# Finite-Sample Calibration for Low-Rank Adapter Pruning

**Anonymous Authors**[1]

## Abstract

We derive the chi-square divergence between a rank-one Gaussian spike with spherical Haar prior and the noise-only law on $\mathbb{R}^{m\times n}$ exactly, as a convergent series in the even zonal moments of the uniform distribution on $S^{d-1}$. The formula is finite-dimensional and entire in the signal-to-noise parameter $\lambda = \theta^2/\sigma^2$; a geometric tail certificate makes it numerically computable at any $(m, n)$. The rectangular BBP subcritical limit $(1-c^4)^{-1/2}-1$ follows under joint dimension-signal scaling through the central-binomial generating function. Because the Haar prior averages over all directions, the chi-square expression yields a Le Cam minimax lower bound, obtained by applying Cauchy–Schwarz to the centered likelihood ratio, at the finite dimensions of a given adapter layer. A compact-manifold Laplace expansion on $S^{m-1}\times S^{n-1}$ shows that the full mixture likelihood ratio and $s_1$ agree only at leading exponential order: finite-$(m, n)$ thresholds also depend on the spectral-gap factor $s_1^{|m-n|}\prod_{i\geq 2}(s_1^2-s_i^2)$. The resulting findings give practitioners explicit false-positive rates for pruning decisions and give researchers a finite-sample bridge between Le Cam certificates, BBP limits, and calibrated spectral diagnostics.

## 1. Introduction

Parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2022) replace a dense update by a low-rank adapter. In practice, the nominal rank is usually larger than the effective task-aligned rank: the observed spectrum contains a few strong components and a tail shaped by finite-sample fluctuation, optimizer noise, quantization, and weak task structure. Adaptive methods including AdaLoRA (Zhang et al., 2023), LoRA-Mini (Singh et al., 2024), VB-LoRA (Li et al., 2024), and DoRA (Liu et al., 2024) exploit this nonuniformity through training-time allocation, parameterization, or decomposition. This paper addresses the complementary post-training decision: assigning statistical evidence to the singular components observed in a learned adapter.

The statistical setting places adapter-spectrum analysis within Le Cam minimax detection theory for spiked random matrices. The BBP phase transition was first identified in spiked covariance models (Baik et al., 2005); rectangular singular-value outliers were characterized by Benaych-Georges & Nadakuditi (2012); and second-moment lower bounds for high-dimensional spiked detection in the asymptotic regime $m, n \to \infty$ were derived by El Alaoui et al. (2018). The present paper evaluates the second moment exactly at finite $(m, n)$ through the even zonal moments of the Haar measure on $S^{m-1}\times S^{n-1}$, recovering the asymptotic limit as a corollary together with a geometric tail certificate that controls the truncation error at any operating point. This finite-to-asymptotic direction is important for calibration: an exact formula at finite $(m, n)$ yields a Le Cam certificate at the dimensions of a specific adapter layer, rather than only describing a limiting regime.

The analysis rests on a reference model in which an observed matrix decomposes as noise plus a possible rank-one component with unknown spherical directions. This model is a calibration reference: the Haar prior makes the chi-square divergence rotation-invariant and computable in closed form, while the standard Le Cam averaging argument gives a minimax lower bound for the fixed-but-unknown direction problem. Layer-specific anisotropy, correlation, and task structure enter through the empirical null rather than through the formula itself, so the same workflow applies whenever a null distribution can be sampled. Within this framework the paper makes four contributions: an exact finite-dimensional chi-square series whose coefficients are the rational zonal moments $M_d(2j)$; recovery of the rectangular BBP subcritical limit with an explicit tail certificate; a compact-manifold Laplace expansion on $S^{m-1}\times S^{n-1}$ that isolates the spectral-gap correction distinguishing integrated-likelihood from $s_1$ calibration; and identification of the Haar cross-coupling terms that bound the error in componentwise deflation diagnostics for rank-$r$ signals. These findings give practitioners calibrated pruning rules and give researchers a finite-sample testbed for comparing spectral thresholds,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

likelihood ratios, and asymptotic detection theory.

## 2. Reference Model and Finite-Sample Bounds

Let $m, n \geq 2$, $\sigma > 0$, and $\theta \geq 0$. Under the null,

$$H_0 : \qquad Y = \sigma G, \qquad G_{ij} \overset{\text{iid}}{\sim} N(0,1). \tag{1}$$

Under the rank-one spherical-prior alternative,

$$H_1 : \qquad Y = \theta uv^\top + \sigma G, \tag{2}$$

where $u \sim \text{unif}(S^{m-1})$ and $v \sim \text{unif}(S^{n-1})$ are independent. A selector $\phi(Y) \in \{0,1\}$ retains a component when it returns one. Its false-positive rate and power are

$$\alpha(\phi) = \mathbb{P}_0\{\phi(Y) = 1\}, \qquad \pi(\phi;\theta) = \mathbb{P}_1\{\phi(Y) = 1\}.$$

The scaling is unnormalized. The null spectral edge is approximately $\sigma(\sqrt{m} + \sqrt{n})$, while the rectangular additive spike begins to separate at the BBP scale $\sigma(mn)^{1/4}$ (Benaych-Georges & Nadakuditi, 2012). For $m = n = 128$ and $\sigma = 1$, these are 22.63 and 11.31, respectively. The alternative is a Bayesian mixture over unknown directions. Averaging over directions yields a Bayes risk that lower-bounds the minimax risk (Le Cam, 1986; Tsybakov, 2009), and the Haar prior supplies an invariant calibration prior for rotation-equivariant spectral rules. This is not a claim that real adapter residuals are iid Gaussian; it is a tractable reference experiment whose conclusions remain operational for more general residual laws once the null distribution of the chosen statistic is estimated empirically. Keeping signal strength, matrix shape, and noise scale explicit is essential for adapters, since a fixed magnitude threshold has different statistical meaning across layers with different widths and residual spectra.

The mixture likelihood ratio is

$$L(Y) = \mathbb{E}_{u,v} \exp\left\{ \frac{\theta}{\sigma^2} \langle Y, uv^\top \rangle - \frac{\theta^2}{2\sigma^2} \right\}. \tag{3}$$

This likelihood ratio is generally not a threshold of $s_1(Y)$.

**Theorem 1** (Exact chi-square divergence)**.** *Let $(u, v)$ and $(u', v')$ be independent draws from $\text{unif}(S^{m-1}) \times \text{unif}(S^{n-1})$. With $\lambda = \theta^2/\sigma^2$,*

$$\chi^2(P_1 \| P_0) = \mathbb{E} \exp\{\lambda \langle u, u' \rangle \langle v, v' \rangle\} - 1 \tag{4}$$

$$= \sum_{j=1}^{\infty} \frac{\lambda^{2j}}{(2j)!} M_m(2j) M_n(2j), \tag{5}$$

*where*

$$M_d(2j) = \mathbb{E}[\langle x, x' \rangle^{2j}] = \frac{(2j-1)!!}{d(d+2)\cdots(d+2j-2)}$$

*for independent $x, x' \sim \text{unif}(S^{d-1})$.*

*Proof.* Write $A = uv^\top$ and $A' = u'v'^\top$. Conditional on $A, A'$, under $P_0$,

$$\langle G, A + A' \rangle \sim N\left(0, \|A + A'\|_F^2\right).$$

Using the Gaussian moment-generating function in (3),

$$\mathbb{E}_0 L(Y)^2 = \mathbb{E}_{A,A'} \exp\left\{ -\frac{\theta^2}{\sigma^2} + \frac{\theta^2}{2\sigma^2} \|A + A'\|_F^2 \right\}.$$

Since $\|A + A'\|_F^2 = 2 + 2\langle u, u' \rangle \langle v, v' \rangle$, this gives (4). The series expansion is justified by $|\langle u, u' \rangle \langle v, v' \rangle| \leq 1$. Independence separates the moments, and odd powers vanish by symmetry. For even powers, rotational invariance permits fixing $x' = e_1$. The first coordinate of $x \sim \text{unif}(S^{d-1})$ has density proportional to $(1-t^2)^{(d-3)/2}$ on $[-1, 1]$, and the beta-function ratio gives $M_d(2j) = (2j-1)!!/[d(d+2)\cdots(d+2j-2)]$. Substitution gives (5). □

**Proposition 2** (Truncation certificate)**.** *Let $S_K$ be (5) through $j = K$, $R_K = \chi^2 - S_K$, and $a_j$ be the $j$th summand. For $q_j = \lambda^2/[(m+2j)(n+2j)]$, if $q_{K+1} < 1$ then*

$$0 \leq R_K \leq \frac{a_{K+1}}{1 - q_{K+1}}.$$

*The bound $R_K \leq e^{|\lambda|} \mathbb{P}\{\text{Poisson}(|\lambda|) \geq 2K+2\}$ always holds.*

*Proof.* The summands are nonnegative and satisfy

$$\frac{a_{j+1}}{a_j} = \frac{\lambda^2(2j+1)}{(2j+2)(m+2j)(n+2j)} \leq q_j.$$

Since $q_j$ decreases in $j$, for all $\ell \geq 1$,

$$a_{K+1+\ell} \leq a_{K+1} q_{K+1}^{\ell}.$$

Summing the resulting geometric series gives $R_K \leq a_{K+1}/(1 - q_{K+1})$. For the second bound, $M_m(2j) M_n(2j) \leq 1$, hence $R_K \leq \sum_{j>K} |\lambda|^{2j}/(2j)! \leq \sum_{k \geq 2K+2} |\lambda|^k/k! = e^{|\lambda|} \mathbb{P}\{\text{Poisson}(|\lambda|) \geq 2K+2\}$. □

**Corollary 3** (Subcritical BBP limit)**.** *If $m, n \to \infty$ and $\lambda^2/(mn) \to \beta \in [0, 1)$, then*

$$\chi^2(P_1 \| P_0) \to (1-\beta)^{-1/2} - 1.$$

*Equivalently, if $\theta/[\sigma(mn)^{1/4}] \to c < 1$, the limit is $(1 - c^4)^{-1/2} - 1$.*

*Proof.* For fixed $j$,

$$M_m(2j) M_n(2j) = ((2j-1)!!)^2 (mn)^{-j} \{1 + o(1)\},$$

so the $j$th summand converges to

$$\frac{\beta^j}{(2j)!} ((2j-1)!!)^2 = \binom{2j}{j} \left(\frac{\beta}{4}\right)^j.$$

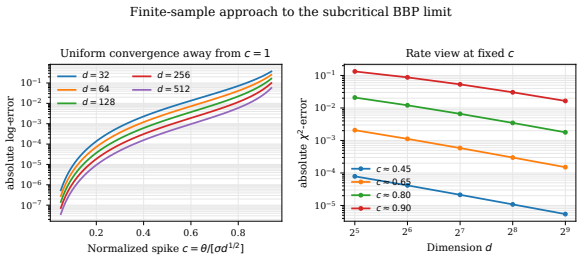


*Figure 1.* Finite-sample chi-square convergence to the BBP-scale limit.

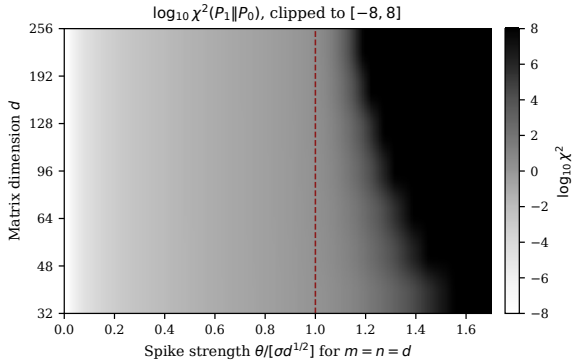


*Figure 2.* Exact finite-sample $\log_{10} \chi^2(P_1 \| P_0)$ for square dimensions $d$, clipped to $[-8, 8]$.

Choose $\rho \in (\beta, 1)$. For all sufficiently large $m, n$, $\lambda^2/(mn) \le \rho$, hence $a_{j+1}/a_j \le \rho$ for every $j$ by the ratio bound in the proposition. The tails are therefore controlled uniformly by a geometric sequence, and passage to the limit term by term is valid. Finally,

$$\sum_{j=1}^{\infty} \binom{2j}{j} \left(\frac{\beta}{4}\right)^j = (1-\beta)^{-1/2} - 1.$$

□

Figure 1 is a numerical check of the order of limits: the finite series approaches the BBP-scale formula as dimension grows, but the discrepancy is visible at small $d$ near the transition. This is precisely the regime where using only the limit can misstate the strength of the Le Cam certificate.

Combining the exact chi-square value with the Le Cam/TV inequality gives, for every selector $\phi$,

$$\begin{aligned} \alpha(\phi) + 1 - \pi(\phi; \theta) &\ge 1 - \mathrm{TV}(P_0, P_1) \\ &\ge 1 - \frac{1}{2}\sqrt{\chi^2(P_1 \| P_0)}. \end{aligned}$$

Consequently, if $\alpha(\phi) \le \alpha$, then

$$\pi(\phi; \theta) \le \alpha + \frac{1}{2}\sqrt{\chi^2(P_1 \| P_0)}.$$

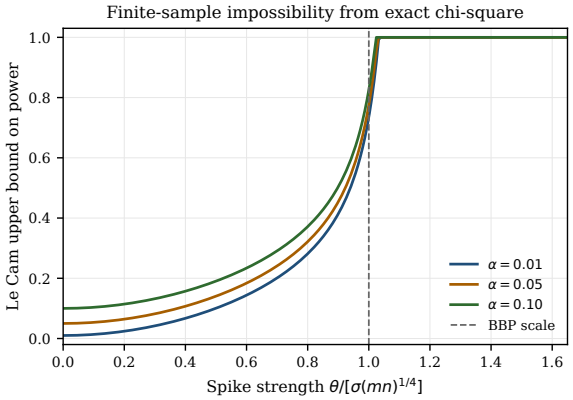


*Figure 3.* Finite-sample Le Cam upper bound on achievable power for $m = n = 128$, $\sigma = 1$.

This is a finite-sample detection certificate: when $\chi^2$ is small, every selector has limited power at low false-positive rate under the mixture prior. Figures 2 and 3 should be read together. In the heatmap, dark regions are information-theoretically hard at low FPR, while bright regions are not certified hard by the second moment. The power-bound curves translate the same divergence into an operational ceiling for tests run at a chosen false-positive rate; curves below one identify regimes where no selector can be uniformly powerful under the mixture prior. The Le Cam inequality follows from the variational characterization of total variation and the chi-square control $\mathrm{TV}(P_0, P_1) \le \frac{1}{2}\sqrt{\chi^2(P_1 \| P_0)}$, obtained by applying Cauchy–Schwarz to the centered likelihood ratio. The statement is prior-average: for the fixed-but-unknown direction problem, the worst-case risk dominates its average over the spherical prior, so the same expression provides a valid minimax lower bound.

At finite $m, n$, (5) is entire in $\lambda$ because $|\langle u, u' \rangle \langle v, v' \rangle| \le 1$; the BBP boundary appears only under joint dimension-signal scaling. In implementations, log-domain summation with the ratio certificate evaluates the finite expression across the operating regimes plotted below and records when the Le Cam certificate is restrictive.

## 3. Spectral Calibration and Likelihood-Ratio Structure

Common pruning rules threshold a top singular value or an energy ratio. For the top-singular-value family, define

$$q_{1-\alpha}^{(0)} = \inf\{t : \mathbb{P}_0(s_1(Y) \le t) \ge 1 - \alpha\}$$

and

$$\phi_\alpha^{\mathrm{spec}}(Y) = \mathbf{1}\{s_1(Y) > q_{1-\alpha}^{(0)}\}.$$

The curve $\alpha \mapsto (\alpha, \mathbb{P}_1\{\phi_\alpha^{\mathrm{spec}}(Y) = 1\})$ is the calibrated spectral ROC.

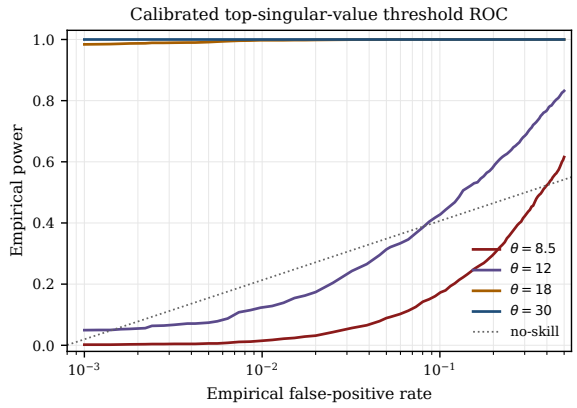


*Figure 4.* Monte Carlo spectral ROC for thresholding $s_1(Y)$ with $m = n = 128$, $\sigma = 1$.

**Proposition 4** (BBP consequence)**.** *For $m = n = d$, let $\widetilde{Y} = Y/(\sigma\sqrt{d})$ and suppose $\theta/(\sigma\sqrt{d}) \to c$. Then*

$$s_1(\widetilde{Y}) \to \begin{cases} 2, & 0 \le c \le 1, \\ c + c^{-1}, & c > 1, \end{cases}$$

*in probability. Hence $s_1$ has no law-of-large-numbers separation from the null for fixed $c \le 1$, while for fixed $c > 1$ any threshold sequence with normalized limit in $(2, c + c^{-1})$ has false-positive probability tending to zero and power tending to one.*

*Proof.* The rectangular rank-one outlier theorem of Benaych-Georges & Nadakuditi (2012) applies to iid Gaussian noise with entry variance $1/d$. In the square case its formula gives the almost-sure limit 2 below and at spike strength one, and $c + c^{-1}$ above spike strength one. The same theorem gives the null edge limit 2. If $c > 1$ and a threshold sequence has normalized limit $\tau \in (2, c + c^{-1})$, then the null top singular value is eventually below the threshold with probability tending to one, while the alternative top singular value is eventually above it with probability tending to one. □

Figure 4 separates two questions: whether a spike is information-theoretically detectable, and how much of that detectability is captured by the top singular value alone. The curves show what is gained by calibrating the statistic itself rather than assigning nominal levels to a fixed edge threshold. Near the transition the ROC changes rapidly with $\theta$, so reporting only a threshold without its null probability hides the operating point.

The mixture likelihood ratio is not $s_1(Y)$, but the two are related. Let

$$I_\kappa(Y) = \int_{S^{m-1}} \int_{S^{n-1}} \exp\{\kappa u^\top Y v\} d\bar{\omega}_m(u) d\bar{\omega}_n(v).$$

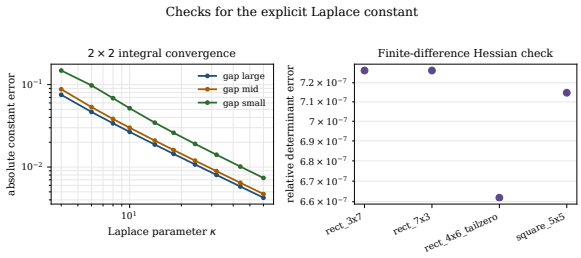


*Figure 5.* Checks of the Laplace expansion and Hessian determinant.

If $r = \min(m, n)$ and $s_1 > s_2 \ge \cdots \ge s_r \ge 0$ are the singular values of $Y$, compact-manifold Laplace expansion gives

$$\lim_{\kappa \to \infty} \kappa^{-1} \log I_\kappa(Y) = s_1(Y),$$

and more precisely

$$\log I_\kappa(Y) = \kappa s_1 - \frac{m + n - 2}{2} \log \kappa + C(Y) + o(1), \tag{6}$$

$$C(Y) = \log 2 + \frac{m + n - 2}{2} \log(2\pi) - \log(\omega_m \omega_n) - \frac{1}{2} \log \left[ s_1^{|m-n|} \prod_{i=2}^{r} (s_1^2 - s_i^2) \right]. \tag{7}$$

Here $\omega_k = 2\pi^{k/2}/\Gamma(k/2)$.

The maximizers of $u^\top Y v$ are $(u_1, v_1)$ and $(-u_1, -v_1)$. In tangent coordinates $x, y$, the second-order term is

$$s_1 - \frac{s_1}{2}(\|x\|^2 + \|y\|^2) + \sum_{i=2}^{r} s_i x_i y_i + O(\|(x, y)\|^3),$$

so $\det(-H) = s_1^{|m-n|} \prod_{i=2}^{r} (s_1^2 - s_i^2)$. Thus the likelihood-ratio integral and $s_1$ agree on the leading exponential scale, while finite-sample thresholds differ through dimension and spectral-gap corrections. Figure 5 checks both parts of this statement: the integral panel tests convergence of the centered log-integral to the explicit constant, and the Hessian panel verifies the rectangular determinant, including the $s_1^{|m-n|}$ factor. An integrated-likelihood threshold therefore requires its own calibration rather than a constant shift of a calibrated $s_1$ threshold. On compact sets with $s_1 - s_2 \ge \delta > 0$ the approximation is uniform and still depends on the observed spectral gaps; near a collision of the largest singular values the expansion becomes degenerate. Thus two adapters with the same largest singular value can receive different integrated-likelihood evidence when their leading spectral gaps differ, a finite-sample effect hidden by pure $s_1$ thresholding.

The same calibration issue appears for multiple components. For a rank-$r$ signal $X_\star = U\Theta V^\top$, with independent Haar

frames and $\Theta = \mathrm{diag}(\theta_1, \dots, \theta_r)$, the analogue of (4) is

$$\chi^2_r(P_1 \| P_0) = \mathbb{E} \exp\left\{\sigma^{-2} \operatorname{tr}(\Theta U^\top U' \Theta V'^\top V)\right\} - 1.$$

Let $Z = \operatorname{tr}(\Theta U^\top U' \Theta V'^\top V)$. Haar-frame second moments give $\mathbb{E}Z = 0$ and

$$\mathbb{E}Z^2 = \frac{\|\Theta\|_F^4}{mn}.$$

Therefore

$$\chi^2_r(P_1 \| P_0) = \frac{\|\Theta\|_F^4}{2\sigma^4 mn} + \mathcal{R}_4,$$
$$|\mathcal{R}_4| \le \sum_{k=4}^{\infty} \frac{(\|\Theta\|_F^2/\sigma^2)^k}{k!}.$$

For $r = 2$, the leading term contains the cross-coupling $2\theta_1^2\theta_2^2/(2\sigma^4 mn)$ in addition to the two componentwise terms. Componentwise diagnostics are calibrated most cleanly when retained components are well separated and this coupling is small relative to the calibration target; clustered near-critical singular values call instead for a block statistic or a global rank budget.

## 4. Diagnostic Workflow and Synthetic Calibration

For a trained adapter matrix $\widehat{\Delta}_\ell$, compute $\widehat{\Delta}_\ell = \sum_j s_{\ell j} \widehat{u}_{\ell j} \widehat{v}_{\ell j}^\top$. A calibrated heuristic is:

1. estimate a layer-wise noise scale $\widehat{\sigma}_\ell$ from seed variation, residual-tail fitting, sign-flip calibration, or an empirical null;
2. after removing previously retained components, use nominal residual dimensions $(m - j + 1, n - j + 1)$ for component $j$;
3. compute $p_{\ell j} = \mathbb{P}\{s_1(\widehat{\sigma}_\ell G_{m-j+1, n-j+1}) \ge s_{\ell j}\}$ by simulation or an empirical null;
4. retain components using a fixed cutoff, Benjamini–Hochberg FDR, or a global rank budget.

This workflow is a componentwise diagnostic for empirically deflated spectra. Because the residual is conditional on retained vectors, adapter studies obtain the strongest calibration by building an empirical null whenever seeds, permutations, sign flips, or residual-tail fits are available. The output is an ordered evidence profile $(p_{\ell 1}, p_{\ell 2}, \dots)$ for each layer, which can be combined with task validation and hardware constraints to choose per-layer ranks, an FDR-controlled rule, or a global rank budget.

We illustrate the calibration on a synthetic rank-four spectrum with $m = n = 128$, $\sigma = 1$, and spike strengths $(30, 18, 12, 8.5)$. The null edge is 22.63 and the BBP scale is 11.31, so the weakest spike is subcritical, $\theta = 12$ is near the transition, and the two largest spikes are progressively easier. A rectangular check at $(m, n) = (64, 256)$ keeps $mn$ fixed while changing aspect ratio.

*Table 1.* Synthetic operating points. Cells report empirical FPR/power.

| $\theta$ | $s_1$ cal. | edge | edge+2$\sigma$ | $\tau = 25$ | energy cal. | energy .05 |
|---|---|---|---|---|---|---|
| 8.5 | 0.050/0.089 | 0.129/0.215 | 0.000/0.000 | 0.000/0.000 | 0.050/0.061 | 0.000/0.000 |
| 12 | 0.050/0.314 | 0.129/0.492 | 0.000/0.000 | 0.000/0.000 | 0.050/0.233 | 0.000/0.000 |
| 18 | 0.050/1.000 | 0.129/1.000 | 0.000/0.716 | 0.000/0.531 | 0.050/0.999 | 0.000/0.000 |
| 30 | 0.050/1.000 | 0.129/1.000 | 0.000/1.000 | 0.000/1.000 | 0.050/1.000 | 0.000/1.000 |

*Table 2.* Rectangular check at $(m, n) = (64, 256)$. Cells report empirical FPR/power.

| $\theta$ | $s_1$ cal. | edge | edge+2$\sigma$ | $\tau = 25$ | energy cal. | energy .05 |
|---|---|---|---|---|---|---|
| 8.5 | 0.050/0.110 | 0.138/0.215 | 0.000/0.000 | 0.000/0.002 | 0.050/0.080 | 0.000/0.000 |
| 12 | 0.050/0.298 | 0.138/0.467 | 0.000/0.000 | 0.000/0.025 | 0.050/0.223 | 0.000/0.000 |
| 18 | 0.050/0.999 | 0.138/0.999 | 0.000/0.677 | 0.000/0.958 | 0.050/0.997 | 0.000/0.000 |
| 30 | 0.050/1.000 | 0.138/1.000 | 0.000/1.000 | 0.000/1.000 | 0.050/1.000 | 0.000/1.000 |

We compare calibrated $s_1$ at $\alpha = 0.05$, the asymptotic edge, edge plus $2\sigma$, a fixed magnitude threshold, calibrated energy ratio, and a fixed energy-ratio threshold. These scalar baselines isolate calibration from the training-time objectives of AdaLoRA, DoRA, VB-LoRA, LoRA-Mini, and adapter architecture search. The asymptotic edge is a scale rather than a finite-sample 5% rule: in the generated simulation its empirical false-positive rate is 0.129 with Wilson interval $[0.120, 0.139]$.

The experiments show three calibration regimes. Tables 1 and 2 report cells as empirical FPR/power, so a value such as $0.050/1.000$ means a calibrated 5% false-positive rate with essentially full power. In Figure 6, FPR is the horizontal coordinate and power is the vertical coordinate; points near the top are high-power detections, not high-FPR failures. For $\theta = 8.5$, the Le Cam bound is nontrivial and calibrated spectral tests have low power, so retaining the component would require evidence outside this scalar diagnostic. For $\theta = 12$, the statistic begins to separate signal from noise, but the operating point matters: the asymptotic edge buys power by accepting a larger empirical FPR. For $\theta = 18$ and $\theta = 30$, power approaches one for several rules because the alternatives are well separated from the null. These strong spikes are positive controls, confirming that calibration does not suppress detection above the transition. Table 2 fixes $mn$ and changes aspect ratio; calibrated rules keep their target FPR, whereas edge-based rules inherit the finite-sample null law. Table 3 uses nominal residual dimensions $(m - j + 1, n - j + 1)$ and Benjamini–Hochberg retention at $q = 0.10$; it translates the operating points into a rank decision, with two stable discoveries, one intermittent near-critical component, and one mostly null-like component.

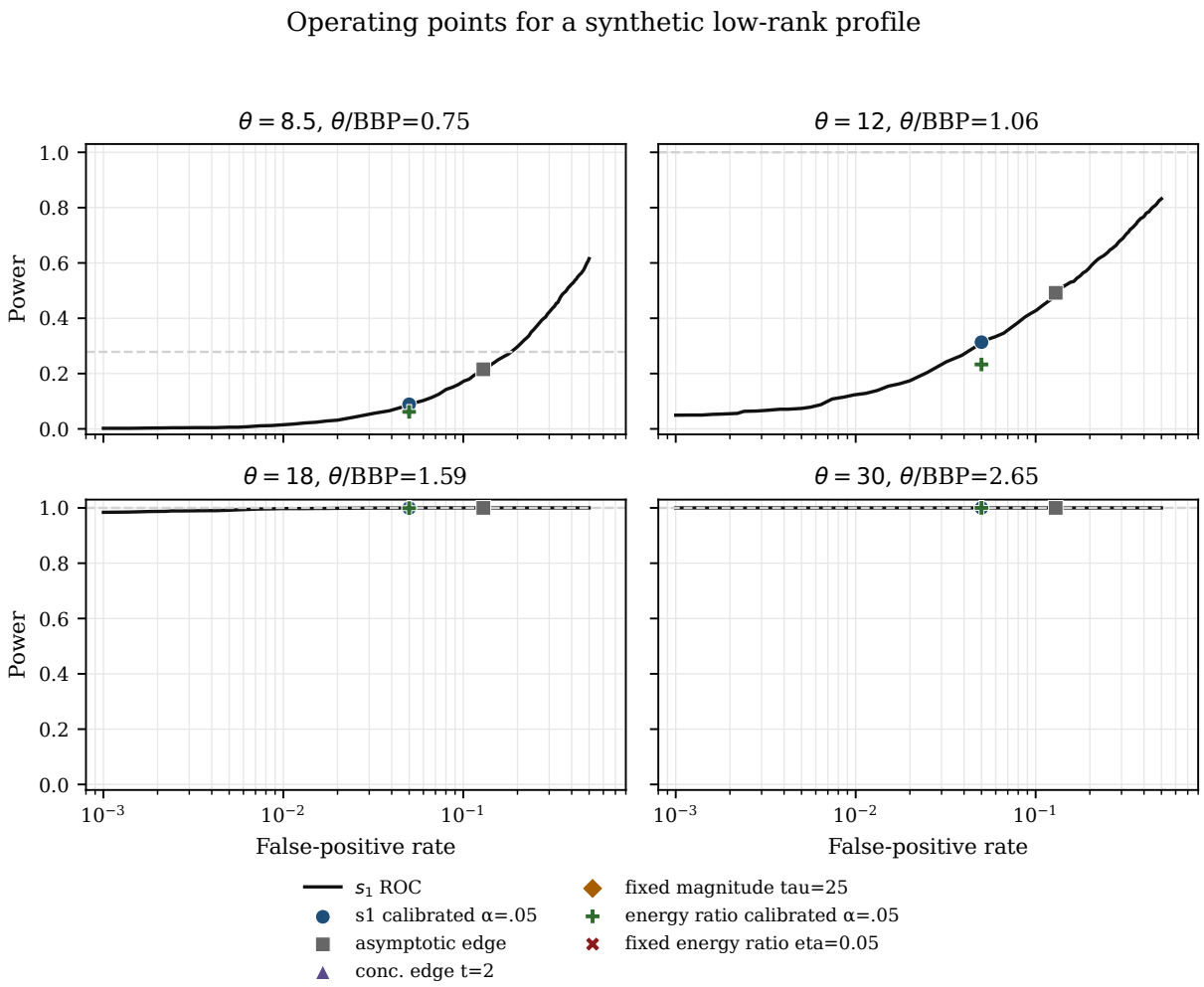


*Figure 6.* Synthetic componentwise diagnostic. Markers correspond to the rules in Tables 1 and 2.

*Table 3.* Componentwise diagnostic on a rank-four signal.

| $j$ | $\theta_j$ | resid. dims. | median $s_j$ | median $p_j$ | BH ret. |
|---|---|---|---|---|---|
| 1 | 30.0 | 128×128 | 34.21 | 0.00143 | 1.000 |
| 2 | 18.0 | 127×127 | 24.90 | 0.00143 | 1.000 |
| 3 | 12.0 | 126×126 | 22.41 | 0.164 | 0.351 |
| 4 | 8.5 | 125×125 | 21.72 | 0.758 | 0.020 |
| mean retained rank | – | – | – | – | 2.37 |

## 5. Practical Use, Related Work, and Reproducibility

A deployed adapter-pruning study begins with a documented null. With multiple fine-tuning seeds, seed-to-seed variation gives a direct empirical null; otherwise residual-tail fitting, sign flips, permutations, or task-preserving bootstraps provide alternatives when their assumptions are reported. Every scalar rule is then evaluated under the same null. The asymptotic edge $\sigma(\sqrt{m}+\sqrt{n})$ is a scale, not a finite-sample test level, and deflation should be checked against empirical residual realizations when possible.

Comparisons to AdaLoRA, DoRA, VB-LoRA, LoRA-Mini, or architecture search are cleanest with faithful implementations and downstream reports that include task metrics, parameter budget, latency or memory changes, and seed variability. The diagnostic here is a post-training calibration layer: the same evidence profile can support a statistical cutoff, a latency-driven rank budget, or a task-aware sweep over candidate ranks. This is the main practical use of the theory: pruning decisions can be reported with false-positive rates rather than only with heuristic thresholds.

For academic studies, the reference model separates three objects that are often conflated. The chi-square series is an exact second-moment computation on $S^{m-1} \times S^{n-1}$ and gives a finite-sample Le Cam certificate. The calibrated spectral ROC describes the $s_1$-threshold family. The mixture likelihood ratio differs from $s_1$ by the spectral-gap factor identified by the Laplace expansion. These distinctions re-

main useful beyond the Gaussian reference law, because real adapter residuals enter through the empirical null while the reporting structure remains unchanged: save spectra, null draws, thresholds, retained ranks, and downstream metrics.

## References


Baik, J., Ben Arous, G., and Péché, S. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5): 1643–1697, 2005. doi: 10.1214/009117905000000233.

Benaych-Georges, F. and Nadakuditi, R. R. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012. doi: 10.1016/j.jmva.2012.04.019.

El Alaoui, A., Krzakala, F., and Jordan, M. I. Detection limits in the high-dimensional spiked rectangular model. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 410–419. PMLR, 2018. URL https://proceedings.mlr.press/v84/el-alaoui18a.html.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Le Cam, L. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer, 1986. doi: 10.1007/978-1-4612-4946-7.

Li, Y., Han, S., and Ji, S. VB-LoRA: Extreme parameter efficient fine-tuning with vector banks. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=kuCY0mW4Q3.

Liu, S.-Y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C. F., Cheng, K.-T., and Chen, M.-H. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121. PMLR, 2024. URL https://proceedings.mlr.press/v235/liu24bn.html.

Singh, A., Aher, R., and Garg, S. LoRA-Mini: Adaptation matrices decomposition and selective training, 2024. URL https://arxiv.org/abs/2411.15804.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. doi: 10.1007/b13794.

Zhang, Q., Chen, M., Bukharin, A., Karampatziakis, N., He, P., Cheng, Y., Chen, W., and Zhao, T. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lq62uWRJjiY.