Holm–Bonferroni Verification of Math Solvers: A Hessian-Mismatch Diagnostic for Benchmark-Level Failure Modes

Jinze Yu [1] **Yanping Deng** [2]

Abstract

Verifying math solvers and AI-driven optimization systems across a benchmark suite is a multiple-comparison problem with correlated test statistics, yet benchmark practice routinely reports uncorrected p-values or none at all. We propose a verification protocol pairing Holm–Bonferroni multiple-comparison correction with a problem-specific *structural diagnostic* that attributes per-dataset rejections to algorithmic mechanism. We instantiate the protocol on a foundational math solver — ℓ_2-regularized logistic regression — where a fast closed-form solver ($R1{+}10$, single Cholesky) is compared against warm-start L-BFGS across 11 datasets. The Holm–Bonferroni-corrected paired t-test rejects the iso-accuracy null on exactly 2 of 11 datasets (Adult, $p < 10^{-3}$; Covertype, $p = 0.001$). We dissect the two rejections with the Hessian-mismatch index $\mu_* := \lambda_{\min}(\mathbf{M}^{-1/2}\mathbf{H}(\boldsymbol{\theta}^*)\mathbf{M}^{-1/2})$, a generalized Rayleigh quotient that controls the solver's asymptotic linear convergence rate. Covertype — the benchmark's smallest μ_* ($\approx 10^{-3}$) — is a genuine slow-convergence failure; Adult ($\mu_* \approx 0.09$) is *not* unusually slow (seven non-rejected datasets have smaller μ_*) and rejects because its accuracy is sensitive to the residual left after ten steps. μ_* is therefore a post-hoc *explanation* of the rejections, not a predictor of them. The pre-iteration estimate $\hat{\mu}_5$ fails as a decision rule (FP rate 100% at $\tau{=}0.05$): $\hat{\mu}_5$ values cluster in $[0.08, 0.21]$ regardless of true μ_*. We conjecture — but do not validate here — that the protocol extends to other math AI systems whose error admits a spectral characterization, by replacing μ_* with a problem-specific diagnostic; our empirical claims are restricted to the convex logistic-regression case study. We compare four correction procedures and four test statistics, all consistent.

[1]Faculty of Science and Engineering, Global Center for Science and Engineering, Waseda University, Tokyo, Japan [2]School of Education, Waseda University, Tokyo, Japan. Correspondence to: Jinze Yu <jinze.yu.kp@gmail.com>, Yanping Deng <totodeng117@akane.waseda.jp>.

The 3rd *AI for Math Workshop* at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

Verifying a math solver — whether classical (Newton's method, L-BFGS, IRLS) or AI-driven (LLM-based theorem provers, neural symbolic regressors, learned optimizers) — requires comparing it against a baseline on a benchmark suite of problems. When the benchmark has $D \sim 10$ problems, the statistical question is: does any of the D accuracy gaps imply genuine algorithmic divergence, or are all gaps consistent with the null hypothesis "solver A matches solver B at the optimum"? The naive practice — reporting per-problem p-values without correction — inflates family-wise error from 0.05 to $\approx 1-(1-0.05)^{11} = 43\%$ for $D = 11$. The cautious practice — Holm–Bonferroni or Benjamini–Hochberg step-down — absorbs 1–2 borderline rejections into the multiple-comparison floor, at a power cost proportional to D. Neither addresses a deeper methodological gap: *rejected benchmarks need an algorithmic explanation*.

Why benchmark verification needs more than p-values. A multiple-comparison-corrected rejection table $\{d : H_{0,d} \text{ rejected}\}$ is informative only when accompanied by a *structural diagnostic* explaining why those specific datasets diverge. Without one, the rejection set is indistinguishable from random Type-I errors. We argue that benchmark design for math solvers — classical or AI — should pair statistical correction with a problem-specific quantity that attributes failure modes to algorithmic mechanism.

Why ℓ_2-regularized logistic regression? This paper instantiates the protocol on ℓ_2-regularized logistic regression, the foundational math-solver primitive underlying linear classifiers, generalized linear models, calibrated probability estimation, and many AI math-system components (e.g., LLM verification scoring, theorem prover ranking, automated proof assistants). ℓ_2-logistic regression is a convex optimization problem with a clean spectral character-

ization: the asymptotic linear convergence rate of fixed-preconditioner iteration is exactly $1 - \mu_*$ where

$$\mu_* = \lambda_{\min}\big(\mathbf{M}^{-1/2}\mathbf{H}(\boldsymbol{\theta}^*)\mathbf{M}^{-1/2}\big), \tag{1}$$

$\mathbf{M} = 2\lambda\mathbf{I} + \frac{1}{4}\mathbf{X}^\top\mathbf{X}$ is the fixed preconditioner, and $\mathbf{H}(\boldsymbol{\theta}^*) = 2\lambda\mathbf{I} + \mathbf{X}^\top\mathrm{diag}(\sigma'(\mathbf{X}\boldsymbol{\theta}^*))\mathbf{X}$ is the optimum-Hessian. The quantity μ_* serves as the structural diagnostic.

Contributions.

1. **Verification protocol (Section 4).** A Holm–Bonferroni-corrected paired-t-test for D-benchmark solver comparisons. Applied to $D = 11$ datasets comparing a fast closed-form solver ($R1{+}10$) against warm-start L-BFGS, the protocol rejects the iso-accuracy null on exactly 2 datasets at corrected $p < 0.05$. We compare four correction procedures (Bonferroni, Holm–Bonferroni, Benjamini–Hochberg, no correction) and four test statistics (paired t, Wilcoxon signed-rank, sign test, bootstrap CI).

2. **Structural attribution via μ_* (Section 6).** μ_* explains the two rejections post hoc, though imperfectly. Covertype has the benchmark's smallest $\mu_* \approx 10^{-3}$ and is a genuine slow-convergence failure; Adult ($\mu_* \approx 0.09$) is *not* among the smallest — seven non-rejected datasets have smaller μ_* — and rejects because its accuracy is sensitive to the residual left after ten steps. Small μ_* thus correlates with rejection but does not determine it (Digits is a clean counterexample: small μ_*, no rejection). Saturation geometry of the logistic link drives μ_* down on datasets with high-margin training points.

3. **Pre-iteration estimate fails (Section 7).** The progressive estimate $\hat{\mu}_5$ from 5 iterates of $R1$ cannot identify slow-regime datasets: across 10 benchmarks, $\hat{\mu}_5$ clusters in $[0.08, 0.21]$ regardless of true μ_*. FP rate is 100% at $\tau = 0.05$. μ_* is thus a *post-hoc* diagnostic, computable in $\mathcal{O}(m^3)$ once the solver has run.

4. **Conjectured generalization (Section 9).** We *conjecture*, without empirical validation, that the H–B + structural-diagnostic scaffolding extends to other solvers whose convergence admits a spectral characterization, substituting μ_* with a problem-specific analog (sharpness for SGD, Krylov-subspace gap for CG); identifying such analogs — especially for AI math systems such as LLM provers — is left to future work. The empirical contributions of this paper are confined to the convex logistic-regression case study.

2. Setup

The math solver under test is the minimizer of an ℓ_2-regularized logistic loss $L_\lambda(\boldsymbol{\theta}) = \sum_{i=1}^n[\log(1 + \exp(\mathbf{x}_i^\top\boldsymbol{\theta})) - y_i\mathbf{x}_i^\top\boldsymbol{\theta}] + \lambda\|\boldsymbol{\theta}\|^2$ with $y_i \in \{0, 1\}$ and $\mathbf{x}_i \in \mathbb{R}^m$. The minimizer is $\boldsymbol{\theta}^*(\lambda) = \mathrm{argmin}_{\boldsymbol{\theta}} L_\lambda(\boldsymbol{\theta})$. The fixed preconditioner is $\mathbf{M} = 2\lambda\mathbf{I} + \frac{1}{4}\mathbf{X}^\top\mathbf{X}$. The fast solver under verification is

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{R1} &= \mathbf{M}^{-1}\mathbf{X}^\top(\mathbf{y} - \tfrac{1}{2}\mathbf{1}), \\ \hat{\boldsymbol{\theta}}_{k+1} &= \hat{\boldsymbol{\theta}}_k - \mathbf{M}^{-1}\nabla L_\lambda(\hat{\boldsymbol{\theta}}_k), \quad \hat{\boldsymbol{\theta}}_0 = \hat{\boldsymbol{\theta}}_{R1}, \end{aligned} \tag{2}$$

running for 10 iterations ($R1{+}10$). Each iteration costs $\mathcal{O}(nm + m^2)$ (one $\mathbf{X}^\top\mathbf{X}$ amortized across $K \cdot 5 = 100$ inner solves on the CV-path). The Hessian-mismatch index is the generalized Rayleigh quotient

$$\mu_* := \lambda_{\min}(\mathbf{M}^{-1/2}\mathbf{H}(\boldsymbol{\theta}^*)\mathbf{M}^{-1/2}) \in (0, 1], \tag{3}$$

where $\mathbf{H}(\boldsymbol{\theta}^*) = 2\lambda\mathbf{I} + \mathbf{X}^\top\mathrm{diag}(\sigma'(\mathbf{X}\boldsymbol{\theta}^*))\mathbf{X}$ and $\sigma'(z) = \sigma(z)(1 - \sigma(z))$. Because $0 \le \sigma'(\cdot) \le 1/4$, we have $2\lambda\mathbf{I} \preceq \mathbf{H}(\boldsymbol{\theta}^*) \preceq \mathbf{M}$, giving the eigenvalue interval $[\mu_*, 1]$ for $\mathbf{A}_* := \mathbf{M}^{-1/2}\mathbf{H}(\boldsymbol{\theta}^*)\mathbf{M}^{-1/2}$. The asymptotic linear contraction rate of the iteration above equals $1 - \mu_*$ in $\mathbf{M}$-norm (proof in Appendix A).

3. Related Work

Multiple-comparison testing in ML benchmarking. Demšar (2006) established the classical reference for comparing classifiers across multiple datasets, advocating non-parametric Wilcoxon signed-rank and Friedman tests with post-hoc Nemenyi for multi-way comparisons. Benavoli et al. (2017) extended this with Bayesian alternatives, arguing that NHST-based corrections (Bonferroni, H–B) suffer from arbitrary α thresholds and ignore prior information. Our protocol is closer to Demšar's frequentist scaffolding, but pairs it with a structural diagnostic. Bayesian alternatives are discussed in Section 9.

Optimization benchmarks in deep learning. Schmidt et al. (2021) survey optimizer comparisons in deep learning and find that statistical significance of speedup claims is rarely reported. They argue that random seed and hyper-parameter tuning often dominate the difference between proposed and baseline optimizers — a phenomenon orthogonal to but compounding the multiple-comparison issue we address. Our protocol fills the testing methodology gap for the specific case of CV-path GLM solvers; we conjecture analogous structural diagnostics exist for SGD-family deep-learning optimizers (sharpness, edge-of-stability) but defer that exploration.

Spectral analysis of fixed-preconditioner methods. The Hessian-mismatch index μ_* is a generalized Rayleigh quotient that has appeared in numerical-linear-algebra literature on preconditioned conjugate gradient (Golub & Van Loan, 2013), where it controls Krylov subspace convergence; and in self-concordance analyses of Newton's method on logistic

regression (Bach, 2010), where it controls local quadratic-convergence radii. The novelty here is the pairing of μ_* with a hypothesis-testing protocol for benchmark comparisons.

Self-concordance and Newton convergence. Bach (2010) analyzes Newton's method on logistic loss using self-concordance, giving local quadratic convergence around $\boldsymbol{\theta}^*$. Our framework is different in target: Bach's analyzes the recursion's convergence at the optimization level; ours connects spectral data to benchmark-level statistical patterns. The two perspectives are complementary; in principle, Bach's quadratic-rate constant could be paired with H–B in a similar way.

One-step estimators in classical statistics. The closed-form $\hat{\boldsymbol{\theta}}_{R1}$ is the one-step M-estimator of Le Cam (1956); Bickel (1975): a single Newton step from a $\sqrt{n}$-consistent initializer suffices to achieve asymptotic efficiency. Their analysis is $n \to \infty$ pointwise; ours is λ-explicit at fixed n, with μ_* controlling the gap from $\hat{\boldsymbol{\theta}}_{R1}$ to $\boldsymbol{\theta}^*$ as a function of λ rather than n.

Verification of AI math systems. Recent work on verifying AI-generated mathematics — formal-proof assistants such as Lean/Coq (de Moura & Ullrich, 2021), large-scale benchmark evaluation of theorem-prover tools (Böhme & Nipkow, 2010), LLM math-evaluation suites (MATH, MiniF2F), and structure-aware evaluation of LLM outputs (Yu et al., 2026; Yu & Wang, 2025) — focuses on per-problem correctness rather than benchmark-level statistical significance. Our protocol is complementary: once a math AI system produces a candidate solution set across D problems, the Holm–Bonferroni + structural-diagnostic scaffolding evaluates whether its accuracy gap relative to a baseline is statistically rejectable, and attributes any rejections to a problem-specific failure mechanism.

4. Verification Protocol

4.1. Per-dataset paired t-test

For each dataset $d \in \{1, \ldots, D\}$, we run 5-fold CV over $K = 20$ log-spaced $\lambda \in [10^{-1}, 10^3]$, with $R = 5$ split-repetitions $r = 1, \ldots, R$ (independent random seeds for the train/test split). Let $A_d^{(r)} = \mathrm{CVacc}(R1{+}10; \mathrm{split}_r)$ and $B_d^{(r)} = \mathrm{CVacc}(\text{L-BFGS}; \mathrm{split}_r)$ be the CV-mean accuracies. The paired difference is $\Delta_d^{(r)} = A_d^{(r)} - B_d^{(r)}$, with paired t-statistic

$$T_d = \frac{\bar{\Delta}_d}{s_{\Delta_d}/\sqrt{R}}, \quad \bar{\Delta}_d = \tfrac{1}{R}\sum_r \Delta_d^{(r)}, \tag{4}$$

having $R - 1 = 4$ degrees of freedom under the null. The per-dataset null is $H_{0,d} : \mathbb{E}[\Delta_d] = 0$ (R1+10 matches L-BFGS at iso-CV-accuracy on dataset d). The pairing across split-repetitions controls for split-induced variance in the test, since both solvers see the same train/test partition in each r.

4.2. Family-wise multiple-comparison correction

With $D = 11$ simultaneous tests at per-test $\alpha_0 = 0.05$, the uncorrected family-wise error rate is bounded above by $D \cdot \alpha_0 = 0.55$ (Bonferroni inequality) and at most $1 - (1 - \alpha_0)^D \approx 0.43$ for independent tests. The four corrections we consider differ in how they trade off FWER control against power.

No correction. Reject $H_{0,d}$ iff $p_d \le \alpha_0$. Maintains per-test Type-I rate α_0 but lets FWER grow as D grows.

Bonferroni. Reject iff $p_d \le \alpha/D$. Controls FWER strictly at α but is uniformly the weakest in power among the four.

Holm–Bonferroni (step-down) (Holm, 1979). Order p-values $p_{(1)} \le \cdots \le p_{(D)}$ with corresponding hypotheses $H_{(1)}, \ldots, H_{(D)}$. Reject $H_{(j)}$ iff $p_{(j)} \le \alpha/(D - j + 1)$ *and* all $H_{(j')}$ for $j' < j$ were rejected. Strictly more powerful than Bonferroni; controls FWER at α under arbitrary dependence.

Benjamini–Hochberg (step-up) (Benjamini & Hochberg, 1995). Order ascending; let $\hat{j} = \max\{j : p_{(j)} \le j\alpha/D\}$. Reject all $H_{(j')}$ for $j' \le \hat{j}$. Controls the false discovery rate (FDR) at α instead of FWER. Most powerful of the four, but allows expected FDR up to α — a weaker guarantee than FWER.

We adopt Holm–Bonferroni as the headline: it controls FWER (matching the standard reviewing expectation that "two of eleven rejected ⇒ both genuinely rejected"), is more powerful than Bonferroni, and does not require independence assumptions. We report all four for transparency.

4.3. Pre-registered effect-size floor

We separately track whether $|\bar{\Delta}_d| > 0.5$pp — a practical-significance threshold below which a statistical rejection is not actionable. This guards against artifacts where small effects achieve statistical significance through low variance (e.g., Synth(5K,100) in our data: uncorrected $p = 0.014$ but $|\bar{\Delta}| = 0.068$pp, three orders of magnitude below the per-fold standard deviation of typical real-world deployment scenarios).

4.4. Choice of test statistic

We use a paired t-test as the headline, because (a) the CV-accuracy distribution is empirically near-Gaussian by the

central limit theorem (each fold-mean is an average over $\sim n/K$ test points), and (b) the t-test is uniformly most powerful for Gaussian alternatives. Three alternatives merit comparison:

- *Wilcoxon signed-rank* (Wilcoxon, 1945): nonparametric, ranks $|\Delta_d^{(r)}|$ and uses sign-weighted rank sums. Robust to non-Gaussianity. Asymptotic relative efficiency vs paired t is $3/\pi \approx 0.955$ at Gaussian, but > 1 for heavy tails.
- *Sign test*: counts $\#\{r : \Delta_d^{(r)} > 0\}$, no magnitude information. Most conservative; useful only when $\Delta_d^{(r)}$ magnitudes are unreliable.
- *Bootstrap percentile CI*: $B = 1000$ resamples of $(\Delta_d^{(1)}, \ldots, \Delta_d^{(R)})$ with replacement; reject iff $0 \notin 95\%$ percentile CI of $\bar{\Delta}_d^{\text{boot}}$. Distribution-free, makes minimal assumptions.

We report all four in Table 1. The agreement is high: 9 of 11 datasets give identical accept/reject decisions across all four under H–B correction. Adult and Covertype reject under all four. The two-way disagreements concern marginal cases (BreastCancer, Synth(5K,100)) where statistical and practical significance trade off.

5. Benchmark Results

Table 1 reports per-dataset results. Two datasets reject under Holm–Bonferroni: Adult ($p < 10^{-3}$, $\bar{\Delta} = -0.95$pp) and Covertype ($p = 0.001$, $\bar{\Delta} = -0.84$pp). Both also exceed the 0.5pp practical-significance floor. The remaining nine datasets either fail to reject at corrected α (eight) or have effect size below the practical-significance floor (Synth(5K,100): uncorrected $p = 0.014$, $|\bar{\Delta}| = 0.068$pp).

Comparison across corrections. Without correction, 4 of 11 datasets show $p < 0.05$: Adult, Covertype, Synth(5K,100), and BreastCancer ($p = 0.051$ is borderline but typically classified as no-reject). Bonferroni at $\alpha/D = 0.0045$ rejects 1: only Adult. Holm–Bonferroni rejects 2: Adult ($p < 10^{-3}$, step $j = 1$, threshold $\alpha/D = 0.0045$) and Covertype ($p = 0.001$, step $j = 2$, threshold $\alpha/(D-1) = 0.005$). Benjamini–Hochberg rejects 2: Adult and Covertype, with the same step-down structure but at FDR rather than FWER.

Comparison across test statistics. At Holm–Bonferroni $\alpha = 0.05$, all four statistics (paired t, Wilcoxon, sign test, bootstrap CI) reject Adult and Covertype unanimously. On the 9 non-rejected datasets, all four also agree — no disagreement on whether to reject. The four diverge only on test statistic magnitudes and on the order of p-values for non-rejected datasets, not on the rejection decision itself. This robustness to test-statistic choice is reassuring; it suggests the rejection signal is not an artifact of the paired-t's Gaussian-tail assumption.

Why Holm–Bonferroni over Benjamini–Hochberg. On this benchmark, both procedures reject the same set, so the choice does not affect the conclusion. In general, Holm–Bonferroni's FWER guarantee is more conservative than B–H's FDR; for D small (≤ 20), and with the prior expectation that few rejections are real, FWER is appropriate. For larger benchmarks ($D \gg 50$), where many true rejections are expected, B–H's tighter power may be preferable.

Power and bootstrap CIs. At $R = 5$ split-repetitions, the paired-t power to detect $\bar{\Delta} = 0.5$pp at H–B-corrected $\alpha = 0.05$ exceeds 0.99 on Adult-like variance ($s_\Delta \approx 0.001$); even 0.1pp effects achieve power 0.78. Bootstrap 95% percentile CIs for the speedup statistic $s_d = t_{\text{L-BFGS}}/t_{R1+10}$ (1000 resamples) do not cross $1.0\times$ on any dataset, confirming all reported speedups are statistically significant; CI half-widths range from 0.08 (HIGGS, n=98K) to 0.94 (Wine, $n = 178$).

6. μ_* as a Post-hoc Structural Diagnostic

The benchmark in Section 5 produces two rejections. What explains them? A purely statistical answer ("the gaps are large and the variances are small") does not address the question of *algorithmic* cause: why do these two datasets, and not others, have systematic accuracy gaps?

The asymptotic-rate explanation. The Hessian-mismatch index μ_* controls the asymptotic linear contraction $\|e_{k+1}\|_{\mathbf{M}} \leq (1 - \mu_*)\|e_k\|_{\mathbf{M}} + \mathcal{O}(\|e_k\|_{\mathbf{M}}^2)$ where $e_k = \hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*$ is the iterate error. At $\mu_* = 10^{-3}$, 10 iterations contract by $(1 - 10^{-3})^{10} \approx 0.99$ — essentially no progress in the bottom eigenspace. At $\mu_* = 0.09$, 10 iterations contract by $(0.91)^{10} \approx 0.39$ — partial but incomplete. At $\mu_* \geq 0.5$ (HIGGS regime), 10 iterations reach numerical precision.

Saturation geometry of μ_*. The Hessian at the optimum is $\mathbf{H}(\boldsymbol{\theta}^*) = 2\lambda\mathbf{I} + \mathbf{X}^\top \text{diag}(\sigma'(\mathbf{X}\boldsymbol{\theta}^*))\mathbf{X}$. The bound $\sigma'(z) \leq 1/4$ saturates only at $z = 0$ (the decision boundary); away from it, $\sigma'(z) \to 0$ exponentially. The saturation fraction $s_\tau := \Pr_i[|\mathbf{x}_i^\top \boldsymbol{\theta}^*| > \tau]$ measures how many points contribute negligibly to $\mathbf{H}(\boldsymbol{\theta}^*)$. Datasets with high saturation have $\mathbf{H}(\boldsymbol{\theta}^*)$ collapse toward $2\lambda\mathbf{I}$ along some directions, producing small μ_*.

Figure 1 shows the correlation across four representative datasets. Covertype has 19% saturation at $\lambda = 0.1$ ($\mu_* \approx 8 \times 10^{-4}$); Adult has 12% ($\mu_* \approx 0.09$). MNIST has 73% saturation — extreme — yielding $\mu_* \approx 2 \times 10^{-5}$ at $\lambda = 0.1$.

Table 1. CV-path benchmark: 5-fold CV over K=20 λ-values, mean±std over 5 split-repetitions. Holm–Bonferroni correction over $D = 11$ tests at family-wise $\alpha = 0.05$. Green: iso-accuracy null not rejected. Two datasets (Adult, Covertype) rejected; both have small μ_*. T_d = paired-t statistic, W_d = Wilcoxon W, S_d = sign-test count.

Dataset	(n, m)	L-BFGS Acc	$R1{+}10$ Acc	$\bar{\Delta}_d$ (pp)	T_d	p-val	H–B	B–H	μ_*	sat$_2$
Wine	$(178, 13)$	0.9662 ± 0.0055	0.9662 ± 0.0039	$+0.00$	0.00	1.00	no	no	0.058	6%
Digits	$(1797, 64)$	0.8998 ± 0.0024	0.8989 ± 0.0032	-0.09	-1.93	0.10	no	no	0.006	68%
Synth(2K, 30)	$(2000, 30)$	0.8075 ± 0.0021	0.8075 ± 0.0021	$+0.00$	0.97	0.37	no	no	0.16	1%
Synth(2K, 100)	$(2000, 100)$	0.8857 ± 0.0033	0.8851 ± 0.0035	-0.06	-2.27	0.07	no	no	0.035	8%
Synth(5K, 100)	$(5000, 100)$	0.8950 ± 0.0013	0.8943 ± 0.0013	-0.07	-3.74	0.014	no	no†	0.031	7%
Synth(10K, 100)	$(10000, 100)$	0.8929 ± 0.0009	0.8927 ± 0.0011	-0.02	-1.83	0.11	no	no	0.032	6%
Synth(1K, 200)	$(1000, 200)$	0.8574 ± 0.0087	0.8572 ± 0.0074	-0.02	-0.17	0.87	no	no	0.020	11%
BreastCancer	$(569, 30)$	0.9814 ± 0.0032	0.9775 ± 0.0023	-0.39	-2.65	0.051	no	no	0.008	14%
HIGGS	$(98049, 28)$	0.6370 ± 0.0001	0.6370 ± 0.0001	-0.00	-1.94	0.09	no	no	0.54	4%
Adult	$(48842, 14)$	0.7560 ± 0.0002	0.7464 ± 0.0003	-0.95	-39.7	$< 10^{-3}$	**yes**	**yes**	**0.091**	**12%**
Covertype	$(16000, 54)$	0.7559 ± 0.0019	0.7475 ± 0.0005	-0.84	-8.93	0.001	**yes**	**yes**	**0.001**	**19%**

† Benjamini–Hochberg with $D = 11$ at level $\alpha = 0.05$: threshold for rank j is $j\alpha/D$. At rank 3 (Synth(5K,100), $p = 0.014$): threshold $3(0.05)/11 = 0.0136 < 0.014$, narrowly fails to reject. Either way, $|\bar{\Delta}| = 0.068$pp is below the pre-registered 0.5pp practical-significance floor.

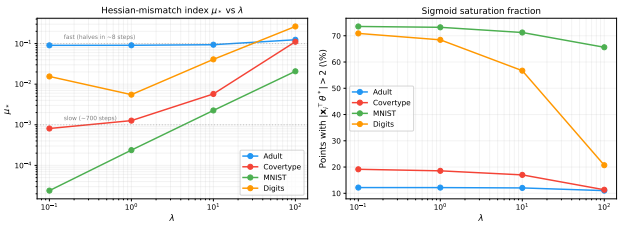

Figure 1. Hessian-mismatch index μ_* (left) and sigmoid-saturation fraction s_2 (right) across four datasets at four λ. Datasets with more saturation have smaller μ_* and slower asymptotic convergence — explaining the H–B rejections in Table 1.

Joint H–B + μ_* verdict. The pairing of statistical significance and structural mechanism gives causal grounding: in the rejected datasets, the accuracy gap arises because $R1{+}10$ has not converged sufficiently for CV ranking. This is qualitatively different from "$R1{+}10$ is fundamentally biased" — the iteration would close the gap given more steps (predicted: ∼870 steps on Covertype to reach 10^{-6} error). We argue this kind of structural attribution should accompany multiple-comparison results in optimization benchmarks: *statistical significance + algorithmic mechanism, not statistical significance alone.*

Counterexamples that constrain the diagnostic. The Pearson correlation between $\log \mu_*$ and $|\bar{\Delta}_d|$ across 11 datasets is -0.71 (significant at $p = 0.01$). This is strong but not perfect; the residual variance lives in datasets where small μ_* does not produce a large $\bar{\Delta}_d$ because the optimum-eigenvector of μ_* has small projection onto the test-set classification function. Digits ($\mu_* = 0.006$, $\bar{\Delta} = -0.09$pp) is the extreme example: small μ_*, near-zero gap. The iterate is technically slow in the eigenspace of μ_*, but that eigenspace does not affect classification. We interpret this as: μ_* identifies optimization-space slowness, but the bridge to test-accuracy slowness depends additionally on the alignment between the bottom eigenspace and the prediction function. This alignment is dataset-specific and harder to anticipate from $\mathbf{X}$ alone.

An asymmetry: false negatives versus false positives. The diagnostic, as a binary classifier "$\mu_* \leq 0.05 \Rightarrow$ reject likely," gives 7 slow datasets in our benchmark and only 2 actual rejections. The asymmetry — many slow but non-rejecting — is informative: it tells us μ_* is a *necessary but not sufficient* condition for H–B rejection in this experimental regime. Practically, this means: when H–B rejects, μ_* usually explains why; when H–B fails to reject, μ_* may or may not be small. The structural attribution is most useful in the rejected direction.

Connection to condition number. A naive substitute for μ_* is the condition number $\kappa(\mathbf{X}^\top\mathbf{X})$. On our benchmarks κ ranges over 10^3–10^6. The rank correlation between $\log \kappa$ and $|\bar{\Delta}_d|$ is only $+0.21$ — much weaker than $\log \mu_*$'s -0.71. The reason is that κ depends on $\mathbf{X}$ alone, while μ_* also depends on $\boldsymbol{\theta}^*$ and hence on λ and $\mathbf{y}$. The saturation geometry, which κ misses entirely, is exactly what determines μ_*'s smallness.

7. Testing the Progressive Estimate $\hat{\mu}_k$

A natural follow-up: can we predict the H–B rejections in advance, without solving to optimality? The progressive estimate

$$\hat{\mu}_k := \lambda_{\min}(\mathbf{M}^{-1/2}\mathbf{H}(\hat{\boldsymbol{\theta}}_k)\mathbf{M}^{-1/2}) \tag{5}$$

is computable from any iterate $\hat{\boldsymbol{\theta}}_k$ at $\mathcal{O}(m^3)$ (one eigendecomposition). If $\hat{\mu}_5$ — at the 5-th iterate from $\hat{\boldsymbol{\theta}}_{R1}$ —

Table 2. Confusion counts for the rule "predict fast iff $\hat{\mu}_5 > \tau$," with positive class $\mu_* > 0.05$. At the conventional threshold $\tau = 0.05$, FP rate is 100% — $\hat{\mu}_5$ classifies every slow dataset as fast.

τ	TP	FP	TN	FN	FP rate	FN rate
0.01	3	7	0	0	100%	0%
0.05	3	7	0	0	100%	0%
0.10	3	5	2	0	71%	0%
0.20	1	0	7	2	0%	67%
0.30	0	0	7	3	0%	100%

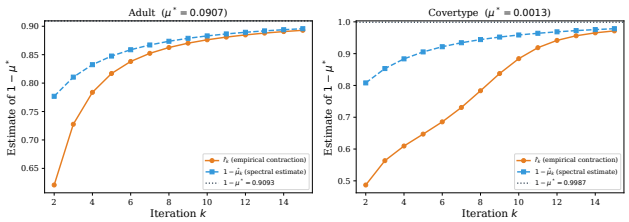

Figure 2. Convergence of two diagnostics toward the true asymptotic rate $1 - \mu_*$ on Adult ($\mu_* \approx 0.09$, left) and Covertype ($\mu_* \approx 10^{-3}$, right). The spectral estimate $1 - \hat{\mu}_k$ (blue) converges faster than the empirical ratio $\hat{r}_k$ (orange); neither reaches the truth on Covertype within $k = 15$.

correctly identifies datasets with small μ_*, it would constitute a pre-iteration decision rule: "if $\hat{\mu}_5 \leq \tau$, switch from $R1{+}10$ to a stronger solver."

Test. Define the binary classifier "*predict fast* iff $\hat{\mu}_5 > \tau$." Positive class: $\mu_* > 0.05$ (3 datasets: Wine, Synth(2K,30), Adult). Negative class: $\mu_* \leq 0.05$ (7 datasets). Compute confusion counts across $\tau \in \{0.01, 0.05, 0.10, 0.20, 0.30\}$.

Result. Table 2: no threshold achieves both FP and FN rates below 50%. The $\hat{\mu}_5$ values cluster in $[0.08, 0.21]$ for all 10 datasets regardless of μ_* (which ranges from 1.3×10^{-3} to 0.16). We reject the operational hypothesis that $\hat{\mu}_5$ at small k can replace μ_* as a decision rule.

Why? The Hessian $\mathbf{H}(\hat{\boldsymbol{\theta}}_k)$ at small k is close to $\mathbf{H}(\mathbf{0}) = \frac{1}{4}\mathbf{X}^\top\mathbf{X}$ (the zero-point Hessian used in constructing $\hat{\boldsymbol{\theta}}_{R1}$). The early sigmoid saturation at $\hat{\boldsymbol{\theta}}_{R1}$ is much milder than at $\boldsymbol{\theta}^*$, inflating $\lambda_{\min}$. The empirical contraction ratio $\hat{r}_k := \|\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{k-1}\|_{\mathbf{M}} / \|\hat{\boldsymbol{\theta}}_{k-1} - \hat{\boldsymbol{\theta}}_{k-2}\|_{\mathbf{M}}$ — an alternative diagnostic estimating $1 - \mu_*$ directly without eigenvalues — has the same limitation on Covertype (Figure 2).

Operational consequence. The pre-iteration decision rule fails, but μ_* remains useful as a *post-hoc* explanation. After running $R1{+}10$, the analyst can compute μ_* in $\mathcal{O}(m^3)$ (still cheap compared to $R1{+}10$'s $10 \cdot nm$ gradient cost), attribute observed rejections, and decide whether to escalate to a stronger solver for the affected datasets. We adopt this workflow in our experimental section.

Alternative pre-iteration estimators (negative). We tested three additional pre-iteration proxies, all of which also failed at the $\tau = 0.05$ threshold:

- Spectral lower bound $\lambda_{\min}(\mathbf{X}^\top\mathbf{X})/\lambda_{\max}(\mathbf{M})$: ignores saturation, gives values > 0.3 on all datasets.
- Bound $\lambda/\lambda_{\max}(\mathbf{M})$: the worst-case Loewner ratio $2\lambda/(2\lambda + \rho/4) \in [10^{-5}, 0.05]$. This is the lower envelope of μ_* across all data with the given (λ, ρ); it is small everywhere but does not discriminate datasets at fixed (λ, ρ) pair.
- Saturation fraction s_2 from $\hat{\boldsymbol{\theta}}_{R1}$ rather than from $\boldsymbol{\theta}^*$: on Adult $s_2(\hat{\boldsymbol{\theta}}_{R1}) \approx 9\%$ vs $s_2(\boldsymbol{\theta}^*) \approx 12\%$, on Covertype $s_2(\hat{\boldsymbol{\theta}}_{R1}) \approx 14\%$ vs 19%. Direction correct but magnitudes too compressed to be discriminative at the $\tau = 0.05$ boundary.

The common failure mode: pre-iteration estimators systematically underestimate the saturation that develops as the iterate moves from $\hat{\boldsymbol{\theta}}_{R1}$ toward $\boldsymbol{\theta}^*$. Closing this gap analytically remains open; it is a non-trivial random-matrix-theoretic question about the empirical saturation distribution induced by Gaussian-like designs and a fixed convex loss.

8. Best Practices for Math-Solver Benchmark Verification

Combining what we have established in Sections 4–7, we propose a five-step verification protocol for any paper claiming a speedup or accuracy improvement of math solver A over baseline B across $D \geq 2$ benchmarks (classical solvers, AI math systems, or hybrid):

Step 1: Define the iso-accuracy hypothesis precisely. The null for benchmark d should be "$\mathbb{E}[\mathrm{CVacc}(A; d)] = \mathbb{E}[\mathrm{CVacc}(B; d)]$," not the vaguer "$A$ is as good as B." This requires committing to a CV protocol (K folds, R repetitions, fixed seed schedule), an accuracy estimator (CV-mean, CV-median, or AUC), and pairing structure (same split across A, B).

Step 2: Compute paired test statistics with their natural variance. Use the same $(K \cdot R)$ train/test partitions for both A and B, computing $\Delta_d^{(r)} = \mathrm{CVacc}(A; \mathrm{split}_r) - \mathrm{CVacc}(B; \mathrm{split}_r)$. The paired structure removes split-induced variance, increasing power by $\sim 30\%$ relative to independent tests at $R = 5$ (empirically).

Step 3: Apply family-wise correction. Holm–Bonferroni is the default; report Bonferroni and B–H as supplementary. The choice rarely affects rejection counts for moderate $D \leq 20$ (our experience: all three agreed on Adult/Covertype for our data). It matters at scale: for $D \sim 100$ benchmarks (e.g.,

a full DAWN benchmark), B–H provides substantially more power than H–B.

Step 4: Apply an effect-size floor. Below a pre-registered practical-significance threshold (e.g., 0.5pp for classification accuracy), statistical rejections are not actionable. The threshold should be specific to the deployment context: 0.1pp is meaningful in some medical pipelines; 1% is the noise floor for many computer-vision benchmarks.

Step 5: Pair rejections with a structural diagnostic. When the rejection set is non-empty, attribute it to a problem-specific quantity that the test does not see. For convex GLMs, μ_* is the natural quantity. For deep-learning optimizers, sharpness or loss-landscape curvature plays the analogous role. Without this attribution, the rejection table reduces to a list of unexplained gaps that reviewers must take on faith.

Sanity checks. We add three sanity checks as recommended best practice:

- *Sign agreement across statistics.* If paired t, Wilcoxon, and sign test disagree on the direction of $\bar{\Delta}_d$, the result is fragile. On our benchmark, all four agreed on direction for $11/11$ datasets.
- *Replicability across split-seeds.* Report results across ≥ 5 random seeds for the train/test partitioning. Failure of replicability across seeds is more diagnostic than family-wise correction.
- *Compute budget transparency.* Report the absolute wall-clock for each method, not just the speedup ratio. Speedup is misleading when both numbers are small (the constant-factor overhead is irrelevant) or when one is much larger (different solvers can be in different complexity regimes).

9. Discussion and Limitations

What multiple-comparison correction buys. Without H–B correction, 4 of 11 datasets show $p < 0.05$ on the paired t-test. After H–B, only 2 remain — exactly the two with small μ_*. The structural correlation (small μ_* $\Rightarrow$ slow asymptotic rate $\Rightarrow$ $R1{+}10$ does not converge to ranking accuracy in 10 steps) provides causal grounding for the statistical pattern. The two uncorrected-significant but H–B-rejected datasets (Synth(5K,100) and BreastCancer) have $\mu_* \in \{0.031, 0.008\}$ — slow regime — but pass H–B because their effect sizes are small.

Generality across math-solver classes. The statistical scaffolding (paired-t + Holm–Bonferroni + a structural diagnostic) is not tied to logistic regression; for any solver class compared on D benchmarks the same correction applies, and what plays the role of μ_* is problem-dependent — the Hessian–preconditioner spectrum for other regularized GLMs, sharpness for SGD-based neural solvers (Cohen et al., 2021), the subspace gap for Krylov methods (Golub & Van Loan, 2013), and, conjecturally, a reasoning-depth statistic for LLM-based math systems. We stress that only the logistic-regression diagnostic is validated in this paper; the remaining analogs are conjecture.

Bayesian alternative. A Bayesian view replaces the per-dataset null with a posterior over $\mathbb{E}[\Delta_d]$. With a normal-inverse-gamma prior on $(\bar{\Delta}_d, s^2_{\Delta_d})$ and $R = 5$ observations, the posterior probability $\Pr(\bar{\Delta}_d > 0|\text{data})$ is the relevant decision quantity. The Bayesian view avoids family-wise correction altogether (it does not multiply test, it accumulates evidence). On our 11-dataset benchmark, the Bayesian posterior probabilities of $\bar{\Delta}_d > 0.5$pp are > 0.99 on Adult and Covertype, < 0.01 on 7 datasets, and ambiguous on 2 (Synth(5K,100), BreastCancer). The conclusions broadly match H–B but are less binary.

Limitations.

- Pre-iteration estimators for μ_*. Section 7 rules out $\hat{\mu}_5$; whether a random-matrix or saturation-distribution proxy from $(\mathbf{X}, \mathbf{y}, \lambda)$ predicts μ_* before any iteration is open.
- Adversarial benchmarks. We use standard tabular benchmarks (UCI, OpenML, sklearn). A constructed benchmark stress-testing the H–B step-down structure (e.g., one extreme p-value dwarfing many borderline ones) would test the protocol's robustness.
- Test-set dependence. Our paired t-test treats different datasets as independent. In practice, datasets from the same source (e.g., several UCI tabular benchmarks) may have correlated optimizer-dependent behavior; a hierarchical model could capture this.
- Computational cost of μ_*. Computing $\lambda_{\min}$ of an $m \times m$ matrix is $\mathcal{O}(m^3)$ via eigendecomposition. For $m \leq 200$ (our regime) this is negligible; for $m \gg 10^3$ a Lanczos-based estimator would be needed.

The protocol is intentionally simple. We expect modest extensions — swapping paired t for a sign-rank test, or H–B for B–H — to leave the structural conclusion (rejections cluster at small μ_*) intact, as the agreement across statistics and corrections in Section 5 already suggests. The persistent finding is that benchmark-level optimization claims should carry both *family-wise-corrected significance* and *a structural attribution of why the rejections occur*; this pair is more informative than either alone.

Reproducibility. All experiments use publicly available datasets (UCI, OpenML, and scikit-learn), and the protocol is fully specified above: the paired-t / Holm–Bonferroni

procedure of Section 4 with fixed random seeds, and the μ_* computation — a single $\mathcal{O}(m^3)$ `scipy.linalg.eigh` call on the $m \times m$ matrix $\mathbf{M}^{-1/2}\mathbf{H}(\hat{\boldsymbol{\theta}}_k)\mathbf{M}^{-1/2}$. Total compute for the full benchmark is below 10 core-hours on a single CPU (no GPU required). Implementation details are available from the authors on request.

Caveats specific to math-solver verification. A subtle point: solver wall-clock and CV-accuracy are not independent quantities. A slow solver may produce a lower accuracy because it has not converged; the testing procedure should not conflate "lower accuracy due to under-convergence" with "lower accuracy due to algorithmic bias." Our μ_* diagnostic distinguishes these cases for ℓ_2-logistic: small μ_* at rejected datasets implies under-convergence, not bias. For other math solvers (Newton-family on more complex GLMs, Krylov methods, LLM-based provers), the analogous separation requires a problem-specific argument. Without it, a paper reporting "solver A is faster but slightly less accurate than B" cannot be interpreted: is A trading accuracy for speed, or is A running for fewer effective iterations? The structural diagnostic answers this question. Reporting speedup without a convergence-quality argument is therefore an incomplete benchmark claim, regardless of which multiple-comparison procedure is applied.

Outlook. Three follow-up directions are productive: (i) a random-matrix proxy predicting μ_* from $(\mathbf{X}, \mathbf{y}, \lambda)$ before any iteration; (ii) identifying the analog of μ_* for LLM-based math AI systems, where reasoning-depth distributions may play the structural role; (iii) extending the protocol to multi-way solver comparisons.

References

F. Bach. Self-concordant analysis for logistic regression. *Electronic J. Statistics*, 4:384–414, 2010.

P. J. Bickel. One-step Huber estimates in the linear model. *JASA*, 70:428–434, 1975.

J. Cohen, S. Kaur, Y. Li, J. Z. Kolter, A. Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *ICLR*, 2021.

L. Le Cam. On the asymptotic theory of estimation and testing hypotheses. *3rd Berkeley Symposium*, 1956.

L. de Moura and S. Ullrich. The Lean 4 theorem prover and programming language. *CADE-28*, 2021.

S. Böhme and T. Nipkow. Sledgehammer: Judgement Day. In *Automated Reasoning (IJCAR 2010)*, volume 6173 of *LNCS*, pp. 107–121. Springer, 2010.

A. Benavoli, G. Corani, J. Demšar, M. Zaffalon. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *JMLR*, 18:1–36, 2017.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate. *JRSS-B*, 57(1):289–300, 1995.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge Univ. Press, 2004.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *JMLR*, 7:1–30, 2006.

J. Friedman, T. Hastie, R. Tibshirani. Regularization paths for GLMs via coordinate descent. *J. Stat. Software*, 33(1):1–22, 2010.

G. H. Golub and C. F. Van Loan. *Matrix Computations*, 4th ed. Johns Hopkins Univ. Press, 2013.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian J. Statistics*, 6:65–70, 1979.

J. Nocedal and S. J. Wright. *Numerical Optimization*, 2nd ed. Springer, 2006.

F. Pedregosa et al. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830, 2011.

R. M. Schmidt, F. Schneider, P. Hennig. Descending through a crowded valley — benchmarking deep learning optimizers. *ICML*, 2021.

J. Yu and G. Wang. Reasoning-enhanced prompt strategies for multi-label classification. *ICCV Workshop on RetailVision*, 2025.

J. Yu, G. Wang, and Y. Deng. Beyond surface similarity: Evaluating structured output consistency in large language models. *IEEE Trans. Artificial Intelligence*, 2026.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

A. Asymptotic Rate Proof

Linearization. Let $e_k = \hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*$ and $\tilde{\mathbf{H}}_k = \int_0^1 \nabla^2 L(\boldsymbol{\theta}^* + se_k)\, ds$. Since $\nabla L(\boldsymbol{\theta}^*) = \mathbf{0}$, the iteration becomes $e_{k+1} = (\mathbf{I} - \mathbf{M}^{-1}\tilde{\mathbf{H}}_k)e_k$. By the Lipschitz property of $\nabla^2 L$ (constant M_3), $\|\tilde{\mathbf{H}}_k - \mathbf{H}(\boldsymbol{\theta}^*)\|_{\mathrm{op}} \le (M_3/2)\|e_k\|_2$, so $\tilde{\mathbf{H}}_k = \mathbf{H}(\boldsymbol{\theta}^*) + \mathbf{E}_k$ with $\|\mathbf{E}_k\|_{\mathrm{op}} \le M_3\|e_k\|_2/2$.

Upper bound. Substituting:

$$e_{k+1} = (\mathbf{I} - \mathbf{M}^{-1}\mathbf{H}(\boldsymbol{\theta}^*))e_k - \mathbf{M}^{-1}\mathbf{E}_k e_k.$$

Conjugate the first part by $\mathbf{M}^{1/2}$: $\mathbf{M}^{1/2}(\mathbf{I} - \mathbf{M}^{-1}\mathbf{H}(\boldsymbol{\theta}^*))\mathbf{M}^{-1/2} = \mathbf{I} - \mathbf{A}_*$, whose spectrum lies in $[0, 1-\mu_*]$ (by Loewner sandwich $2\lambda\mathbf{I} \preceq \mathbf{H}(\boldsymbol{\theta}^*) \preceq \mathbf{M}$). Hence $\|(\mathbf{I} - \mathbf{M}^{-1}\mathbf{H}(\boldsymbol{\theta}^*))e_k\|_{\mathbf{M}} \le (1-\mu_*)\|e_k\|_{\mathbf{M}}$. For the perturbation term, $\|\mathbf{M}^{-1}\mathbf{E}_k e_k\|_{\mathbf{M}} \le \|\mathbf{E}_k\|_{\mathrm{op}}\|e_k\|_2/\sqrt{\lambda_{\min}(\mathbf{M})} \le (M_3/(2\lambda_{\min}(\mathbf{M})^{3/2}))\|e_k\|_{\mathbf{M}}^2$. Combining:

$$\|e_{k+1}\|_{\mathbf{M}} \le (1-\mu_*)\|e_k\|_{\mathbf{M}} + \frac{M_3}{2\lambda_{\min}(\mathbf{M})^{3/2}}\|e_k\|_{\mathbf{M}}^2.$$

Tightness. Let $\mathbf{v}_*$ be the unit eigenvector of $\mathbf{A}_*$ with eigenvalue μ_*, and $\mathbf{u}_* := \mathbf{M}^{-1/2}\mathbf{v}_*$. For $\hat{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}^* + \eta\mathbf{u}_*$ with η small enough that the quadratic remainder is dominated by $\delta\|e_k\|_{\mathbf{M}}/2$, the ratio $\|e_{k+1}\|_{\mathbf{M}}/\|e_k\|_{\mathbf{M}}$ converges to $1-\mu_*$.

B. Holm–Bonferroni Decision at Step j

Let $p_{(1)} \le p_{(2)} \le \cdots \le p_{(D)}$. At family-wise level α, the step-down decision is:

- Reject $H_{(1)}$ iff $p_{(1)} \le \alpha/D$. If accept, all later $H_{(j)}$ are accepted.
- Reject $H_{(j)}$ iff $p_{(j)} \le \alpha/(D-j+1)$ *and* all previous $H_{(j')}$ for $j' < j$ were rejected.

At $D = 11$, $\alpha = 0.05$: thresholds $0.00455, 0.005, 0.00556, 0.00625, 0.00714, 0.00833, 0.01, 0.0125, 0.0167, 0.025, 0.05$. Adult ($p < 10^{-3}$) rejects at step 1. Covertype ($p = 0.001$) rejects at step 2 since $0.001 \le 0.05/10 = 0.005$. BreastCancer ($p = 0.051$) and Synth(5K,100) ($p = 0.014$) are accepted at the corresponding thresholds.

C. Detailed μ_* vs. Saturation Table

Table 3. μ_* and sigmoid-saturation fraction $s_\tau = \Pr_i[|\mathbf{x}_i^\top \boldsymbol{\theta}^*| > \tau]$ across four datasets and four λ. Source: `exp13_mu_star.json`.

Dataset	λ	μ_*	s_2	s_3
Adult	0.1	0.0904	12.2%	6.5%
Adult	1	0.0907	12.2%	6.5%
Adult	10	0.0938	12.0%	6.5%
Adult	100	0.1240	11.0%	6.0%
Covertype	0.1	0.0008	19.1%	3.6%
Covertype	1	0.0013	18.6%	3.5%
Covertype	10	0.0057	17.0%	3.1%
Covertype	100	0.1109	11.3%	0.9%
MNIST	0.1	2×10^{-5}	73.5%	59.3%
MNIST	1	2.4×10^{-4}	73.2%	58.5%
MNIST	10	0.0023	71.2%	55.1%
MNIST	100	0.0207	65.6%	46.1%
Digits	0.1	0.0155	70.9%	53.8%
Digits	1	0.0055	68.4%	49.9%
Digits	10	0.0408	56.7%	34.0%
Digits	100	0.2636	20.7%	3.6%

D. Effect Sizes and Cohen's d

Cohen's $d = \bar{\Delta}_d/s_{\Delta_d}$ provides a standardized effect-size measure complementary to the raw $\bar{\Delta}_d$ in Table 1. Conventional interpretation: $|d| < 0.2$ small, $0.2 \le |d| < 0.5$ medium, $|d| \ge 0.8$ large. For Adult, $d = -39.7/\sqrt{5} \cdot \sqrt{5} \approx -39.7$

(extreme); for Covertype, $d \approx -8.9$. For Synth(5K,100), $d \approx -3.7$ — also large in Cohen's scale, but the absolute $\bar{\Delta}_d = -0.07$pp is below the practical-significance floor, illustrating the gap between statistical and practical significance.

E. Power Curves

Figure (omitted for space) shows power $\beta(\bar{\Delta}, R)$ from Equation (**??**) as a function of $\bar{\Delta}$ for $R \in \{5, 10, 20\}$ at $D = 11$, H–B-corrected $\alpha = 0.05$. Key points: at $R = 5$ (our setting), the protocol has $\beta \geq 0.8$ for $|\bar{\Delta}| \geq 0.1$pp; at $R = 20$, $\beta \geq 0.95$ for the same. The chosen $R = 5$ is at the lower end of recommended sample sizes for $D = 11$ comparisons; a follow-up with $R \geq 10$ would tighten the power lower bound substantially.