \kappa_{n,\gamma}^{d_{\mathrm{eff}}/2}\, \exp\!\Big(-\tfrac{16\pi M_1^2 \varepsilon^2 \ln n}{\gamma}\Big), \tag{5}$$

with $c_d > 0$ *a constant depending only on* d *and* $d_{\mathrm{eff}} = d_{n,\gamma,\varepsilon}(\mathcal{M}_{B_\varepsilon(\bar{\theta})})$.

Proof sketch. A standard application of Hoeffding's inequality plus a covering argument (Lemma 6 of Abbas et al. (2021)) gives, for any $\xi > 0$,

$$\Pr\Big(\sup_{\theta \in B_\varepsilon(\bar{\theta})} |R(\theta) - R_n(\theta)| \ge \xi\Big) \le 2\, N\!\Big(\tfrac{\xi}{4M_1}\Big) \exp\!\Big(-\tfrac{n\xi^2}{2}\Big),$$

using that h_θ is bounded in $[0,1]$, M_1-Lipschitz in θ, and i.i.d. across samples for fixed θ. The Fisher-based covering bound (Lemma 7 of Abbas et al. (2021)) controls the cover at scale $\varepsilon/\sqrt{\kappa_{n,\gamma}}$ by $c_d(1+\varepsilon\Lambda)^d \kappa_{n,\gamma}^{d_{\mathrm{eff}}/2}$. Choosing $\xi = 4M_1\varepsilon/\sqrt{\kappa_{n,\gamma}}$ matches the two scales, and using $n/\kappa_{n,\gamma} = 2\pi \ln n/\gamma$ in the exponent yields β as in (5). The full argument is in Appendix D. □

The deviation $4M_1\varepsilon/\sqrt{\kappa_{n,\gamma}} \sim \varepsilon\sqrt{\ln n/n}$ does not depend on d_{eff}. The complexity d_{eff} enters only through the confidence β.

Equivalence testing. Specializing to two models that agree on every test input gives the guarantee that motivates this paper. Suppose $\theta_1^*, \theta_{\mathrm{ref}} \in \Theta$ satisfy

$$p(y \mid x_i; \theta_1^*) = p(y \mid x_i; \theta_{\mathrm{ref}}) \qquad \forall y,\ i = 1, \ldots, n, \tag{6}$$

i.e. the candidate and reference produce identical soft predictions on every drawn input. Then $h_{\theta_1^*}(x_i) = 0$ for all i and $R_n(\theta_1^*) = 0$.

Corollary 4 (Equivalence-testing guarantee)**.** *Under the assumptions of Theorem 3 with* $\bar{\theta} = \theta_{\mathrm{ref}}$, *and with* $\varepsilon \ge \|\theta_1^* - \theta_{\mathrm{ref}}\|$, *suppose* (6) *holds. Then with probability at least* $1-\beta$,

$$\boxed{R(\theta_1^*) \;\le\; \frac{4M_1\varepsilon}{\sqrt{\kappa_{n,\gamma}}}} \tag{7}$$

where β *is given by* (5).

Proof. $\theta_1^* \in B_\varepsilon(\theta_{\mathrm{ref}})$ by assumption and $R_n(\theta_1^*) = 0$ by (6). Apply Theorem 3. □

The bound certifies that the expected predictive total-variation distance under $\mathcal{D}$ is at most $\eta := 4M_1\varepsilon/\sqrt{\kappa_{n,\gamma}}$. Section 5 translates this into the corresponding sample-complexity statement. By Markov's inequality, $\Pr_{x\sim\mathcal{D}}[h_{\theta_1^*}(x) > t] \le \eta/t$ for every $t > 0$. The translation to hard-decision (argmax) decoding requires either a coupling argument under sampling decoding, in which case the disagreement probability is bounded by $2R(\theta_1^*)$, or a margin assumption under argmax decoding; we discuss both in Appendix C (Remark 11).

Distributional, not pointwise, equivalence. The corollary provides a *distributional* guarantee: it bounds $\mathbb{E}_{x\sim\mathcal{D}}[h_{\theta_1^*}(x)]$, not $h_{\theta_1^*}(x)$ for every x, and not $\theta_1^* = \theta_{\mathrm{ref}}$. Two models can disagree on a set $S \subset \mathcal{X}$ of positive measure as long as either $\mathcal{D}(S)$ is small (the disagreement region is rare under the data distribution) or the disagreement on S is mild. Pointwise equivalence is strictly stronger and is not what the corollary provides; parameter equivalence is stronger still, and our bound is useful precisely when we do not know whether the parameters are identical.

False-positive control. A "false positive" — incorrectly certifying equivalence — occurs when $R(\theta_1^*) > \eta$ but the bound declares otherwise; this happens with probability at most β. Reducing β by increasing n or adjusting ε controls the false-positive rate. The bound does not directly control false negatives (failing to certify two models that are in fact equivalent): a conservative bound may declare the models non-equivalent simply because n is insufficient. By contrast, hypothetical pointwise equivalence testing has no false positives but requires evaluating on the entire input space, which is infeasible. Our distributional approach trades exact pointwise checking for a probabilistic guarantee that holds on average under $\mathcal{D}$.

5. Sample Complexity

We now derive the number n of i.i.d. test inputs needed to certify $R(\theta_1^*) \le \eta$ at confidence $1-\delta$. The bound has two ingredients to control simultaneously: the deviation $4M_1\varepsilon/\sqrt{\kappa_{n,\gamma}} \le \eta$ and the failure probability $\beta \le \delta$. Both are required for a meaningful guarantee: a tiny error bound with $\beta = 0.99$ is useless, as is $\beta = 10^{-100}$ paired with an error bound of 1. The error bound controls *how small* the population disagreement is on the high-probability event; the failure probability controls *how likely* that event is.

The deviation condition rearranges to $\gamma n/(2\pi \ln n) \;\ge$

$16M_1^2\varepsilon^2/\eta^2$, with clean sufficient form

$$n \geq \frac{32\pi M_1^2\varepsilon^2}{\gamma\eta^2}\ \ln n. \tag{8}$$

The technical constraint $\varepsilon > 1/\sqrt{n}$ from Theorem 3 is automatically absorbed by (8) in the regime where the error bound is non-vacuous: substituting $\varepsilon = 1/\sqrt{n}$ into (8) gives a quadratic constraint on n that is implied by the same bound for any larger ε.

Taking logarithms of (5) and asking $\beta \leq \delta$ gives

$$\frac{16\pi M_1^2\varepsilon^2 \ln n}{\gamma} \ \geq\ \frac{d_{\text{eff}}}{2}\ln\kappa_{n,\gamma} + d\ln(1+\varepsilon\Lambda) + \ln\frac{2c_d}{\delta}. \tag{9}$$

For large n, $\ln\kappa_{n,\gamma} \sim \ln n$, so the dominant requirement is $16\pi M_1^2\varepsilon^2/\gamma > d_{\text{eff}}/2$, i.e. $\varepsilon \geq \varepsilon_{\min}$ with

$$\varepsilon_{\min}\ :=\ \sqrt{\frac{\gamma\, d_{\text{eff}}}{32\pi M_1^2}}. \tag{10}$$

The quantity $\varepsilon_{\min}$ is the minimum ball radius for which the exponential suppression $\exp(-C\varepsilon^2\ln n)$ in β is strong enough to overcome the polynomial growth $\kappa^{d_{\text{eff}}/2}$ of the covering number, ensuring $\beta \to 0$ as $n \to \infty$.

Choosing ε. ε is a free parameter of Corollary 4, and the two terms it enters pull in opposite directions. The deviation $4M_1\varepsilon/\sqrt{\kappa_{n,\gamma}}$ in (7) grows linearly in ε, so smaller ε tightens the guarantee. The exponent in β, however, scales as $-\varepsilon^2\ln n$, while the polynomial covering term grows only as $(1+\varepsilon\Lambda)^d\,\kappa_{n,\gamma}^{d_{\text{eff}}/2}$, so larger ε tightens the confidence. The feasible range is $\varepsilon \in [\max(\|\theta_1^* - \theta_{\text{ref}}\|, \varepsilon_{\min}, 1/\sqrt{n}),\ 2\sqrt{d}]$, and within it the choice that minimizes the deviation while keeping $\beta \leq \delta$ is

$$\varepsilon^* := \max\big(\|\theta_1^* - \theta_{\text{ref}}\|,\ \varepsilon_{\min}\big),$$

under which Corollary 4 delivers $R(\theta_1^*) \leq 4M_1\varepsilon^*/\sqrt{\kappa_{n,\gamma}}$ with probability at least $1-\delta$.

Corollary 5 (Sample complexity)**.** *Under the hypotheses of Corollary 4, taking $\varepsilon = \varepsilon^*$, the guarantee $R(\theta_1^*) \leq \eta$ holds at confidence $\geq 1-\delta$ provided*

$$n\ \geq\ \frac{32\pi M_1^2(\varepsilon^*)^2}{\gamma\eta^2}\ln n \quad \textit{and (9) at } \varepsilon = \varepsilon^*. \tag{11}$$

The two cases:

1. ***Parameter-close regime.*** *If $\|\theta_1^* - \theta_{\text{ref}}\| \leq \varepsilon_{\min}$ (in particular if $\theta_1^* = \theta_{\text{ref}}$), then $\varepsilon^* = \varepsilon_{\min}$ and (11) simplifies to*

$$n\ =\ \widetilde{\Theta}\Big(\frac{d_{\text{eff}}}{\eta^2}\Big). \tag{12}$$

Explicitly, $n = \Theta\big(\frac{d_{\text{eff}}}{\eta^2}\ \ln\frac{d_{\text{eff}}}{\eta^2}\big)$ suffices.

2. ***Parameter-distant regime.*** *If $\|\theta_1^* - \theta_{\text{ref}}\| > \varepsilon_{\min}$, then $\varepsilon^* = \|\theta_1^* - \theta_{\text{ref}}\|$ and*

$$n\ =\ \widetilde{\Theta}\Big(\frac{M_1^2\,\|\theta_1^* - \theta_{\text{ref}}\|^2}{\gamma\eta^2}\Big). \tag{13}$$

The proof is in Appendix E. This dichotomy is the central point: in the parameter-close regime, the cover scale $\sim 1/\sqrt{\kappa}$ is the binding constraint and d_{eff} enters linearly; in the parameter-distant regime, the cover scale $\sim\varepsilon$ is set by the candidate's deviation and the covering term $(1+\varepsilon\Lambda)^d$ is absorbed into the constant. The data-aware advantage of d_{eff} is a near-equivalence phenomenon: when the two models are nearly identical, the relevant complexity is the intrinsic dimension of the family near $\bar{\theta}$; when they are far apart, the dominant effect is the Lipschitz-bounded deviation.

6. Worked Example: Binary Logistic Regression

We instantiate the theorem on a setting where all hypotheses can be verified in closed form. Let $\mathcal{X} = \{x \in \mathbb{R}^p : \|x\| \leq R\}$, $\mathcal{Y} = \{0,1\}$, $d = p+1$, $\theta = (w,b)$, and $p(1 \mid x;\theta) = \sigma(w^\top x + b)$ with σ the sigmoid.

The Lipschitz condition 1 holds with $M_1 = (R+1)/4$, since $|\sigma(w_1^\top x + b_1) - \sigma(w_2^\top x + b_2)| \leq \frac{1}{4}|(w_1 - w_2)^\top x + (b_1 - b_2)| \leq \frac{R+1}{4}\|\theta_1 - \theta_2\|$ using $\sigma' \leq 1/4$ and $\sqrt{R^2+1} \leq R+1$. The full-rank Fisher condition 2 and the bounded log-Fisher operator condition 3 are verified in Appendix F; the latter holds with $\Lambda = 2(R+1)/c_*$ where $c_* = \sigma(R+1)(1-\sigma(R+1))$.

Numerical example. Take $p = 10$ (so $d = 11$), $R = 5$, $\gamma = 0.01$, and suppose two logistic regressions agree on $n = 50{,}000$ test points. Then $M_1 = 1.5$ and $\kappa_{n,\gamma} = 500/(2\pi\cdot 10.82) \approx 7.35$. If the local effective dimension is $d_{\text{eff}} = 5$ (five of eleven directions are "active" at scale $1/\kappa \approx 0.14$), then $\varepsilon_{\min} = \sqrt{0.01\cdot 5/(32\pi\cdot 1.5^2)} \approx 0.0149$ and the deviation bound is $4M_1\varepsilon_{\min}/\sqrt{\kappa_{n,\gamma}} \approx 0.033$. The exponent in β evaluates to about 27.2, giving $e^{-27.2} \approx 1.5\times 10^{-12}$; the polynomial prefactor $\kappa^{d_{\text{eff}}/2} = 7.35^{2.5} \approx 146$ and $(1+\varepsilon_{\min}\Lambda)^d$ remain bounded, so $\beta \ll 10^{-9}$. For a target $\eta = 0.05$ at $\delta = 10^{-2}$, Corollary 5(1) gives $n \geq (d_{\text{eff}}/\eta^2)\ln(d_{\text{eff}}/\eta^2) \approx 1.5\times 10^4$, i.e. about 15,000 i.i.d. test inputs.

7. Discussion

What the bound certifies, and what it does not. Corollary 4 bounds the expected predictive total-variation distance under $\mathcal{D}$. This is a statement about distributional equivalence at the level of soft outputs and inherits two limitations. First, it is not pointwise: by Markov's inequality, the fraction of inputs on which the two predictive distributions differ

by more than t is bounded by η/t, but no input is individually certified. Second, the bound is on soft disagreement and translating it to hard disagreement requires a coupling argument under sampling decoding (free) or a margin assumption under argmax decoding (Remark 11). Neither of these is a defect of the bound; they reflect what TV distance does and does not control.

Why the absolute constants are pessimistic. The bound carries two sources of looseness inherited from the covering construction. The Lipschitz constant M_1 is worst-case, and for natural-data inputs the typical Lipschitz behavior $\mathbb{E}_x[\|\nabla_\theta h_\theta(x)\|^2]^{1/2}$ is often orders of magnitude smaller than the worst-case $\sup_x \|\nabla_\theta h_\theta(x)\|$ that enters our analysis. The covering term $(1+\varepsilon\Lambda)^d$ grows exponentially in the parameter dimension and dominates the bound for deep networks unless $\varepsilon\Lambda \ll 1/d$. Both terms are an artifact of the covering proof and might be replaceable by sharper quantities under a chaining or Rademacher-based argument. Importantly, neither of these terms involves d_{eff}: they are parameter-dimension constants that bound the *regularity* of the model class, while d_{eff} controls the *intrinsic complexity*. Empirically, we expect d_{eff} to track the sample-complexity scaling much more tightly than the parameter dimension, and verifying this gap experimentally is a natural next step.

Comparison with VC-based equivalence testing. The classical VC analysis of binary-classifier equivalence yields $\Pr_{x\sim\mathcal{D}}[f_{\theta_1^*}(x) \neq f_{\theta_{\text{ref}}}(x)] = O\big(\sqrt{(d_{\text{VC}} + \ln(1/\delta))/n}\big)$ under zero empirical disagreement. Our bound differs in three ways. First, what is being measured: the VC bound controls hard-classification disagreement, while (7) controls expected soft TV disagreement; the soft quantity is finer-grained, since two models can agree on every hard label while having different confidence levels, or have small soft disagreement while occasionally disagreeing on hard predictions near a decision boundary. The two are related (Remark 11 in Appendix C) but not equivalent. Second, the complexity measure: d_{VC} is a worst-case property of the hypothesis class, independent of data, while d_{eff} depends on the Fisher information and incorporates the data distribution and the trained parameter $\bar{\theta}$. For overparametrized models it may well occur that $d_{\text{eff}} \ll d \ll d_{\text{VC}}$. Third, where the complexity enters: d_{VC} appears in the error rate, while d_{eff} enters only the failure probability β and not the deviation. Both bounds have the same $O(1/\sqrt{n})$ asymptotic rate up to log factors.

The unobservable parameter distance. The bound depends on $\varepsilon \geq \|\theta_1^* - \theta_{\text{ref}}\|$, which a black-box verifier does not know. We see two practical takes on this. The first is to read the corollary as a conditional certificate: "if the candidate lies within ε of the reference, then with confidence $1-\delta$ its expected TV disagreement is at most η." This makes transparent what black-box agreement on a sample does and does not establish, and is appropriate when there are additional constraints — training history, deployment context, partial information about model parameters — that bounds ε a priori. The second is to take the worst case $\varepsilon = \text{diam}(\Theta) = 2\sqrt{d}$; the bound remains non-vacuous whenever case (b) of Corollary 5 dominates, because then d_{eff} rather than ε controls the rate. Neither perspective removes the dependence on ε; they can only clarify how the bound is to be used.

When d_{eff} helps and when it does not. The dichotomy in Corollary 5 sharpens what the effective-dimension framework buys. In the parameter-close regime — the regime that matters for near-equivalence claims about fine-tuned, quantized, or random-seed–perturbed models — the rate is $\tilde{\Theta}(d_{\text{eff}}/\eta^2)$ and d_{eff} takes over the role classically played by the VC or parameter dimension. In the parameter-distant regime, d_{eff} drops out: the test is driven by the worst-case Lipschitz constant of the family and the parameter-space deviation, neither of which is data-aware. This is consistent with the intuition behind effective dimension: the geometry of the model manifold near a fixed point is informative for tasks that probe local behavior, and the ε-ball equivalence test is exactly such a task.

Beyond i.i.d. inputs. The i.i.d. assumption rules out adaptive testing and per-token verification of language models, where consecutive tokens are dependent. Appendix G give some ideas on how to extend the bound to martingale-difference deviations via Azuma–Hoeffding, which suffices for autoregressive verification under mild boundedness assumptions. The price is a slightly worse constant in the deviation; the d_{eff}-dependence is unchanged. Covariate shift, where the test distribution differs from the training or commitment distribution, is a more serious extension and would require a separate density-ratio analysis; we leave this to future work.

Future directions. Three directions seem most promising. First, replacing the covering argument with a chaining or Rademacher analysis might remove the $(1+\varepsilon\Lambda)^d$ factor and yield bounds that are less loose on realistic deep architectures. Second, the bound is currently stated for soft-prediction disagreement in TV distance; a parallel analysis for KL or for f-divergences in general would let the test compose with downstream calibration or distillation guarantees. Third, empirically verifying the d_{eff}-vs.–query-budget scaling on quantized-vs.–full-precision and fine-tuned-vs.–base model pairs would test the parameter-close regime and is a natural follow-up to this theoretical analysis.

References

Abbas, A., Sutter, D., Zoufal, C., Lucchi, A., Figalli, A., and Woerner, S. The power of quantum neural networks. *Nature Computational Science*, 1(6):403–409, 2021.

Angluin, D. Learning regular sets from queries and counterexamples. *Information and Computation*, 75(2):87–106, 1987.

Balasubramanian, V. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation*, 9(2):349–368, 1997.

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Batu, T., Fortnow, L., Rubinfeld, R., Smith, W. D., and White, P. Testing closeness of discrete distributions. *Journal of the ACM*, 60(1):1–25, 2013.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. Learnability and the Vapnik–Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

Canonne, C. L. A survey on distribution testing: your data is big, but is it blue? *Theory of Computing*, Graduate Surveys, 9:1–100, 2020.

Chow, T. S. Testing software design modeled by finite-state machines. *IEEE Transactions on Software Engineering*, SE-4(3):178–187, 1978.

Daskalakis, C., Dikkala, N., and Kamath, G. Concentration of multilinear functions of the Ising model with applications to network data. In *Advances in Neural Information Processing Systems*, 2017.

Daskalakis, C., Diakonikolas, I., Servedio, R. A, Valiant, G. and Valiant P. Testing k-modal distributions: optimal algorithms via reductions. In *SODA*, 2013.

DeMillo, R. A. and Lipton, R. J. A probabilistic remark on algebraic program testing. *Information Processing Letters*, 7(4):193–195, 1978.

Hopcroft, J. E., Motwani, R., and Ullman, J. D. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, 3rd edition, 2006.

Isberner, M., Howar, F., and Steffen, B. The TTT algorithm: a redundancy-free approach to active automata learning. In *Runtime Verification*, volume 8734 of *LNCS*, pp. 307–322. Springer, 2014.

Kim, I., Ramdas, A., Singh, A., and Wasserman, L. Classification accuracy as a proxy for two-sample testing. *Annals of Statistics*, 49(1):411–434, 2021.

Lloyd, J. R. and Ghahramani, Z. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*, 2015.

Merten, M., Steffen, B., Howar, F., and Margaria, T. Next generation LearnLib. In *TACAS*, volume 5505 of *LNCS*, pp. 220–223. Springer, 2011.

Myung, I. J., Balasubramanian, V., and Pitt, M. A. Counting probability distributions: differential geometry and model selection. *PNAS*, 97(21):11170–11175, 2000.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Rissanen, J. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, 1996.

Schuirmann, D. J. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

Schwartz, J. T. Fast probabilistic algorithms for verification of polynomial identities. *Journal of the ACM*, 27(4):701–717, 1980.

Vapnik, V. N. and Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

Wellek, S. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. CRC Press, 2nd edition, 2010.

Zippel, R. Probabilistic algorithms for sparse polynomials. In *EUROSAM*, volume 72 of *LNCS*, pp. 216–226. Springer, 1979.

A. Hoeffding's Inequality

We use the following standard concentration inequality.

Lemma 6 (Hoeffding's inequality)**.** *Let $Z_1, \ldots, Z_n$ be independent random variables with $Z_i \in [a_i, b_i]$ almost surely. Let $\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$ and $\mu = \mathbb{E}[\bar{Z}]$. Then for any $t > 0$,*

$$\Pr(|\bar{Z} - \mu| \geq t) \;\leq\; 2\exp\Big(-\tfrac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big).$$

In the i.i.d. case with $Z_i \in [0, B]$ this simplifies to $\Pr(|\bar{Z} - \mu| \geq t) \leq 2\exp(-2nt^2/B^2)$.

In our application (proof of Theorem 3, Step 1), we apply this with $Z_i = h_\theta(x_i) \in [0,1]$, $\bar{Z} = R_n(\theta)$, $\mu = R(\theta)$, and $B = 1$, giving $\Pr(|R_n(\theta) - R(\theta)| \geq t) \leq 2\exp(-2nt^2)$ at each fixed θ.

B. Setup and Notation

We restate the setup with all remarks for completeness.

Joint vs. conditional Fisher. Because $\mathcal{D}$ is independent of θ, $\nabla_\theta \log p(x, y; \theta) = \nabla_\theta \log p(y \mid x; \theta)$ and the joint and conditional Fisher information matrices coincide. We use the conditional form throughout (equation (1) in the body).

Remark 7 (Local vs. global). We use only the *local* effective dimension (Definition 4 in Abbas et al. (2021)), because Theorem 3 concerns $\theta \in B_\varepsilon(\bar{\theta})$. Locality is what makes d_{eff} depend on the trained model $\bar{\theta}$ and hence on the training data, in contrast with the global definition.

Remark 8 (I.i.d. assumption). Theorem 3 requires $x_1, \ldots, x_n \overset{\text{iid}}{\sim} \mathcal{D}$. This is satisfied when test inputs are drawn from a held-out test set or a known input distribution. It is *not* satisfied for adaptive testing, autoregressive token-level testing of language models (consecutive tokens are dependent), or under covariate shift. Appendix G sketches a martingale-difference extension.

Remark 9 (Same-architecture restriction). Both models must belong to the same parameterized family $\mathcal{M}_\Theta$. The bound applies to comparing a model before and after fine-tuning, before and after quantization, or two same-architecture training runs that differ only in random seed. It does not apply across architectures (different parameter spaces).

C. Soft Prediction Disagreement

Lemma 10 (Properties of h_θ)**.** *Suppose*

$$\sup_{x \in \mathcal{X}} \frac{1}{2} \sum_{y \in \mathcal{Y}} \big| p(y \mid x; \theta_1) - p(y \mid x; \theta_2) \big| \;\leq\; M_1 \|\theta_1 - \theta_2\| \tag{14}$$

for all $\theta_1, \theta_2 \in \Theta$. Then

1. *$h_\theta(x) \in [0,1]$ for every $\theta \in \Theta$ and $x \in \mathcal{X}$;*
2. *$|h_{\theta_1}(x) - h_{\theta_2}(x)| \leq M_1 \|\theta_1 - \theta_2\|$ for every $x \in \mathcal{X}$ and $\theta_1, \theta_2 \in \Theta$;*
3. *for fixed $\theta \in \Theta$, $h_\theta(x_1), \ldots, h_\theta(x_n)$ are i.i.d.*

Proof. **(a)**: $h_\theta(x) \leq \frac{1}{2}\sum_y (p(y \mid x; \theta) + p(y \mid x; \theta_{\text{ref}})) = 1$ and $h_\theta(x) \geq 0$ trivially.

(b): For any θ_1, θ_2,

$$\begin{aligned}
&|h_{\theta_1}(x) - h_{\theta_2}(x)| \\
&= \tfrac{1}{2}\Big|\textstyle\sum_y \big(|p(y \mid x; \theta_1) - p(y \mid x; \theta_{\text{ref}})| \\
&\quad - |p(y \mid x; \theta_2) - p(y \mid x; \theta_{\text{ref}})|\big)\Big| \\
&\leq \tfrac{1}{2} \sum_y \big|\, |p(y \mid x; \theta_1) - p(y \mid x; \theta_{\text{ref}})| \\
&\quad - |p(y \mid x; \theta_2) - p(y \mid x; \theta_{\text{ref}})| \,\big| \\
&\leq \tfrac{1}{2} \sum_y \big|p(y \mid x; \theta_1) - p(y \mid x; \theta_2)\big| \\
&\leq M_1 \|\theta_1 - \theta_2\|,
\end{aligned}$$

using triangle inequality, the reverse triangle inequality $\big||a| - |b|\big| \leq |a - b|$, and (14).

(c): $h_\theta(x_i)$ is a deterministic function of x_i alone (since $\theta, \theta_{\text{ref}}$ are fixed); the i.i.d. property of the x_i transfers. □

Remark 11 (Sampling vs. argmax decoding). *Sampling decoding.* If the served model emits $Y \sim p(\cdot \mid x; \theta)$, the probability that two models disagree on the sampled token at input x under the maximal coupling is bounded pointwise by $\Pr_{\text{couple}}[Y_1 \neq Y_{\text{ref}} \mid x] \leq 2h_\theta(x)$, hence in expectation by $2R(\theta_1^*) \leq 8M_1\varepsilon/\sqrt{\kappa_{n,\gamma}}$. This is the standard total-variation–coupling identity.

Argmax decoding, binary case. Let $f_\theta(x) := \mathbf{1}[p(1 \mid x; \theta) \geq 1/2]$ be the deterministic argmax classifier. For $\mathcal{Y} = \{0, 1\}$ the pointwise inequality

$$\mathbf{1}[f_{\theta_1^*}(x) \neq f_{\theta_{\text{ref}}}(x)] \leq 2\, h_{\theta_1^*}(x)$$

holds for every $x \in \mathcal{X}$: when the argmaxes differ, one model assigns $p(1 \mid x) \geq 1/2$ and the other $p(1 \mid x) < 1/2$, so $h_{\theta_1^*}(x) = |p(1 \mid x; \theta_1^*) - p(1 \mid x; \theta_{\text{ref}})| \geq 1/2$ and the deterministic factor of 2 closes the gap; when the argmaxes agree, the inequality is $0 \leq 2h_{\theta_1^*}(x)$. Taking expectations under $\mathcal{D}$ and applying Corollary 4,

$$\Pr_{x \sim \mathcal{D}}[f_{\theta_1^*}(x) \neq f_{\theta_{\text{ref}}}(x)] \;\leq\; 2\, R(\theta_1^*) \;\leq\; \tfrac{8M_1\varepsilon}{\sqrt{\kappa_{n,\gamma}}}.$$

The Lipschitz condition is applied only to the soft disagreement h_θ, never to the discontinuous indicator $\mathbf{1}[f_{\theta_1^*} \neq f_{\theta_{\text{ref}}}]$; the hard-disagreement bound follows from the deterministic pointwise inequality after taking expectations. *Argmax decoding, $K \geq 3$.* The factor-of-2 argument above relies on the binary structure of the argmax, where the two predictions are forced to lie on opposite sides of the threshold $1/2$. For $K \geq 3$ this fails: the argmax can flip between two non-top classes without any one probability crossing a fixed threshold. Concretely, with $K = 3$ and $p_{\theta_1^*}(\cdot \mid x) = (0.34, 0.33, 0.33)$, $p_{\theta_{\text{ref}}}(\cdot \mid x) = (0.33, 0.34, 0.33)$, the argmaxes differ but $h_{\theta_1^*}(x) = 0.01$, so no constant prefactor can yield a hard-disagreement bound from h alone. A

margin assumption recovers a quantitative bound: if there exists $\mu > 0$ such that

$$\Pr_{x\sim\mathcal{D}}\big[\, p_{\theta_{\mathrm{ref}}}(y^{(1)} \mid x) - p_{\theta_{\mathrm{ref}}}(y^{(2)} \mid x) < \mu \,\big] \le \rho,$$

where $y^{(1)}, y^{(2)}$ are the top-1 and top-2 labels of the reference at x, then on the high-margin event a flip in argmax forces $h(x) \ge \mu/2$, and Markov's inequality yields $\Pr_{x\sim\mathcal{D}}[f_{\theta_1^*}(x) \ne f_{\theta_{\mathrm{ref}}}(x)] \le \rho + 2R(\theta_1^*)/\mu$.

D. Proof of the Main Theorem

Executive summary. The result combines two lemmas of Abbas et al. (2021): a Hoeffding + covering uniform-convergence bound at scale ξ, and a Fisher-based covering-number bound that replaces the d-dimensional volume term with $\kappa^{d_{\mathrm{eff}}/2}$. Choosing ξ to match the cover scale balances the two; the resulting deviation is $\Theta(M_1\varepsilon/\sqrt{\kappa})$ and the failure probability decays exponentially in $M_1^2\varepsilon^2 \ln n/\gamma$.

Proof of Theorem 3. **Step 1 (Covering argument + Hoeffding's inequality).** We invoke Lemma 6 of Abbas et al. (2021): for any per-sample loss $\ell(\theta; x)$ that is M-Lipschitz in θ uniformly in x, bounded in $[0, B]$, and i.i.d. across samples for fixed θ,

$$\Pr\left(\sup_{\theta\in B_\varepsilon(\bar\theta)} |R(\theta) - R_n(\theta)| \ge \xi\right) \le 2\, N_{B_\varepsilon(\bar\theta)}\left(\frac{\xi}{4M}\right) \exp\left(-\frac{n\xi^2}{2B^2}\right). \tag{15}$$

where $N_S(r)$ is the r-covering number of S in Euclidean norm.

Sketch of (15). Let $\{\theta_1, \dots, \theta_N\}$ be the centers of an r-covering of $B_\varepsilon(\bar\theta)$ with $r = \xi/(4M)$ and $N = N_{B_\varepsilon(\bar\theta)}(r)$. For any $\theta \in B_\varepsilon(\bar\theta)$ there exists θ_j with $\|\theta - \theta_j\| \le r$; the Lipschitz property gives $|\ell(\theta; x) - \ell(\theta_j; x)| \le Mr = \xi/4$ for all x, hence $|R(\theta) - R(\theta_j)| \le \xi/4$ and similarly $|R_n(\theta) - R_n(\theta_j)| \le \xi/4$. Combining, $|(R(\theta) - R_n(\theta)) - (R(\theta_j) - R_n(\theta_j))| \le \xi/2$, so $|R(\theta) - R_n(\theta)| \ge \xi$ implies $|R(\theta_j) - R_n(\theta_j)| \ge \xi/2$. A union bound over the N centers, combined with Hoeffding's inequality (Lemma 6) at each center, yields (15).

Applicability to h_θ. Set $\ell = h_\theta$. By Lemma 10, the loss is bounded by $B = 1$, is M_1-Lipschitz, and is i.i.d. for fixed θ. Hence

$$\Pr\left(\sup_{\theta\in B_\varepsilon(\bar\theta)} |R(\theta) - R_n(\theta)| \ge \xi\right) \le 2\, N_{B_\varepsilon(\bar\theta)}\left(\frac{\xi}{4M_1}\right) \exp\left(-\frac{n\xi^2}{2}\right). \tag{16}$$

Step 2 (Fisher cover bound). Lemma 7 of Abbas et al. (2021), under conditions 2–3, controls the cover when rescaled to scale $\varepsilon/\sqrt{\kappa}$:

$$N_{B_\varepsilon(\bar\theta)}\Big(\tfrac{\varepsilon}{\sqrt{\kappa_{n,\gamma}}}\Big) \;\le\; c_d\,(1+\varepsilon\Lambda)^d\,\kappa_{n,\gamma}^{d_{\mathrm{eff}}/2}. \tag{17}$$

This is a purely geometric statement about $B_\varepsilon(\bar\theta)$ under the Fisher metric: it does not involve h_θ or any loss function, only conditions 2–3.

Step 3 (Combination of lemmas). Choose $\xi = 4M_1\varepsilon/\sqrt{\kappa_{n,\gamma}}$, so that $\xi/(4M_1) = \varepsilon/\sqrt{\kappa_{n,\gamma}}$ matches (17). Substituting into (16) and using $n/\kappa_{n,\gamma} = 2\pi \ln n/\gamma$,

$$\Pr\Big(\sup_{\theta\in B_\varepsilon(\bar\theta)} |R(\theta) - R_n(\theta)| \ge \tfrac{4M_1\varepsilon}{\sqrt{\kappa_{n,\gamma}}}\Big) \le 2\, c_d\,(1+\varepsilon\Lambda)^d\,\kappa_{n,\gamma}^{d_{\mathrm{eff}}/2}\,\exp\Big(-\tfrac{16\pi M_1^2\varepsilon^2 \ln n}{\gamma}\Big) = \beta.$$

Since $\theta_1 \in B_\varepsilon(\bar\theta)$, on the event

$$\mathcal{E} := \Big\{\sup_{\theta\in B_\varepsilon(\bar\theta)} |R(\theta) - R_n(\theta)| < \tfrac{4M_1\varepsilon}{\sqrt{\kappa_{n,\gamma}}}\Big\},$$

which has $\Pr(\mathcal{E}) \ge 1 - \beta$, the supremum bound implies $|R(\theta_1) - R_n(\theta_1)| < 4M_1\varepsilon/\sqrt{\kappa_{n,\gamma}}$. □

E. Proof of the Sample-Complexity Corollary

Proof of Corollary 5. **Deviation:** substitute $\varepsilon = \varepsilon^*$ in (8), giving $n \ge (32\pi M_1^2(\varepsilon^*)^2/(\gamma\eta^2)) \ln n$.

Confidence: for $\varepsilon \ge \varepsilon_{\min}$ the leading term of (9) is satisfied; the subdominant terms $d\ln(1+\varepsilon\Lambda) + \ln(2c_d/\delta)$ are $O(1)$ in n and absorbed by the $\ln n$-factor on the left for n at least the threshold above.

Case 1. Substitute $\varepsilon^* = \varepsilon_{\min} = \sqrt{\gamma d_{\mathrm{eff}}/(32\pi M_1^2)}$ into (11); the M_1 and γ factors cancel, leaving $n \ge (d_{\mathrm{eff}}/\eta^2)\ln n$. The asymptotic $\widetilde\Theta(d_{\mathrm{eff}}/\eta^2)$ follows from $n \ge C\ln n \Rightarrow n = \Theta(C\ln C)$ with $C = d_{\mathrm{eff}}/\eta^2$.

Case 2. Substitute $\varepsilon^* = \|\theta_1^* - \theta_{\mathrm{ref}}\|$ into (11), giving $n \ge (32\pi M_1^2\|\theta_1^* - \theta_{\mathrm{ref}}\|^2/(\gamma\eta^2)) \ln n$, hence the stated $\widetilde\Theta$. □

Auxiliary remarks.

Remark 12 (The unobservable $\|\theta_1^* - \theta_{\mathrm{ref}}\|$). The verifier has black-box access only and does not know $\|\theta_1^* - \theta_{\mathrm{ref}}\|$. Two practical reading variants: **(i) Hypothesis form:** treat ε as a hypothesis, reading the corollary as "conditional on the candidate lying within ε of the reference." **(ii) Worst case:** take $\varepsilon = \mathrm{diam}(\Theta) = 2\sqrt{d}$ and accept the resulting bound; this remains useful whenever case 1 of Corollary 5 dominates (since then d_{eff} rather than ε governs the rate).

Remark 13 (When the bound is vacuous). The error bound exceeds 1 (the threshold where the bound becomes vacuous) iff $n \le 32\pi M_1^2\varepsilon^2 \ln n/\gamma$. The confidence β exceeds 1 (vacuous) iff $\varepsilon < \varepsilon_{\min}$ defined in (10). Both conditions tighten with increasing n.

Remark 14 (Why the empirical scaling can be much better than the bound). Two sources of looseness make the absolute constants pessimistic for deep networks without diminishing the value of d_{eff} as a complexity measure. **(L1)** M_1 is a worst-case Lipschitz constant; for natural-data inputs the typical Lipschitz behavior $\mathbb{E}_x[\|\nabla_\theta h(x)\|^2]^{1/2}$ can be orders of magnitude smaller. **(L2)** $(1+\varepsilon\Lambda)^d$ grows exponentially in d. For deep nets this term dominates unless $\varepsilon\Lambda \ll 1/d$. Both come from the covering argument and are absent from the empirical d_{eff}-vs.–query-budget scaling.

F. Verification of Hypotheses for Logistic Regression

We complete the verification of the three hypotheses of Theorem 3 for the binary logistic-regression family of Section 6.

F.1. Lipschitz condition (1)

For any $\theta_1 = (w_1, b_1)$, $\theta_2 = (w_2, b_2)$ and $\|x\| \le R$,

$$\begin{aligned}
&|\sigma(w_1^\top x + b_1) - \sigma(w_2^\top x + b_2)| \\
&\le \tfrac{1}{4}|(w_1 - w_2)^\top x + (b_1 - b_2)| \\
&\le \tfrac{1}{4}(R\|w_1 - w_2\| + |b_1 - b_2|) \\
&\le \tfrac{R+1}{4}\|\theta_1 - \theta_2\|,
\end{aligned}$$

using $\sigma' \le 1/4$ and $\sqrt{R^2+1} \le R+1$. Hence $M_1 = (R+1)/4$.

F.2. Full-rank Fisher (2) and bounded log-Fisher operator (3)

Let $\tilde{x} = (x^\top, 1)^\top \in \mathbb{R}^{p+1}$ and $z = w^\top x + b$. The score is $\nabla_\theta \log p(y \mid x; \theta) = (y - \sigma(z))\tilde{x}$, so

$$F(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\big[\sigma(z)(1-\sigma(z))\,\tilde{x}\tilde{x}^\top\big].$$

Full rank. For $\theta \in \operatorname{int}(\Theta)$ and $\|x\| \le R$, $|z| \le R+1$, so $\sigma'(z) \ge \sigma(R+1)(1-\sigma(R+1)) =: c_* > 0$. Whenever the data distribution makes $\{\tilde{x} : x \in \operatorname{supp}(\mathcal{D})\}$ span $\mathbb{R}^{p+1}$ — a generic-position assumption that holds e.g. if $\mathcal{D}$ has a density on an open set of $\mathbb{R}^p$ — $v^\top F(\theta) v \ge c_* \mathbb{E}_x[(v^\top \tilde{x})^2] > 0$ for any $v \ne 0$.

Bounded log-Fisher operator. Differentiating F in θ, $\partial_\theta F = \mathbb{E}_x[\partial_\theta(\sigma'(z))\,\tilde{x}\tilde{x}^\top]$ with $\partial_\theta \sigma'(z) = \sigma'(z)(1 - 2\sigma(z))\,\tilde{x}$. Operator norms are bounded by $\|F^{-1}\partial_\theta F\|_{\text{op}} \le 2(R+1)/c_*$ on Θ, so we may take $\Lambda = 2(R+1)/c_*$. (See Lemma 7 of Abbas et al. (2021) for the exact transfer to $\nabla_\theta \log \bar{F}$.)

G. Extension to Non-i.i.d. Inputs

The i.i.d. assumption in Theorem 3 is used in Step 1 of the proof, specifically through Hoeffding's inequality applied to the sum $\sum_i h_\theta(x_i)$. Replacing Hoeffding with Azuma–Hoeffding should the result to martingale-difference sequences, which covers several practically important non-i.i.d. settings, most notably those settings of autoregressive token-level verification of language models in which consecutive tokens are dependent but the per-token disagreements $h_\theta(x_t) - \mathbb{E}[h_\theta(x_t) \mid x_{<t}]$ form a bounded martingale difference sequence. Working out the constants and the precise filtration is beyond the scope of this paper; we view it as a natural follow-up.