Finite Resources False Discovery Rate Control in Structured Hypothesis Spaces

Anonymous Author(s)
Affiliation
Address
email

Abstract

Scientific discovery relies on large-scale hypothesis testing. Identifying true discoveries while controlling false ones faces two challenges: reference (null) data are costly to obtain, leaving finite-data uncertainty, and the hypothesis space often carries structure the procedure should exploit. We present a framework for FDR control when each hypothesis is evidenced only by a finite count of null draws, leaving its p-value uncertain, and the hypothesis space carries arbitrary structure represented through a reproducing kernel. We present two decision rules that are both robust to structural mis-specification, yet offer a distinct trade-off between the assumptions behind FDR control and statistical power. The first rule controls the FDR without requiring the structure to be correct, only that the ordering it induces is not inverted relative to the truth; the second maximizes power by adapting mirror-statistic control into count space, utilizing an analytical framework to assess FDR control when exact mirror symmetry is relaxed. The tractability of the RKHS framework also yields per-hypothesis uncertainty, which we use to allocate null samples efficiently.

1 Introduction

Consider a scientist analyzing a cohort of patients: given observational data Z_{obs}, the researcher formulates hypotheses and tests each by comparing Z_{obs} against samples drawn from reference patients, yielding a p-value per hypothesis. This pipeline faces three challenges. First, testing many hypotheses simultaneously inflates the risk of false discoveries. Second, dependence among hypotheses (for example, spatial proximity) calls for dedicated methods. Third, obtaining enough reference samples per hypothesis (usually matched to the experiment's covariates) is a severe practical barrier: even with 1,000 hypotheses and a target FDR of 0.05, guidelines suggest approximately 10,000 reference(H_0) samples *per hypothesis* [24]. We address this with two components: a count-based likelihood that uses the finite null sample directly, with no point estimate of the latent p-value, and a prior on the hypothesis-space structure through which each hypothesis borrows strength from its neighbors. While both components reduce the sampling cost on their own, they complement each other to further lower the per-hypothesis sampling requirement, as we empirically validate. To the best of our knowledge, this paper presents the first unified framework to address these three challenges simultaneously, serving as the first finite-sample extension of the Bayesian FDR framework that actively leverages structured hypothesis spaces [18].

The remainder of this paper follows: Section 2 reviews related work; Section 3 presents the problem setup, generative model, and non-spatial solution; Section 4 presents the spatial estimator and two decision rules: Rule 1 controls FDR without a mirror (flip-invariance) assumption, requiring only that the spatial ordering not be inverted relative to the truth, while Rule 2 requires it for exactness and otherwise holds up to a slack that we characterize analytically; both share a ranking that is optimal

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

under correct specification. Section 5 derives per-hypothesis uncertainty and an adaptive allocation policy; Section 6 presents the evaluation.

2 Preliminary Definitions and Related works

False Discovery Rate (FDR) is the expected proportion of false positives (V) among all rejected hypotheses(R): FDR $= E\left[\frac{V}{R} \mid R > 0\right] P(R > 0)$. **Local False Discovery Rate (lfdr):** the posterior null probability of a *specific hypothesis* given its statistic z under a two-group mixture [11]: $\text{lfdr}(z) = P(H_0 \mid Z = z) = \frac{\alpha_0 f_0(z)}{\alpha_0 f_0(z) + \alpha_1 f_1(z)}$ with priors α_0, α_1 and null and alternative densities f_0, f_1. For *related hypotheses*, the literature has transitioned from treating structure as a nuisance [7], to leveraging it to boost statistical power. Structured FDR methods fall into three branches: explicit dependency models, which encode joint or conditional distributions but need sparse, pre-defined dependency graphs; adaptive weighting, where LAWS [9] and STRAW [23] re-weight p-values by their spatial neighbors and AdaPT [17] learns rejection thresholds from covariates; and regularization, where SmoothFDR [22] applies total-variation penalties to the prior null probabilities over graph edges, which limits it to discrete graphs. We build on the regularized FDR framework of [18], which models the prior null probability $\alpha(\mathbf{x})$ directly in an RKHS over arbitrarily structured spaces. Our second rule adapts the symmetry principle of the knockoff filter [4], which controls FDR when the null statistics are sign-symmetric ([5] study approximate symmetry): in count space the sign flip becomes the reflection $k \leftrightarrow m - k$, and we quantify the FDR when this symmetry is perturbed by estimating the scoring surface from the counts. **Resource allocation.** Online FDR control [13, 1, 20] allocates an α-budget across a stream of hypotheses; closer to us, [3] allocates a null-sampling budget across hypotheses, but over an unstructured space with no pooling. Our N hypotheses are fixed, decided jointly, and structured.

3 Problem Setup and Methods

We consider a fixed set of N hypotheses, each associated with a location loc_i in a domain $\mathcal{X}$. Formally, $\mathcal{X}$ can be any space admitting a symmetric positive-definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and an associated RKHS $\mathcal{H}_K$. Each hypothesis has a latent state $\theta_i \in \{0, 1\}$, where $\theta_i = 0$ denotes the null $H_{0,i}$ and $\theta_i = 1$ the alternative. For each hypothesis we compute a statistic Z_i, draw m_i independent samples from its null distribution, and record the number k_i of draws at least as extreme as Z_i; the observed data are the tuples (loc_i, m_i, k_i). We build on the RKHS Bayesian-FDR framework of [18], in which the prior null probability is a smooth function over a kernel geometry; our departure is that each hypothesis's evidence is a finite count of null draws, entering as a count likelihood rather than a resolved p-value. The generative model is a three-level hierarchy,

$$\text{loc}_i \xrightarrow{\alpha(\cdot)} \theta_i \xrightarrow{f_{\theta_i}} p_i^* \xrightarrow{\text{Binomial}(m_i, \cdot)} k_i,$$

developed below.

Structural layer. The probability that hypothesis i is null depends on its location through an unknown function $\alpha : \mathcal{X} \to [0, 1]$: $\Pr(\theta_i = 0 \mid \text{loc}_i) = \alpha(\text{loc}_i)$.

Latent p-value layer. Conditional on θ_i, the hypothesis generates a latent p-value p_i^* from the corresponding component of a two-group mixture:$p_i^* \mid \theta_i = 0 \sim f_0 \quad ; \quad p_i^* \mid \theta_i = 1 \sim f_1$, where $f_0 = U[0, 1]$ and f_1 is concentrated near zero: the standard two-group model [11].

Observation layer. Only the count k_i is observed, and by construction $k_i \mid p_i^*, m_i \sim \text{Binomial}(m_i, p_i^*)$.

3.1 The non-spatial case

The lfdr is defined in terms of the latent p_i^*; standard methods substitute the plug-in $\hat{p}_i = (k_i + 1)/(m_i + 1)$ and treat it as exact, ignoring finite-sample uncertainty. In GWAS [15], for example, FDR significance for 10^6 markers requires $p < 10^{-7}$, which would demand $m \gtrsim 10^8$ null draws just to resolve the plug-in; the budget m_i bounds the attainable evidence. We instead *marginalize out* p_i^*, integrating the observation model against the latent density $f_j(p)$, which gives exact count

likelihoods under each mixture component:

$$P_j(k,m) = \int_0^1 \binom{m}{k} p^k (1-p)^{m-k}\, f_j(p)\, dp, \qquad j \in \{0,1\}. \tag{1}$$

For a general f_j this integral has no closed form. The Beta family is the natural choice: $\mathrm{Beta}(a,b)$ with $a < 1 < b$ concentrates the alternative near $p = 0$, and conjugacy to the binomial makes (1) closed-form. With $f_0 = \mathrm{Beta}(1,1) = U[0,1]$ and $f_1 = \mathrm{Beta}(a,b)$:

$$P_0(k,m) = \frac{1}{m+1}, \qquad P_1(k,m) = \binom{m}{k} \frac{B(k+a,\, m-k+b)}{B(a,b)} \tag{2}$$

where B is the Beta function. The null pmf is uniform in k, hence invariant under the reflection $k \leftrightarrow m-k$, the symmetry Rule 2's mirror construction requires (Section 4); a non-uniform empirical null [11] can be modeled as a Beta mixture, preserving conjugacy. We denote $P_{0,i} = P_0(k_i, m_i)$ and $P_{1,i} = P_1(k_i, m_i; b)$. Theorem 4 (App. C) shows that the marginal lfdr recovers the classical lfdr at rate $O(1/m)$ for $\hat{p} = (k+1)/(m+1)$ away from 0 and 1, and at rate $\Theta(m^{a-1})$ at the lattice edge $k = 0$, where both saturate to zero; the interior rate beats the $O(1/\sqrt{m})$ sampling noise in $\hat{p}$, so count space costs nothing asymptotically there (nor at the edge when $a < 1/2$). The count-space lfdr follows from the marginal mixture:

$$\mathrm{lfdr}_{\mathrm{marg}}(k_i, m_i) = \frac{\hat{\bar{\alpha}}\, P_{0,i}}{\hat{\bar{\alpha}}\, P_{0,i} + (1-\hat{\bar{\alpha}})\, P_{1,i}}, \tag{3}$$

Global mixture parameters. By Theorem 5, the marginal distribution of (k_i, m_i) reduces to the non-spatial two-group mixture, so the global parameters $\bar{\alpha}$ and b can be estimated from the marginal counts alone, independently of the spatial structure. Theorem 5 establishes the two-group form, not identifiability: the count likelihood sees f_1 only through its first m moments, so identifying $(\bar{\alpha}, b)$ requires $m_i \geq 2$ (at $m = 1$, $\Pr(k=1) = \bar{\alpha}/2 + (1-\bar{\alpha})\, a/(a+b)$ is one equation in two unknowns), and estimating a as well would require $m_i \geq 3$; we therefore hold $a < 1$ fixed, encoding the canonical alternative shape (singular at $p = 0$, decreasing on $(0,1]$). We estimate $\bar{\alpha}$ with Storey's estimator [**?**] on $\hat{p}_i$, $\hat{\bar{\alpha}}(\lambda) = \#\{i : \hat{p}_i > \lambda\}/[N(1-\lambda)]$, which is upward biased ($\Pr_0(\hat{p}_i > \lambda) \geq 1-\lambda$ on the lattice, so $\mathbb{E}\,\hat{\bar{\alpha}}(\lambda) \geq \bar{\alpha}$), taking $\hat{\bar{\alpha}} = \min_{\lambda \in \Lambda} \hat{\bar{\alpha}}(\lambda)$ over $\Lambda = \{0.05, 0.10, \ldots, 0.90\}$. Given $\hat{\bar{\alpha}}$, b is the maximizer over a compact $\mathcal{B} \subset (1, \infty)$ of the marginal log-likelihood,

$$\hat{b} = \arg\max_{b \in \mathcal{B}} \sum_{i=1}^{N} \log\Big[\hat{\bar{\alpha}}\, P_{0,i} + (1-\hat{\bar{\alpha}})\, P_1(k_i, m_i; b)\Big], \tag{4}$$

3.2 The spatial case

We now lift the constant $\bar{\alpha}$ to a function $\alpha : \mathcal{X} \to [0,1]$, recovering the structural layer, so the likelihood at hypothesis i becomes

$$P(k_i, m_i \mid \mathrm{loc}_i) = \alpha(\mathrm{loc}_i)\, P_{0,i} + (1 - \alpha(\mathrm{loc}_i))\, P_{1,i}, \tag{5}$$

with $P_{0,i}, P_{1,i}$ as in (2); its population average recovers the non-spatial mixture (Theorem 5, Appendix C.2), so $(\hat{\bar{\alpha}}, \hat{b})$ are reused. We fit $\hat{\alpha}(\mathrm{loc})$ as the regularized maximum-likelihood estimator

$$\hat{\alpha} = \arg\min_{\alpha \in \mathcal{H}_K} \; -\sum_{i=1}^{N} \log\big[\alpha(\mathrm{loc}_i)\, P_{0,i} + (1 - \alpha(\mathrm{loc}_i))\, P_{1,i}\big] \;+\; \lambda\, \|\alpha - \hat{\bar{\alpha}}\|^2_{\mathcal{H}_K} \;+\; \gamma\, \Lambda_{[0,1]}(\alpha). \tag{6}$$

The RKHS penalty $\|\alpha - \hat{\bar{\alpha}}\|^2_{\mathcal{H}_K}$ controls how far α may deviate from the global rate $\hat{\bar{\alpha}}$, at the smoothness scale set by K, so a hypothesis with no nearby information is pulled toward $\hat{\bar{\alpha}}$; $\gamma\, \Lambda_{[0,1]}(\alpha)$ is a soft penalty keeping $\alpha(\mathrm{loc}) \in [0,1]$ (its form, the solver and convergence follow [18]). By the Representer Theorem [16] the minimizer has the form $\hat{\alpha}(\mathrm{loc}) = \hat{\bar{\alpha}} + \sum_{j=1}^{N} c_j\, K(\mathrm{loc}, \mathrm{loc}_j)$, the objective is strictly convex in $\mathbf{c} \in \mathbb{R}^N$, and natural gradient descent [2] minimizes it with no inversion of the Gram matrix K (Appendix B.4).

4 Decision Rules and FDR Control

We consider two decision rules: *Rule 1: Model-Free* uses $\hat{\alpha}$ only to order the hypotheses, then applies Sun–Cai's running-average threshold rule to the marginalized non-spatial lfdr. *Rule 2: Mirror Statistics* uses $\widehat{\mathrm{lfdr}}_{\mathrm{spatial}}$ directly as the rejection score, importing the mirror-statistic machinery of [4] into count space. A variant of Rule 2 has an exact FDR guarantee (Rem. 10, App. E.5); we relax it, trading a bounded slack for power and computation, and bound the resulting FDR in terms of that slack.

We first establish Rule 1, whose FDR control owes nothing to the mirror assumption and does not require the spatial model to be correct, at a cost in power. *For Rule 2*, and as an independent contribution, we address the mirror assumption in the structured setting: control holds up to a slack, removable in principle by a leave-one-out (LOO) construction, and largest for isolated hypotheses, which is not a serious loss since those revert to the global mean $\bar{\alpha}$ by construction. Theorem 3 shows that under correct specification ranking by the spatial lfdr is optimal; both rules share this ranking and differ only in where they stop, and in our experiments Rule 2 spends more of the FDR budget, an inherent control–power tradeoff. *Finally,* we develop an *analytical framework* for the breakdown of the mirror assumption, built on a helper problem, the *folded construction*, that satisfies the assumption exactly and lets a user bound the slack on each specific dataset. We formalize the assumption that separates the two rules:

Definition 1 (Mirror assumption)**.** *A scoring rule $i \mapsto T_i = g(\hat{\alpha}(\mathrm{loc}_i), k_i, m_i)$ satisfies the* mirror assumption *if, for every null i, the surface value $\hat{\alpha}(\mathrm{loc}_i)$ entering T_i is invariant under the flip $k_i \leftrightarrow m_i - k_i$; equivalently, $\hat{\alpha}(\mathrm{loc}_i)$ is measurable with respect to the folded statistic $\breve{k}_i = \min(k_i, m_i - k_i)$.*

Rule 1: Model-Free. **(1)** Compute $\widehat{\mathrm{lfdr}}_{\mathrm{spatial}}(\hat{\alpha}(\mathrm{loc}_i), k_i, m_i)$ and, from the non-spatial fit, $\mathrm{lfdr}_{\mathrm{marg}}(k_i, m_i)$ for every $i \in \{1, \ldots, N\}$. **(2)** Order all N hypotheses by $\widehat{\mathrm{lfdr}}_{\mathrm{spatial}}$ in increasing order. **(3)** Apply the running-average rule [21] along this ordering on the marginal lfdr: reject the largest prefix $\mathcal{R}_1$ whose running average satisfies $\frac{1}{|\mathcal{R}_1|}\sum_{i\in\mathcal{R}_1} \mathrm{lfdr}_{\mathrm{marg}}(k_i, m_i) \le \tau$. The structural model sets the ordering; the accounting is entirely marginal.

The validity of Rule 1 does not depend on $\hat{\alpha}$ being correct, only on the ordering it induces: control holds unless the ordering is systematically inverted relative to the truth, that is, unless it prefers hypotheses that are more likely null than their counts alone suggest. The reason is that each rejected hypothesis adds to the expected number of false rejections exactly its probability of being null given all the data, $\mathrm{lfdr}_i^* = \Pr(\theta_i = 0 \mid \text{all data})$. Rule 1 budgets with $\mathrm{lfdr}_{\mathrm{marg}}$ instead, so $\mathrm{FDR}(\mathcal{R}_1) \le \tau$ as long as, on the rejected set, $\mathrm{lfdr}_{\mathrm{marg}}$ does not underestimate lfdr^* on average:

$$\mathbb{E}\Big[\tfrac{1}{|\mathcal{R}_1|\vee 1}\textstyle\sum_{i\in\mathcal{R}_1}\big(\mathrm{lfdr}_i^* - \mathrm{lfdr}_{\mathrm{marg}}(k_i, m_i)\big)\Big] \;\le\; 0. \tag{PC}$$

A random ordering satisfies (PC) with equality, an informative one with room to spare, and only an inverted ordering violates it. Accordingly, our robustness claim refers to *structural* misspecification: an adversarially inverted ordering can break control. When the structure is informative, the ordering front-loads alternatives, the running average of $\mathrm{lfdr}_{\mathrm{marg}}$ grows slowly, and the rule rejects deeper into the ordering (proof in App. D.1).

Theorem 1 (Rule 1 FDR control)**.** *Under the generative model of Section 3 and Assumption 4, with $\mathrm{lfdr}_{\mathrm{marg}}$ evaluated at the true $(\bar{\alpha}, b)$: **(a)** if* (PC) *holds, $\mathrm{FDR}(\mathcal{R}_1) \le \tau$; **(b)** if $\mathrm{Cov}\big(\mathbb{I}\{i \in \mathcal{R}_1\}, \mathrm{lfdr}_i^* - \mathrm{lfdr}_{\mathrm{marg}}(k_i, m_i)\big) \le 0$ for every i, then $\mathrm{mFDR}(\mathcal{R}_1) = \mathbb{E}|\mathcal{R}_1 \cap \mathcal{H}_0| \,/\, \mathbb{E}|\mathcal{R}_1| \le \tau$. Neither part requires $\hat{\alpha}$ to be correctly specified.*

Rule 2: Mirror Statistics in Count Space. We score each hypothesis with the spatial lfdr, pair it with its count-space mirror, and apply the Barber–Candès step-up threshold rule [4]:

$$\hat{t}_q \;=\; \max\left\{ t \in \{T_i\} \cup \{\tilde{T}_i\} \;:\; \frac{1 + \#\{i : \tilde{T}_i \le t\}}{1 \vee \#\{i : T_i \le t\}} \;\le\; \tau \right\}, \qquad \mathcal{R}_2 = \{i : T_i \le \hat{t}_q\}. \tag{7}$$

$$T_i \;=\; \widehat{\mathrm{lfdr}}_{\mathrm{spatial}}(\hat{\alpha}(\mathrm{loc}_i),\, k_i,\, m_i)\,, \qquad \tilde{T}_i \;=\; \widehat{\mathrm{lfdr}}_{\mathrm{spatial}}(\hat{\alpha}(\mathrm{loc}_i),\, m_i - k_i,\, m_i)\,.$$

The reflection $k \leftrightarrow m - k$ is the count-space form of the Barber–Candès symmetry: the null pmf $P_0(k \mid m) = 1/(m+1)$ is flip-invariant at any *fixed* $\hat{\alpha}$ (App. E). For the data-dependent $\hat{\alpha}$ it holds only approximately, and we quantify the violation by the *flip-one-out stability*

$$\beta_N \;:=\; \max_{i \in \mathcal{H}_0} \left\| \hat{\boldsymbol{\alpha}}^{(i)} - \hat{\boldsymbol{\alpha}} \right\|_\infty, \tag{8}$$

where $\hat{\boldsymbol{\alpha}}^{(i)}$ denotes the spatial estimator refit with k_i replaced by $m_i - k_i$.

Theorem 2 (Rule 2 error control for the deployed plug-in)**.** *Fit $\hat{\alpha}, \hat{b}$ on the raw counts and run* (7)*, and let $\delta = L\,\beta_N + L_b\,|\Delta\hat{b}| = O(1/\lambda) + O(1/N)$ be the single-flip score displacement, with $\beta_N, |\Delta\hat{b}|$ the flip-one-out stabilities of surface and shape and L, L_b the Lipschitz constants of $\widehat{\mathrm{lfdr}}_{\mathrm{spatial}}$ in α, b. Under Assumptions 4–3, the structural conditions of App. E (Assumptions 6–7), and a flip-stability hypothesis on the data-dependent procedure,*

$$\mathrm{mFDR}(\mathcal{R}_2) \le \tau + O(\delta) \qquad \textit{and} \qquad \mathrm{FDR}(\mathcal{R}_2) \le \tau + C_{\mathrm{BC}}\,\delta,$$

the averaged bound under threshold stability (Assumption 8) and the per-realization bound under ranking stability (Assumption 9), with $C_{\mathrm{BC}} = 4$ when a constant fraction of nulls falls below the threshold. Neither hypothesis is implied by the structural conditions; without them an unconditional fallback $\mathrm{mFDR}(\mathcal{R}_2) \le \tau + O(\underline{s})$ holds at the score-grid scale $\underline{s}$, and with a flip-invariant score ($\delta = 0$, the folded construction) both reduce to $\mathrm{FDR}(\mathcal{R}_2) \le \tau$ exactly. Precise statements and constants are in App. E (Thms 7, 8, 6; Prop. 5).

The two stability hypotheses require that a single flip move the data-dependent cutoff (*threshold stability*, Assumption 8) or the boundary ranking (*ranking stability*, Assumption 9) by $O(\delta)$ rather than an $O(1)$ jump; they are the only genuinely load-bearing conditions in the theorem, and the bound degrades gracefully around them. If the scoring surface is held fixed independently of the data being tested (an oracle, or the leave-one-out and folded constructions of Rem. 10) the flip cannot move it, the slack vanishes, and $\mathrm{FDR} \le \tau$ holds exactly. We do not use these constructions, because withholding each hypothesis's own count from its score discards exactly the local information that gives Rule 2 its power, a loss we quantify as a Fisher-information tax (App. E.5, eq. 137). We instead fit on the full data and pay the additive slack above: δ is set by the regularization, $O(1/\lambda)$ at the cross-validated λ and not shrinking with N, and is computable by a single refit, while the discrimination and bounded-boundary conditions (Assumptions 6, 7) confine the slack to hypotheses adjacent to the threshold. What those conditions do *not* deliver is the flip-stability of the cutoff and ranking; that is the hard input, and when it fails the grid-scale fallback above is what survives.

A single flip moves the surface at loc_i by an amount set by the local curvature, so the violation concentrates at hypotheses with few neighbors and shrinks as the design infills around them, down to the regularization floor (Section E.5). Discreteness adds no slack and in fact *reduces* it, a count-space mechanism absent in continuous-data FDR control: the mirror symmetry holds exactly on the lattice $\{0, 1, \ldots, m_i\}$, and a flip smaller than the local grid spacing crosses no decision (Lemma 10), so the slack from learning the surface on its own data is confined to near-threshold hypotheses rather than charged across all of them; the spacing is set by the Beta–Binomial likelihood-ratio increment, bounded away from zero wherever signal and null are distinguishable (App. E), so the margin is genuine.

Proof walkthrough: The argument turns on the null count being symmetric about $m_i/2$ (Assumption 4): $k_i \overset{d}{=} m_i - k_i$, so the side a null lands on, $B_i = \mathbb{I}(k_i > m_i/2)$, is a fair coin, which is what the Barber–Candès reverse martingale needs to certify $\mathrm{FDR} \le \tau$ exactly, provided the statistic used to rank and threshold does not depend on which side the coin showed. That independence is Definition 1. A fixed α and b would give an exact mirror, since a score that does not read the data is unmoved by the flip; our $\hat{\alpha}$ is fit on the counts, so flipping k_i moves the surface at loc_i and Definition 1 holds only approximately. The standard remedy is *masking*: fit the score on flip-invariant information alone, as AdaPT and the knockoff procedures do to obtain exact finite-sample control; in count space it is realized by the leave-one-out and folded constructions (Rem. 10, App. E.5), which pay no slack. *We deliberately do not deploy them:* masking discards information, while the purpose of Rule 2 is to retain power, so we fit on the raw counts and accept a controlled relaxation, with the flip-invariant constructions as the exact baseline against which the slack is measured (Prop. 5). Steps 1–3 of the proof establish exact control when the surface is independent of the coins $\{B_i\}_{i \in \mathcal{H}_0}$ (the exchangeability (133) *does not hold for the plug-in*); Step 4 bounds, by an influence argument,

the displacement δ of every score under a single flip; the averaged bound then follows from a slack identity that reduces the mFDR gap to each null's *own* distance to the cutoff, and the per-realization bound from a perturbed reverse martingale whose one-step defect is δ, both localized by the lattice (Lemma 10) to a thin boundary layer of near-threshold hypotheses.

Theorem 3 (Optimality of the spatial lfdr score)**.** *Assume the compositional model of Section 3 holds with true null-probability function $\alpha^*(\mathrm{loc})$, and that $(\hat{\alpha}, \hat{b}, a) = (\alpha^*, b^*, a^*)$ are correctly specified. At any marginal FDR (mFDR) level τ, rejecting a sublevel set of the spatial score $T_i = \widehat{\mathrm{lfdr}}_{\mathrm{spatial}}(\hat{\alpha}(\mathrm{loc}_i), k_i, m_i)$ achieves at least as many true discoveries in expectation as any rule that ranks by the location-free score $\mathrm{lfdr}_{\mathrm{marg}}(k_i, m_i)$. If $\alpha^*(\mathrm{loc})$ is non-constant on a set of positive measure, the inequality is strict at every level at which, with positive probability, the location-free rule rejects a hypothesis while retaining one with a smaller true posterior null probability.*

Proof sketch (full proof in Appendix E.11). At any mFDR level the most powerful rejection region is a sublevel set of the true posterior null probability [21], which under correct specification is T_i. A location-free ranking cannot order hypotheses whose counts favor one and whose locations favor the other; exchanging such an inverted pair lowers the mFDR and raises the expected number of true discoveries.

Both rules rank by T_i and differ only in where they stop: Rule 1's running average of $\mathrm{lfdr}_{\mathrm{marg}}$ overstates the error of an informative prefix (PC), so it realizes an mFDR below τ, whereas Rule 2 stops by the mirror estimate at the cost of the stability conditions of Theorem 2, and in our experiments operates closer to τ (Figure 2).

Analyzing the violation: the folded construction. The slack in Theorem 2 is the price of fitting $\hat{\alpha}$ on the signed count k_i. To analyze it we pair the problem with a *folded construction*: the same estimator fit on the folded counts $\breve{k}_i = \min(k_i, m_i - k_i)$, which the flip leaves fixed. The folded surface cannot move when a null is flipped, so it satisfies Definition 1 exactly and controls FDR with no slack, and the gap between the two surfaces localizes and bounds the violation on any given dataset (Section E.5). It is not a drop-in replacement: folding merges a large count (strong evidence *for* the null) with the small count that looks like signal, which we show costs a fixed fraction of the per-hypothesis information, worst at small m; it is a tool for analysis and diagnosis, not an alternative rule.

5 Uncertainty and Smart Allocation

5.1 Per-hypothesis confidence intervals

Two kinds of uncertainty enter differently. Finite-count uncertainty is modeled: the marginalization of Section 3.1 carries it into the lfdr itself, exactly. Structural uncertainty in $\hat{\alpha}$ is quantified on top of that model, asymptotically, and enters no FDR guarantee; the intervals below combine the two for classification and allocation, not for error control.

For a per-hypothesis confidence interval on the lfdr, we synthesize the sources of uncertainty efficiently by constructing $1-\gamma/2$ marginal intervals for each component and combining them via a union bound. Monotonicity of the lfdr in both arguments (Lemma 2, Appendix C.6) makes this combination cheap: the extrema of the lfdr over the uncertainty rectangle fall on opposite corners, so the joint interval reduces to two function evaluations. For the spatial prior, the inverse-Hessian (Laplace) variance $\sigma_i^2 = \mathbf{k}_i^\top \hat{H}^{-1} \mathbf{k}_i$ (App. C.4) yields the Gaussian interval $[\alpha_i^{\text{low}}, \alpha_i^{\text{high}}] := \hat{\alpha}(\mathrm{loc}_i) \pm z_{1-\gamma/4}\, \sigma_i$. For the latent p-value, we use the $\gamma/4$ and $1 - \gamma/4$ quantiles of the Beta–Binomial posterior $p_i^* \mid k_i, m_i \sim \mathrm{Beta}(k_i + 1, m_i - k_i + 1)$ ($[10^{-10}, 1 - 10^{-10}]$ when $m_i = 0$) to obtain $[p_i^{\text{low}}, p_i^{\text{high}}]$. By the monotonicity argument above, the joint interval is $\mathrm{lfdr}_i^{\text{low}} = \mathrm{lfdr}(\alpha_i^{\text{low}}, p_i^{\text{low}})$ and $\mathrm{lfdr}_i^{\text{high}} = \mathrm{lfdr}(\alpha_i^{\text{high}}, p_i^{\text{high}})$ and the union bound gives joint coverage $\Pr\big(\mathrm{lfdr}_i^{\text{true}} \in [\mathrm{lfdr}_i^{\text{low}}, \mathrm{lfdr}_i^{\text{high}}]\big) \geq 1 - \gamma$ *asymptotically*, for budgets m_i fixed in advance (adaptive budgets: App. C.14.5).

Floors of uncertainty The per-hypothesis uncertainty $\sigma_i^2 = \mathbf{k}_i^\top \hat{H}^{-1} \mathbf{k}_i$ has an irreducible floor: local resampling at loc_i grows the data-driven Fisher information $\hat{v}_i$ in $\hat{H} = K^\top \mathrm{diag}(\hat{v}) K + 2\lambda K$, shrinking the data contribution to σ_i^2, but the regularization term $2\lambda K$ is fixed by the fact that we

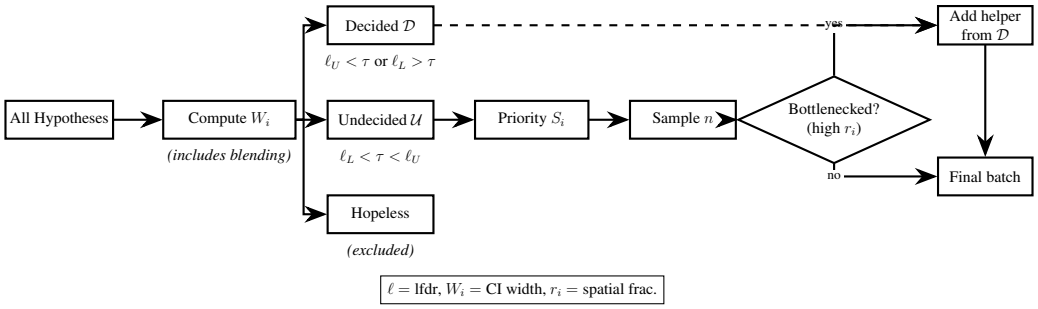


Figure 1: High view of the allocation policy main blocks.

observe at finitely many locations $\{\mathrm{loc}_j\}_{j=1}^N$ and the prior must bridge the gaps between them. We call this the *variance floor* for hypothesis i (Appendix C.9). Its magnitude tracks spatial support: small when loc_i has many near neighbors, and as large as $1/(2\lambda)$ when loc_i is isolated and σ_i^2 collapses to the irreducible uncertainty around the global prior $\hat{\tilde{\alpha}}$ (Appendix C.9). Operationally, what matters is not driving σ_i^2 to zero but its gap to the floor: hypotheses above the floor can be sharpened by more sampling, those at it cannot, and this gap-to-floor distinction drives the allocation rule of Section 5.2.

5.2 From tractable uncertainty to allocation

How should a finite null budget be allocated across hypotheses? Rather than equal effort m everywhere, we direct draws where uncertainty has room to shrink and where the lfdr is most likely to change a rejection decision. As a proof of concept we present a policy for Rule 1, whose marginal-lfdr threshold interacts directly with allocation; for Rule 2 the policy must read only folded counts to preserve the mirror symmetry (App. C.14.5). Implementation details (closed-form saturation and headroom expressions, numerical safeguards, the micro/macro update schedule) are in Appendix C.14.

At any allocation step, each hypothesis lfdr confidence interval $[\mathrm{lfdr}_i^{\mathrm{low}}, \mathrm{lfdr}_i^{\mathrm{high}}]$ (Section 5.1) classify it into one of three classes: ***Decided*** hypotheses have a CI entirely on one side of $\hat{\tau}_q$. ***Ambiguous*** hypotheses have a CI straddling $\hat{\tau}_q$: this is where allocation effort is directed. *Hopeless* hypotheses still straddle $\hat{\tau}_q$ even as σ_i^2 approaches its variance floor (Section 5.1); the irreducible uncertainty rules out a definitive decision, so they are removed from the priority queue but remain available for helper sampling on behalf of neighbors. Figure 1 summarizes this classification.

Priority score among ambiguous hypotheses. Among ambiguous hypotheses, the next batch of null draws goes to those whose uncertainty has the most room to shrink *and* the largest impact on the lfdr at loc_i or its kernel-coupled neighbors. The variance σ_i^2 splits into a local (Fisher) and a spatial (Hessian/regularization) channel, each with its own floor, and two factors per channel govern how much a sample helps: a *saturation factor* sat (how far the channel is from its floor) and a *headroom factor* κ (the channel's share of the current variance). The priority score combines them as

$$S_i \;=\; \underbrace{\beta \cdot \mathrm{sat}_{F,i} \cdot \kappa_{F,i} \cdot W_i^2}_{\text{local (Fisher) channel}} \;+\; \underbrace{(1-\beta) \cdot \mathrm{sat}_{H,i} \cdot \kappa_{H,i} \cdot \sum_{j \in \mathcal{A}} K_{ij}^2\, W_j^2}_{\text{spatial (Hessian) channel}}, \tag{9}$$

with $W_i = \mathrm{lfdr}_i^{\mathrm{high}} - \mathrm{lfdr}_i^{\mathrm{low}}$ the CI width at i, $\mathcal{A}$ the set of ambiguous neighbors and $\beta \in [0,1]$ the balance between the channels. A channel contributes only when both its factors are positive: a saturated Fisher channel drops its local term regardless of the remaining CI width, and likewise the spatial channel (closed forms of sat_F, sat_H, κ_F, κ_H in Appendix C.14). Setting β too high concentrates samples in dense regions and underexplores small clusters of alternatives (the cluster-size bias of [18]); too low over-relies on spatial borrowing and risks missing locally distinguishable signals. We adapt β during a short burn-in from the estimated null fraction $\hat{\alpha}_0$: more local sampling when most hypotheses appear null, more borrowing when alternatives are clustered. Decided and hopeless hypotheses are not scored by (9), but their residual local uncertainty propagates through the kernel to nearby ambiguous hypotheses: when an ambiguous i is bottlenecked by spatial uncertainty

(its Hessian channel dominates), the policy compares the benefit of sampling i directly against that of sampling a non-ambiguous neighbor j and substitutes j if larger, so non-ambiguous hypotheses keep contributing without reopening their own decisions (the β-tuning rule and the benefit comparison are in Appendix C.14). Two pieces of machinery support the policy. First, the joint lfdr CI of Section 5.1 formalizes when a hypothesis becomes decided or hopeless; for a schedule fixed in advance, a union bound over K rounds controls, with probability at least $1 - K\gamma$, the event that any definitive decision is incorrect or later contradicted (Appendix C.5), while under the adaptive policy this fixed-time guarantee does not apply (App. C.14.5); as the bound is conservative, we keep decided hypotheses decided and sample them only through the hopeless mechanism. Second, re-solving (6) as data arrive is handled by Sherman–Morrison micro-updates between full re-optimizations every T batches (Appendix C.13).

6 Evaluations

We evaluate two things: the finite-resource lfdr machinery of Sections 3.1 and 5, on 10 semi-synthetic datasets built from ADBench [14], an anomaly-detection benchmark of difficult, high-dimensional real-world data, on the AlpacaEval 2.0 LLM-as-judge benchmark [19], and on a HIGGS collider search; and the adaptive allocation policy of Section 5.2, again on ADBench.

6.1 Finite-resource lfdr quality

ADBench. We subsample $1{,}000$ points per dataset, cluster them to define null and alternative regions, label unclustered points as "background" (spatially isolated nulls), and flip 20% of labels at random to simulate corruption. Latent p-values are drawn as $p_i^* \sim U(0,1)$ for nulls (including background) and $p_i^* \sim \mathrm{Beta}(0.3, 3)$ for alternatives. The oracle is RegFDR [18] on the true $\{p_i^*\}$; the baselines are RegFDR on point-estimated p-values, ignoring finite-m uncertainty, and Benjamini–Hochberg [6], ignoring both uncertainty and structure. The fitted model is never the generating one, so Figure 2 tests FDR control under misspecification rather than self-consistency. The fit holds the alternative shape at a value a that we vary, estimates b from the counts, and uses a Matérn kernel of fixed smoothness, while the generator varies the latent-field smoothness ν. For each of the 10 datasets we cross 8 values of the fitted $a \in [0.1, 0.8]$ with $\nu \in \{d/2 + k : k = 1, \ldots, 5\}$, where d is the dataset dimension, giving $10 \times 8 \times 5 = 400$ configurations per method; the fitted alternative is misspecified whenever $a \neq 0.3$, and the kernel whenever ν differs from the fitted smoothness (App. F). All methods are evaluated on identical realizations and share every non-method-specific hyperparameter (regularization strength, target FDR level τ, total sampling budget). Solid lines report the mean over the 8 values of a; shaded bands show the corresponding 0.95 quantile envelope.

Both rules keep FDR well below the target 0.1 across the whole sweep (Figure 2), whereas BH and RegFDR on point-estimated p-values violate it substantially in the finite-m regime, a striking outcome since RegFDR is the oracle that generated the labels: the excess FDR is due entirely to the misspecifications and the point-estimation error. Rule 2 (count-space mirror) achieves consistently high power at an FDR closer to the budget but safely within it; Rule 1 (spatially ordered Sun–Cai) is very conservative, leaving substantial budget unused, and its power collapses under stronger misspecifications.

LLM-as-judge benchmark (AlpacaEval 2.0). A real LLM-as-judge benchmark with *no spatial structure* and no information borrowed across hypotheses: $N = 805$ AlpacaEval 2.0 prompts, on which a judge evaluates 12 challenger models against a fixed baseline. The null is that a prompt does not systematically distinguish model quality, so any preference over the baseline is random. To avoid leakage between the test statistic and the ground truth, the 12 challengers are split into two disjoint groups of size $m = 6$: Group A supplies the count k_i of its challengers failing to defeat the baseline, and Group B independently sets the ground truth by majority vote (full setup in Appendix G). BH on standard discrete p-values cannot reject at $m = 6$, since the smallest attainable p-value $1/(m+1) \approx 0.143$ exceeds any conventional FDR target, so we compare against the uncalibrated heuristic $\tilde{p}_i = 1 - \mathrm{wins}_i/m$ fed into BH. Our framework

Method	Disc.	FDR / Power
BH($\tilde{p}_i$)	48	0.021 / 0.092
Rule 1	167	0.042 / 0.312
Rule 2	168	0.042 / 0.315

Table 1: LLM evaluation. $\tilde{p}_i = 1 - \frac{\mathrm{wins}_i}{m}$.

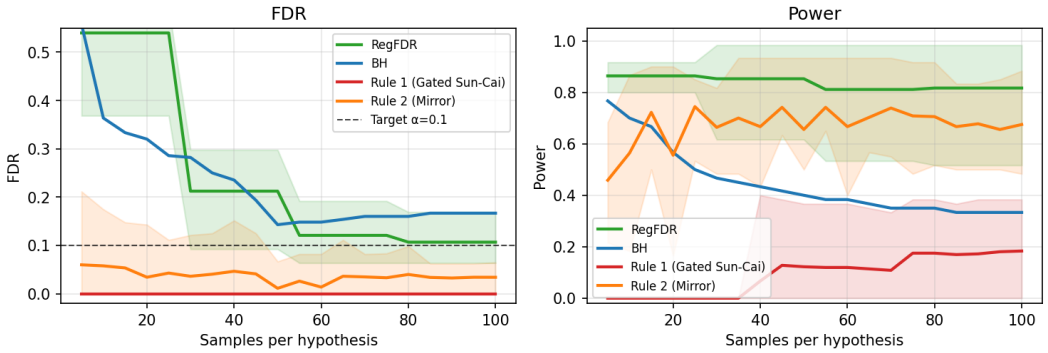


Figure 2: **Sensitivity of FDR Control and Power to Specifications.** FDR and power across number of null samples. The variance areas are built over the 10 datasets and perturbations over the fitted alternative shape a and the generator's smoothness ν. The dashed line marks the target FDR level $\tau = 0.1$.

Method	Disc.	False	True	FDP	Recall
BH	154	39	115	0.253 [0.191, 0.327]	0.564
Rule 1	95	10	85	0.105 [0.058, 0.183]	0.417
Rule 2	92	9	83	0.098 [0.052, 0.176]	0.407

Table 2: HIGGS, $N = 490$ bins, $\tau = 0.1$. Counts are rounded means over 5 repetitions; brackets are Wilson 95% intervals for the FDP given the reported counts.

substantially outperforms it (Table 1), the two rules producing nearly identical results, as expected when no spatial information is available for them to differ on.

Costly null simulation (HIGGS). In a collider search, events are binned by kinematics and each bin is tested for an excess over a Standard-Model background whose null distribution has no closed form: null draws come from simulation, so m is bounded by simulation cost rather than chosen. We use the HIGGS setup of [18] on a 10% subsample ($\approx$ 9,800 of the 98,050 events of [10]), binned by k-means into $N = 490$ cells over (lepton p_T, missing E_T), 204 of them signal-enriched, with a Matérn kernel over bin centroids; labels are used for evaluation only, and our rules see $m \approx 20$ draws per bin. BH is applied to the per-bin one-sided binomial p-values of [18]. Table 2 reports rounded means over 5 repetitions at $\tau = 0.1$: BH exceeds the target by a factor of about 2.5, while both rules hold it at the cost of about a quarter of BH's recall.

6.2 Adaptive allocation

We compare three allocation strategies on the ADbench datasets: *Active*, the policy of Section 5.2, prioritizing ambiguous hypotheses by remaining capacity; *Uniform*, which samples ambiguous hypothe-

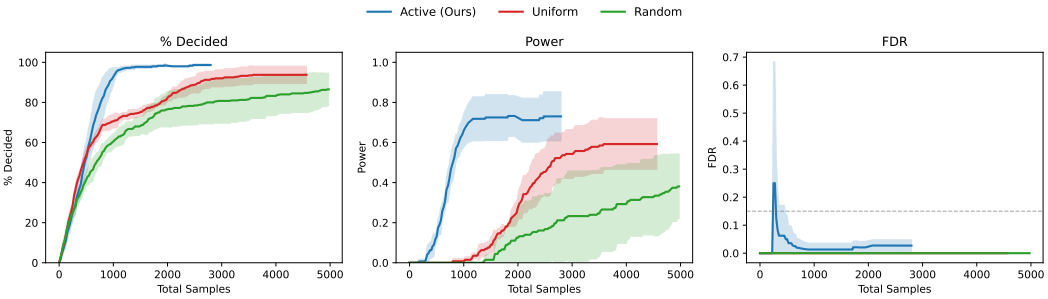


Figure 3: **Comparing allocations on ADbench.** three allocation strategies. (a) Fraction of hypotheses with a decided lfdr CI. (b) Power. (c) False discovery rate; the dashed line marks the target $\tau = 0.15$.

ses uniformly; and *Random*, which samples uniformly across all hypotheses, ignoring prior decisions. Each method receives the same total null-draw budget; we report metrics across 4 random seeds and the 10 datasets simultaneously. As Figure 3 shows, Active makes more decisions per unit budget (panel a) and reaches higher power (panel b) than both baselines, both during transient and saturation. Hypotheses discovered only by Active are likely true alternatives whose signal spatial borrowing resolves where local Fisher information alone cannot. The target $\tau = 0.15$ holds at saturation but is exceeded transiently at small budgets (panel c; see Limitations). Figure 4 shows that Active also converges faster to the oracle α^{oracle} in RKHS distance.

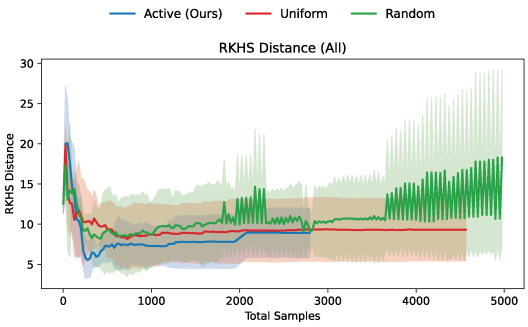


Figure 4: Distance to oracle($\|\hat{\alpha} - \alpha^{\text{oracle}}\|_{\mathcal{H}_\mathcal{K}}$).

7 Conclusions and Limitations

We presented a framework for FDR control that marginalizes the p-value uncertainty of finite null sampling, exploits structure in the hypothesis space, and quantifies per-hypothesis uncertainty tractably enough to drive allocation, with two decision rules whose relationship we characterized. Empirically it improves substantially over baselines under finite resources and allocates samples efficiently. We see strong potential at the intersection with kernel theory, e.g. hierarchical graphs via hyperbolic kernels.

Limitations. (1) We assume uniform p-values under H_0, which might not align with empirical null estimation [11]; this can be addressed by modeling the null as a Beta mixture, which preserves the Beta-Binomial conjugacy. Rule 2 also takes the null count law to be mirror-symmetric [4], quantifying only how the plug-in surface breaks the flip-invariance of the *scores*; a null with $O(1/m_i)$ mirror asymmetry adds $O(1/m_{\min})$ to its bound (Remark 8), and larger departures we do not quantify. (2) The variance floors create a distinctive bandit setting in which the goal shifts from minimizing uncertainty to recognizing when the floor has been reached. We believe this novel problem warrants further investigation. (3) Adaptive allocation makes each budget m_i depend on the counts observed so far. Each allocation is decided before its draw, so the likelihood in (6) is unaffected, but the fixed-m sampling law is not: the marginal factorization behind $(\hat{\tilde{\alpha}}, \hat{b})$, the discrete-uniform null behind Rule 2, and the fixed-time coverage of the intervals of Section 5.1 assume budgets fixed in advance, and the burn-in, which only sets β, does not remove this. We therefore present Section 5.2 as a policy rather than a theorem, report its transient FDR excess (Figure 3c), and in App. C.14.5 state what holds under adaptive allocation and give a folded variant under which Rule 2 remains exact.

8 Societal Impact

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

[1] E. Aharoni and S. Rosset. Generalized α-investing: Definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):771–794, 2014. doi: 10.1111/rssb.12048.

[2] S.-i. Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

[3] R. Ao, H. Chen, D. Simchi-Levi, and F. Zhu. Online resource allocation with average budget constraints, 2025. URL `https://arxiv.org/abs/2402.11425`.

[4] R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5), Oct. 2015. ISSN 0090-5364. doi: 10.1214/15-aos1337. URL `http://dx.doi.org/10.1214/15-AOS1337`.

[5] R. F. Barber, E. J. Candès, and R. J. Samworth. Robust inference with knockoffs, 2019. URL `https://arxiv.org/abs/1801.03896`.

[6] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

[7] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001. doi: 10.1214/aos/1013699998.

[8] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[9] T. T. Cai, W. Sun, and Y. Xia. Laws: A locally adaptive weighting and screening approach to spatial multiple testing. *Journal of the American Statistical Association*, 117(539):1370–1383, Jan. 2021. ISSN 1537-274X. doi: 10.1080/01621459.2020.1859379. URL `http://dx.doi.org/10.1080/01621459.2020.1859379`.

[10] Y. Chen, E. A. Huerta, J. Duarte, P. Harris, D. S. Katz, M. S. Neubauer, D. Diaz, F. Mokhtar, R. Kansal, S. E. Park, V. V. Kindratenko, Z. Zhao, and R. Rusack. A fair and ai-ready higgs boson decay dataset. *Scientific Data*, 9(1), Feb. 2022. ISSN 2052-4463. doi: 10.1038/s41597-021-01109-0. URL `http://dx.doi.org/10.1038/s41597-021-01109-0`.

[11] B. Efron. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, Mar. 2004. ISSN 1537-274X. doi: 10.1198/016214504000000089. URL `http://dx.doi.org/10.1198/016214504000000089`.

[12] W. Fithian and L. Lei. Conditional calibration for false discovery rate control under dependence. *The Annals of Statistics*, 50(6):3091 – 3118, 2022. doi: 10.1214/21-AOS2137. URL `https://doi.org/10.1214/21-AOS2137`.

[13] D. P. Foster and R. A. Stine. α-investing: A procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):429–444, 2008. doi: 10.1111/j.1467-9868.2007.00643.x.

[14] S. Han, X. Hu, H. Huang, M. Jiang, and Y. Zhao. Adbench: Anomaly detection benchmark, 2022. URL `https://arxiv.org/abs/2206.09426`.

[15] R. C. Johnson, G. W. Nelson, J. L. Troyer, J. A. Lautenberger, B. D. Kessing, C. A. Winkler, and S. J. O'Brien. Accounting for multiple comparisons in a genome-wide association study (gwas). *BMC Genomics*, 11(1), Dec. 2010. ISSN 1471-2164. doi: 10.1186/1471-2164-11-724. URL `http://dx.doi.org/10.1186/1471-2164-11-724`.

[16] G. S. Kimeldorf and G. Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, Apr. 1970. ISSN 0003-4851. doi: 10.1214/aoms/1177697089. URL http://dx.doi.org/10.1214/aoms/1177697089.

[17] L. Lei and W. Fithian. Adapt: An interactive procedure for multiple testing with side information, 2018. URL https://arxiv.org/abs/1609.06035.

[18] B. Perets and S. Mannor. Controlling false discovery in arbitrarily structured hypothesis spaces via reproducing kernels, 2026. URL https://arxiv.org/abs/2605.17559.

[19] R. Raju, S. Jain, B. Li, J. Li, and U. Thakker. Constructing domain-specific evaluation sets for llm-as-a-judge, 2024. URL https://arxiv.org/abs/2408.08808.

[20] A. Ramdas, T. Zrnic, M. Wainwright, and M. Jordan. SAFFRON: An adaptive algorithm for online control of the false discovery rate. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4286–4294. PMLR, 2018. URL https://proceedings.mlr.press/v80/ramdas18a.html.

[21] W. Sun and T. T. Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912, 2007. ISSN 01621459. URL http://www.jstor.org/stable/27639933.

[22] W. Tansey, O. Koyejo, R. A. Poldrack, and J. G. Scott. False discovery rate smoothing. 2016. URL https://arxiv.org/abs/1411.6144.

[23] P. Wang, P. Yan, and C. Li. Straw: Structure-adaptive weighting procedure for large-scale spatial multiple testing. 2023. URL https://arxiv.org/abs/2309.15699.

[24] M. J. Zhang, J. Zou, and D. Tse. Adaptive monte carlo multiple testing via multi-armed bandits, 2019. URL https://arxiv.org/abs/1902.00197.

A Assumptions used throughout the appendix

The proofs in this appendix invoke the following four standing assumptions. They are imposed on the data-generating process and the estimator; they are not novel to this paper but are stated here so that downstream theorems can refer to them by label.

Uniform null. The exact discrete-uniform null on the counts ($k_i \mid m_i, H_{0,i} \sim \mathrm{Uniform}\{0, \dots, m_i\}$) is stated as Assumption 4 in App. E (§E.3); the proofs in this appendix invoke it under that label.

Assumption 1 (Bounded kernel and Lipschitz lfdr)**.** *The kernel K is symmetric positive definite, normalized so that $K_{ii} = 1$, and satisfies $\|K\|_{\mathrm{op}} = O(1)$ and $\|\mathbf{k}_i\|_2 = O(1)$ uniformly in i. The lfdr function $\mathrm{lfdr}(\alpha, k, m)$ is Lipschitz in α on the support permitted by Assumption 2.*

Assumption 2 (Bounded mixture residuals)**.** *There exist constants $0 < c_1 \le c_2$ such that for every i, the mixture density $f_i = \hat{\alpha}(\mathrm{loc}_i)P_{0,i} + (1 - \hat{\alpha}(\mathrm{loc}_i))P_{1,i} \ge c_1$ and the residual weight $w_i = (P_{1,i} - P_{0,i})/f_i$ satisfies $|w_i| \le c_2$.*

Assumption 3 (Regularization and kernel non-degeneracy)**.** *The regularization parameter satisfies $\lambda \ge \lambda_0 > 0$ for some constant λ_0 independent of N. The kernel design matrix is non-degenerate, $\lambda_{\min}(\boldsymbol{K}) > 0$.*

These assumptions are mild and standard for kernel-regularized M-estimation in finite-sample empirical Bayes, and we verify their relevance for the experiments on a per-dataset basis.

B Natural Gradient Derivation for Beta-Binomial Likelihoods

We provide the complete derivation of the natural gradient for our discrete observation setting, confirming that the kernel cancellation property established in [18] extends to Beta-Binomial likelihoods.

B.1 Setup and Notation

For each hypothesis $i \in \{1, \dots, N\}$, we observe (k_i, m_i) where k_i is the count of extreme statistics among m_i null samples. The null probability function is modeled as:

$$\alpha_i = (\mathbf{K}\mathbf{c})_i = \sum_{j=1}^{N} K_{ij} c_j \tag{10}$$

where $\mathbf{c} \in \mathbb{R}^N$ is the RKHS dual coefficient vector (App. C.4), the values $\boldsymbol{\alpha} = \mathbf{K}\mathbf{c}$ are the null probabilities, and $\mathbf{K}$ is the symmetric positive definite Gram matrix.

The marginal likelihoods under each hypothesis are:

$$P_{0,i} \triangleq P(k_i \mid m_i, H_0) = \frac{1}{m_i + 1} \tag{11}$$

$$P_{1,i} \triangleq P(k_i \mid m_i, H_1) = \binom{m_i}{k_i} \frac{B(k_i + a, m_i - k_i + b)}{B(a, b)} \tag{12}$$

and the mixture likelihood is:

$$f_i \triangleq \alpha_i P_{0,i} + (1 - \alpha_i) P_{1,i} \tag{13}$$

B.2 Loss Function

The complete objective consists of three terms:

$$\mathcal{L}(\mathbf{c}) = \underbrace{-\sum_{i=1}^{N} \log f_i}_{\mathcal{L}_{\text{data}}} + \underbrace{\lambda_{\text{reg}}\, \mathbf{c}^\top \mathbf{K}\mathbf{c}}_{\mathcal{L}_{\text{reg}}} + \underbrace{\lambda_{\text{bound}} \Lambda_{\text{bound}}(\boldsymbol{\alpha})}_{\mathcal{L}_{\text{bound}}} \tag{14}$$

The boundary penalty enforces $\alpha_i \in [0, 1]$:

$$\Lambda_{\text{bound}}(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \left[\max(0, \alpha_i - 1)^2 + \max(0, -\alpha_i)^2\right] \tag{15}$$

This soft constraint is preferred over hard projection or logistic transformations, as it preserves the convexity of the optimization landscape while being sufficient in practice [18].

B.3 Euclidean Gradient

We compute $\nabla_{\mathbf{c}}\mathcal{L}$ by differentiating each term.

Data term. For a single observation i:

$$\frac{\partial}{\partial c_j}\left[-\log f_i\right] = -\frac{1}{f_i} \cdot \frac{\partial f_i}{\partial c_j} \tag{16}$$

Since $f_i = \alpha_i P_{0,i} + (1 - \alpha_i)P_{1,i}$:

$$\frac{\partial f_i}{\partial \alpha_i} = P_{0,i} - P_{1,i} \tag{17}$$

And since $\alpha_i = \sum_k K_{ik} c_k$:

$$\frac{\partial \alpha_i}{\partial c_j} = K_{ij} \tag{18}$$

Combining via chain rule:

$$\frac{\partial}{\partial c_j}\left[-\log f_i\right] = -\frac{P_{0,i} - P_{1,i}}{f_i} \cdot K_{ij} \tag{19}$$

Summing over all observations and defining the residual weight:

$$w_i \triangleq \frac{P_{1,i} - P_{0,i}}{f_i} = \frac{P_{1,i} - P_{0,i}}{\alpha_i P_{0,i} + (1 - \alpha_i)P_{1,i}} \tag{20}$$

we obtain:

$$\nabla_{\mathbf{c}}\mathcal{L}_{\text{data}} = \sum_{i=1}^{N} w_i \mathbf{k}_i = \mathbf{K}^\top \mathbf{w} = \mathbf{K}\mathbf{w} \tag{21}$$

where $\mathbf{k}_i$ denotes the i-th column of $\mathbf{K}$, and the last equality uses symmetry.

Regularization term.

$$\nabla_{\mathbf{c}}\mathcal{L}_{\text{reg}} = \nabla_{\mathbf{c}}\left[\lambda_{\text{reg}} \mathbf{c}^\top \mathbf{K}\mathbf{c}\right] = 2\lambda_{\text{reg}}\mathbf{K}\mathbf{c} \tag{22}$$

Boundary penalty term. First, compute the gradient with respect to $\boldsymbol{\alpha}$:

$$\left[\nabla_{\boldsymbol{\alpha}}\Lambda_{\text{bound}}\right]_i = \begin{cases} 2(\alpha_i - 1) & \text{if } \alpha_i > 1 \\ -2\alpha_i & \text{if } \alpha_i < 0 \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

Then apply the chain rule. Since $\alpha_i = \sum_k K_{ik} c_k$:

$$\nabla_{\mathbf{c}}\mathcal{L}_{\text{bound}} = \lambda_{\text{bound}}\mathbf{K}^\top \nabla_{\boldsymbol{\alpha}}\Lambda_{\text{bound}} = \lambda_{\text{bound}}\mathbf{K}\nabla_{\boldsymbol{\alpha}}\Lambda_{\text{bound}} \tag{24}$$

Combined Euclidean gradient.

$$\begin{aligned} \nabla_{\mathbf{c}}\mathcal{L} &= \mathbf{K}\mathbf{w} + 2\lambda_{\text{reg}}\mathbf{K}\mathbf{c} + \lambda_{\text{bound}}\mathbf{K}\nabla_{\boldsymbol{\alpha}}\Lambda_{\text{bound}} && (25) \\ &= \mathbf{K}\left(\mathbf{w} + 2\lambda_{\text{reg}}\mathbf{c} + \lambda_{\text{bound}}\nabla_{\boldsymbol{\alpha}}\Lambda_{\text{bound}}\right) && (26) \end{aligned}$$

This reveals that the Gram matrix $\mathbf{K}$ factors out completely from all three terms.

B.4 Natural Gradient and Kernel Cancellation

The natural gradient preconditions by the inverse metric tensor $\mathbf{K}^{-1}$:

$$\tilde{\nabla}_{\boldsymbol{\alpha}}\mathcal{L} = \mathbf{K}^{-1}\nabla_{\boldsymbol{\alpha}}\mathcal{L} \tag{27}$$

Substituting the factored form from Equation (26):

$$\begin{aligned}\tilde{\nabla}_{\mathbf{c}}\mathcal{L} &= \mathbf{K}^{-1}\mathbf{K}\left(\mathbf{w} + 2\lambda_{\text{reg}}\mathbf{c} + \lambda_{\text{bound}}\nabla_{\boldsymbol{\alpha}}\Lambda_{\text{bound}}\right) &(28)\\ &= \mathbf{w} + 2\lambda_{\text{reg}}\mathbf{c} + \lambda_{\text{bound}}\nabla_{\boldsymbol{\alpha}}\Lambda_{\text{bound}} &(29)\end{aligned}$$

Lemma 1. *For the Beta-Binomial spatial FDR objective with boundary constraints, the natural gradient is:*

$$\boxed{\tilde{\nabla}_{\mathbf{c}}\mathcal{L} = \mathbf{w} + 2\lambda_{\text{reg}}\mathbf{c} + \lambda_{\text{bound}}\nabla_{\boldsymbol{\alpha}}\Lambda_{\text{bound}}} \tag{30}$$

where:

- $w_i = (P_{1,i} - P_{0,i})/f_i$ *is the residual weight*
- $f_i = \alpha_i P_{0,i} + (1-\alpha_i)P_{1,i}$ *is the mixture likelihood*
- $[\nabla_{\boldsymbol{\alpha}}\Lambda_{\text{bound}}]_i = 2\max(0, \alpha_i - 1) - 2\max(0, -\alpha_i)$

B.5 Update Rule

The natural gradient descent update is:

$$\mathbf{c}^{(t+1)} = \mathbf{c}^{(t)} - \eta\left(\mathbf{w}^{(t)} + 2\lambda_{\text{reg}}\mathbf{c}^{(t)} + \lambda_{\text{bound}}\nabla_{\boldsymbol{\alpha}}\Lambda^{(t)}_{\text{bound}}\right) \tag{31}$$

The complete algorithm is:

Algorithm 1 Natural Gradient Descent for Beta-Binomial Spatial FDR

Require: Kernel matrix $\mathbf{K}$, observations $\{(k_i, m_i)\}_{i=1}^N$, parameters $\lambda_{\text{reg}}, \lambda_{\text{bound}}, \eta$

1: Initialize coefficients $\mathbf{c}^{(0)}$
2: **for** $t = 0, 1, 2, \ldots$ until convergence **do**
3: $\quad\boldsymbol{\alpha}^{(t)} \leftarrow \mathbf{K}\mathbf{c}^{(t)}$ {Forward pass (values): $O(N^2)$}
4: $\quad$Compute $P_{0,i} = 1/(m_i+1)$ and $P_{1,i} = \text{BetaBinom}(k_i; m_i, a, b)$ for all i
5: $\quad f_i \leftarrow \alpha_i^{(t)} P_{0,i} + (1-\alpha_i^{(t)})P_{1,i}$ for all i
6: $\quad w_i \leftarrow (P_{1,i} - P_{0,i})/f_i$ for all i
7: $\quad g_i \leftarrow 2\max(0, \alpha_i^{(t)} - 1) - 2\max(0, -\alpha_i^{(t)})$ for all i {Boundary gradient}
8: $\quad\tilde{\nabla} \leftarrow \mathbf{w} + 2\lambda_{\text{reg}}\mathbf{c}^{(t)} + \lambda_{\text{bound}}\mathbf{g}$ {Natural gradient (coeff space)}
9: $\quad\mathbf{c}^{(t+1)} \leftarrow \mathbf{c}^{(t)} - \eta\,\tilde{\nabla}$
10: **end for**
11: **return** $\hat{\mathbf{c}}$ and values $\hat{\boldsymbol{\alpha}} = \mathbf{K}\hat{\mathbf{c}}$

The per-iteration complexity is $O(N^2)$, dominated by the matrix-vector product in line 3.

C Convergences Proofs

C.1 Proof of lFDR Convergence

Theorem 4 (Rate of Convergence)**.** *Let $f_1 = \text{Beta}(a, b)$ with $a < 1 < b$, and treat (α, a, b) as known. Write $\hat{p} = (k+1)/(m+1)$ and $lfdr_{class}(p) = \alpha/(\alpha + (1-\alpha)f_1(p))$. There is a constant C depending only on (α, a, b) such that for every $m \geq 1$ and $0 \leq k \leq m-1$,*

$$\left|lfdr_{marg}(k, m) - lfdr_{class}(\hat{p})\right| \;\leq\; C\Big(\frac{1}{k+1} + \frac{1}{m-k}\Big)\min\Big\{1, \frac{\alpha}{(1-\alpha)f_1(\hat{p})}\Big\}. \tag{32}$$

In particular: (i) on $\hat{p} \in [\delta, 1-\delta]$ the error is $O(1/(\delta m))$; (ii) at a fixed count k near the lower edge it is $O\big((k+1)^{-a} m^{a-1}\big)$, and at $k = 0$ it is $\Theta(m^{a-1})$ (Remark 1). At the upper edge $k = m$, both lfdrs tend to 1, with $\big|\mathit{lfdr}_{\mathit{marg}}(m, m) - \mathit{lfdr}_{\mathit{class}}(1)\big| = \Theta(m^{1-b})$.

Proof. The primary objective of this theorem is to quantify the approximation error introduced by working in a discrete space rather than an continuous p-value space. The lfdr is a monotonic function of the likelihood ratio of the alternative distribution to the null distribution: $\text{lfdr}(\cdot) = \frac{\alpha_0}{\alpha_0 + (1-\alpha_0)\Lambda}$ where Λ represents the likelihood engine of the decision rule. Therefore, proving that the discrete marginalized local FDR ($\text{lfdr}_{\text{marg}}$) converges to the classical continuous local FDR ($\text{lfdr}_{\text{class}}$) reduces to proving that the discrete, count-space likelihood ratio $\Lambda_m = P_1(k, m)/P_0(k, m)$ converges pointwise to the continuous alternative density function $f_1(\hat{p})$. The proof strategy is mapped out in three distinct phases: (1) Utilizing Beta-Binomial conjugacy, we express the discrete likelihood ratio Λ_m exactly for any finite count k and sample budget m. By transforming the binomial coefficients and Beta functions into their Gamma function (Γ) equivalents, the discrete ratio is represented as a well-defined, finite product of factors. (2) Notice, that the paper follow the Laplace-smoothed empirical ratio $p = (1+k)/(1+m)$, hence we change variables to the total sample scale $M = m+1$ and $\hat{p} = (k+1)/M$. Instead of approximating continuous slopes, we apply Stirling's asymptotic quotient expansion directly to the discrete Gamma function ratios. A uniform form of this expansion controls the lattice remainder by $O(1/(k+1) + 1/(m-k))$ at every count $0 \le k \le m-1$, which is $O(1/(\delta M))$ on $\hat{p} \in [\delta, 1-\delta]$. (3) Upon multiplying the expanded Gamma quotients, the sample scale parameter M cancels out from the primary exponent via exact linear combination. The remaining terms structurally reassemble into the explicit algebraic definition of the continuous Beta density evaluated at the empirical mean, yielding $\Lambda_m = f_1(\hat{p})(1+\varepsilon)$ with ε controlled by that remainder. **Finally**, the exact identity $\frac{\alpha}{\alpha+Y_1} - \frac{\alpha}{\alpha+Y_2} = \frac{\alpha(Y_2 - Y_1)}{(\alpha+Y_1)(\alpha+Y_2)}$ transfers the likelihood-level rate to the local FDR statistic without any further expansion, explicitly bounding the finite-resource discretization error.

The likelihood ratio:

$$\Lambda_m = \frac{P_1(k, m)}{P_0(k, m)} = (m+1) \cdot P_1(k, m) \tag{33}$$

since $P_0(k, m) = \frac{1}{m+1}$ under the discrete uniform null model.

We list the four independent structural algebraic components before combination:

(i) The binomial coefficient from the combination count:

$$\binom{m}{k} = \frac{m!}{k!(m-k)!} = \frac{\Gamma(m+1)}{\Gamma(k+1)\Gamma(m-k+1)} \tag{34}$$

(ii) The non-normalized Beta integral component of the alternative distribution:

$$B(k+a, m-k+b) = \frac{\Gamma(k+a)\Gamma(m-k+b)}{\Gamma(m+a+b)} \tag{35}$$

(iii) The standard normalizing constant of the Beta distribution:

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{36}$$

This remains as $B(a, b)$ throughout the proof since it does not depend on m or k, acting as a fixed multiplicative constant factor.

(iv) The inverse null likelihood prefactor scaling factor:

$$m+1 = \frac{\Gamma(m+2)}{\Gamma(m+1)} \tag{37}$$

Combining the structural terms (i), (ii), and (iii) yields the complete Beta-Binomial marginal pmf:

$$\begin{aligned} P_1(k, m) &= \binom{m}{k} \cdot \frac{B(k+a, m-k+b)}{B(a, b)} \\ &= \frac{\Gamma(m+1)}{\Gamma(k+1)\Gamma(m-k+1)} \cdot \frac{1}{B(a, b)} \cdot \frac{\Gamma(k+a)\Gamma(m-k+b)}{\Gamma(m+a+b)} \end{aligned} \tag{38}$$

Grouping the terms dynamically by their asymptotic base variables splits the equation into three isolated tail tracks:

$$P_1(k,m) = \frac{1}{B(a,b)} \cdot \frac{\Gamma(m+1)}{\Gamma(m+a+b)} \cdot \frac{\Gamma(k+a)}{\Gamma(k+1)} \cdot \frac{\Gamma(m-k+b)}{\Gamma(m-k+1)} \tag{39}$$

Forming the complete ratio $\Lambda_m = (m+1) \cdot P_1(k,m)$, the scaling fraction $\frac{\Gamma(m+2)}{\Gamma(m+1)}$ multiplies directly into the global m-dependent structural term:

$$\frac{\Gamma(m+2)}{\Gamma(m+1)} \cdot \frac{\Gamma(m+1)}{\Gamma(m+a+b)} = \frac{\Gamma(m+2)}{\Gamma(m+a+b)} \tag{40}$$

This isolates the likelihood ratio completely into three decoupled ratios of Gamma functions:

$$\Lambda_m = \frac{1}{B(a,b)} \cdot \underbrace{\frac{\Gamma(k+a)}{\Gamma(k+1)}}_{=:T_1} \cdot \underbrace{\frac{\Gamma(m-k+b)}{\Gamma(m-k+1)}}_{=:T_2} \cdot \underbrace{\frac{\Gamma(m+2)}{\Gamma(m+a+b)}}_{=:T_3} \tag{41}$$

Let $M = m+1$ and $\hat{p} = \frac{k+1}{M}$. This change of variables yields the exact relationships:

- $k+1 = \hat{p}M \implies k = \hat{p}M - 1$
- $m-k = (m+1)-(k+1) = M - \hat{p}M = (1-\hat{p})M$
- $m+2 = M+1$
- $m+a+b = M+a+b-1$

To evaluate the asymptotic behavior of these factorials as $M \to \infty$, we rewrite each quotient T_j into the canonical form $\frac{\Gamma(z+c)}{\Gamma(z+d)}$, where z represents a large, growing base variable, while c and d serve as small, fixed parameter shifts.

$$T_1 = \frac{\Gamma(k+a)}{\Gamma(k+1)} = \frac{\Gamma((k+1)+a-1)}{\Gamma((k+1)+0)} = \frac{\Gamma(\hat{p}M+a-1)}{\Gamma(\hat{p}M)} \tag{42}$$

$$T_2 = \frac{\Gamma(m-k+b)}{\Gamma(m-k+1)} = \frac{\Gamma((m-k)+b)}{\Gamma((m-k)+1)} = \frac{\Gamma((1-\hat{p})M+b)}{\Gamma((1-\hat{p})M+1)} \tag{43}$$

$$T_3 = \frac{\Gamma(m+2)}{\Gamma(m+a+b)} = \frac{\Gamma(M+1)}{\Gamma(M+a+b-1)} \tag{44}$$

We use the uniform form of Stirling's bound for Gamma quotients. For fixed real c, let $\varphi_c(z) = \Gamma(z+c)/(\Gamma(z)\, z^c)$. It is continuous and positive on $[1,\infty)$ with $\varphi_c(z) = 1 + c(c-1)/(2z) + O(z^{-2})$, so there is C_c with $|\varphi_c(z) - 1| \le C_c/z$ for all $z \ge 1$, and φ_c is bounded above and away from zero on $[1,\infty)$. For $0 \le k \le m-1$ all three base variables are at least 1: $\hat{p}M = k+1$, $(1-\hat{p})M = m-k$ and $M = m+1$. Hence, exactly,

$$T_1 = (\hat{p}M)^{a-1}\, \varphi_{a-1}(\hat{p}M), \tag{45}$$

$$T_2 = \frac{\Gamma((1-\hat{p})M+b)}{(1-\hat{p})M\, \Gamma((1-\hat{p})M)} = ((1-\hat{p})M)^{b-1}\, \varphi_b((1-\hat{p})M), \tag{46}$$

$$T_3 = \frac{M\,\Gamma(M)}{\Gamma(M+a+b-1)} = \frac{M^{2-a-b}}{\varphi_{a+b-1}(M)}. \tag{47}$$

Multiplying, the powers of M cancel, $(a-1)+(b-1)+(2-a-b) = 0$, and the remaining factors reassemble the continuous density:

$$\Lambda_m = f_1(\hat{p})\,(1+\varepsilon), \qquad 1+\varepsilon = \frac{\varphi_{a-1}(k+1)\,\varphi_b(m-k)}{\varphi_{a+b-1}(m+1)}. \tag{48}$$

Since each factor is within C/z of 1 and bounded away from zero,

$$|\varepsilon| \le C_\varepsilon \Big(\frac{1}{k+1} + \frac{1}{m-k} \Big), \qquad 1+\varepsilon \ge c_0 > 0, \tag{49}$$

with C_ε, c_0 depending only on (a, b); the $1/(m+1)$ term is absorbed since $m+1 > k+1$. The only non-algebraic step is this Stirling bound, and it acts on Gamma functions, not on f_1.

Transfer to the lfdr. With $Y_1 = (1-\alpha)\Lambda_m$ and $Y_2 = (1-\alpha)f_1(\hat{p})$, the exact identity gives

$$\text{lfdr}_{\text{marg}}(k, m) - \text{lfdr}_{\text{class}}(\hat{p}) = \frac{\alpha(Y_2 - Y_1)}{(\alpha + Y_1)(\alpha + Y_2)} = \frac{-\alpha(1-\alpha)f_1(\hat{p})\,\varepsilon}{(\alpha + Y_1)(\alpha + Y_2)}. \tag{50}$$

Bounding the denominator below by α^2 gives $|\varepsilon|(1-\alpha)f_1(\hat{p})/\alpha$; bounding it below by $Y_1 Y_2 \geq c_0(1-\alpha)^2 f_1(\hat{p})^2$ gives $|\varepsilon|\,\alpha/(c_0(1-\alpha)f_1(\hat{p}))$. The smaller of the two is at most $\max(1, c_0^{-1})\,|\varepsilon| \min\{1, \alpha/((1-\alpha)f_1(\hat{p}))\}$, which with (49) is (32).

Special cases. On $\hat{p} \in [\delta, 1-\delta]$, $k+1 \geq \delta M$ and $m-k \geq \delta M$, so the bound is $O(1/(\delta m))$. Near the lower edge ($\hat{p} \leq 1/2$), $f_1(\hat{p}) \geq 2^{1-b}\hat{p}^{a-1}/B(a,b)$ and $1/(m-k) \leq 1/(k+1)$, so the bound is $O\big((k+1)^{-1}\hat{p}^{1-a}\big) = O\big((k+1)^{-a}m^{a-1}\big)$; at $k=0$ it is $O(m^{a-1})$, attained by Remark 1.

Upper edge. At $k = m$ the base variable of T_2 is 0 and $f_1(1) = 0$ for $b > 1$, so $\text{lfdr}_{\text{class}}(1) = 1$. Directly, $\Lambda_m = (m+1)B(m+a, b)/B(a, b) = \Theta(m^{1-b}) \to 0$, so $1 - \text{lfdr}_{\text{marg}}(m, m) = (1-\alpha)\Lambda_m/(\alpha + (1-\alpha)\Lambda_m) = \Theta(m^{1-b})$: both lfdrs tend to 1, and neither rejects. □

Remark 1 (Boundary Behavior and LFDR Saturation)**.** *The lower-edge rate of Theorem 4 is attained at $k = 0$, and there the likelihoods themselves do not converge. The base variable of the first Gamma quotient stays at $\hat{p}M = 1$ for every m, so the factor $\varphi_{a-1}(1)$ in* (48) *does not tend to* 1*. Directly,*

$$T_1\big|_{k=0} \;=\; \frac{\Gamma(0+a)}{\Gamma(0+1)} \;=\; \Gamma(a), \tag{51}$$

a fixed constant in m. Applying Stirling to the remaining factors T_2 and T_3 (whose base variables, $(1-\hat{p})M = m$ and $M = m+1$, do tend to infinity) gives

$$\Lambda_m\big|_{k=0} \;=\; \frac{\Gamma(a)}{B(a,b)} \cdot M^{1-a} \cdot \left(1 + O\left(\frac{1}{m}\right)\right), \tag{52}$$

whereas the continuous density at $\hat{p}_{\min} = 1/M$ evaluates to

$$f_1(\hat{p}_{\min}) \;=\; \frac{1}{B(a,b)} \cdot M^{1-a} \cdot \left(1 - \frac{1}{M}\right)^{b-1} \;=\; \frac{1}{B(a,b)} \cdot M^{1-a} \cdot \left(1 + O\left(\frac{1}{m}\right)\right). \tag{53}$$

The ratio $\Lambda_m / f_1(\hat{p}_{\min})$ therefore tends to $\Gamma(a)$ rather than to 1*:*

$$\lim_{m\to\infty} \frac{\Lambda_m\big|_{k=0}}{f_1(\hat{p}_{\min})} \;=\; \Gamma(a), \tag{54}$$

a persistent multiplicative gap that does not vanish. For $a = 1$ (uniform alternative), $\Gamma(a) = 1$ and the gap closes; for $a \neq 1$, the likelihood-level convergence fails at the lattice boundary by a constant factor.

The lfdr-level error nevertheless vanishes, but through a different mechanism. In the discovery regime $a < 1$, both Λ_m and $f_1(\hat{p}_{\min})$ diverge as $M^{1-a} \to \infty$, so both the discrete and continuous local FDR formulas

$$\textit{lfdr}_{\textit{marg}}(k=0, m) \;=\; \frac{\alpha}{\alpha + (1-\alpha)\Lambda_m}, \qquad \textit{lfdr}_{\textit{class}}(\hat{p}_{\min}) \;=\; \frac{\alpha}{\alpha + (1-\alpha)f_1(\hat{p}_{\min})}, \tag{55}$$

saturate to zero. By the exact identity of the proof, with $Y_1 = (1-\alpha)\Lambda_m$ and $Y_2 = (1-\alpha)f_1(\hat{p}_{\min})$, both of order M^{1-a} and $Y_1/Y_2 \to \Gamma(a)$,

$$\big|\textit{lfdr}_{\textit{marg}}(0, m) - \textit{lfdr}_{\textit{class}}(\hat{p}_{\min})\big| = \frac{\alpha|Y_1 - Y_2|}{(\alpha + Y_1)(\alpha + Y_2)} = \frac{\alpha\,|\Gamma(a) - 1|}{(1-\alpha)\,\Gamma(a)\,f_1(\hat{p}_{\min})}\,(1 + o(1)) = \Theta\big(m^{a-1}\big)\,, \tag{56}$$

since $\Gamma(a) \neq 1$ for $a \in (0, 1)$. Thus, while the likelihoods do not align at the boundary, the lfdr remains consistent — not via convergence of Λ_m to $f_1(\hat{p})$, but via saturation of both lfdrs at zero — and the rate $\Theta(m^{a-1})$ in Theorem 4 is attained rather than merely bounded.

Lattice Discretization vs. Continuous Sampling Noise A central motivation for the discrete Beta-Binomial framework is the favorable scaling of its inherent numerical error compared to classical continuous estimation. Here we contrast the exact discretization error of Theorem 4 with the statistical sampling noise inherent to continuous p-value frameworks.

The continuous baseline ($O(1/\sqrt{m})$). Classical continuous FDR control requires a smooth estimate of $f_1(p)$ from the same data used to compute the test statistic. Whether the estimate is parametric or nonparametric, the central limit theorem fixes the estimation error at $O(1/\sqrt{m})$ for the parametric case and slower for the nonparametric case. This noise enters the lfdr formula directly and cannot be removed by computational care; it is a fundamental property of having only m samples to estimate a continuous density.

The interior lattice advantage ($O(1/m)$). The discrete framework bypasses density estimation: $P_0(k, m)$ and $P_1(k, m; b)$ are evaluated in closed form for any (k, m, b), with the only error being the geometric mismatch between the discrete count lattice and the continuous p-value space. Theorem 4 shows this discretization error scales as $O(1/m)$ uniformly on $[\delta, 1-\delta]$ with $O(1/m)$ decays strictly faster than $O(1/\sqrt{m})$. Hence, there is no sampling-noise floor to contend with. **However**, follow the remark, it is worth noticing that at $k = 0$, the rate degrades to $O(m^{a-1})$ (Remark 1). Comparing exponents:

$$O(m^{a-1}) \prec O(m^{-1/2}) \iff a - 1 < -\tfrac{1}{2} \iff a < \tfrac{1}{2}. \tag{57}$$

The discrete framework's boundary rate is faster than the continuous sampling-noise floor precisely when the alternative density has a sharp spike near zero.

C.2 Proof of Theorem: Marginal Likelihood Factorization

Theorem 5 (Marginal Likelihood Factorization)**.** *Under the generative model, when locations are sampled from a distribution μ over $\mathcal{X}$, the marginal likelihood for observation (k_i, m_i) follows a standard two-group mixture:*

$$P(k_i, m_i) = \bar{\alpha} \cdot P(k_i|m_i, H_0) + (1 - \bar{\alpha}) \cdot P(k_i|m_i, H_1; b) \tag{58}$$

where $\bar{\alpha} = \mathbb{E}_\mu[\alpha(loc)]$ is the spatial average of the null probability.

Proof. The joint distribution of the observation and location is:

$$P(k_i, m_i, \text{loc}_i) = P(k_i, m_i|\text{loc}_i) \cdot \mu(\text{loc}_i) \tag{59}$$

where the conditional likelihood at location loc_i is given by the discrete mixture:

$$\begin{aligned} P(k_i, m_i|\text{loc}_i) = \alpha(\text{loc}_i) P(k_i|m_i, H_0) \\ + (1 - \alpha(\text{loc}_i)) P(k_i|m_i, H_1; b) \end{aligned} \tag{60}$$

Marginalizing over locations:

$$\begin{aligned} P(k_i, m_i) &= \int_{\mathcal{X}} P(k_i, m_i|\text{loc})\, d\mu(\text{loc}) \\ &= \int_{\mathcal{X}} \big[\alpha(\text{loc}) P(k_i|m_i, H_0) \\ &\qquad + (1 - \alpha(\text{loc})) P(k_i|m_i, H_1; b)\big]\, d\mu(\text{loc}) \end{aligned} \tag{61}$$

Since the marginal likelihoods $P(k_i|m_i, H_0)$ and $P(k_i|m_i, H_1; b)$ do not depend on location, we can factor them out:

$$\begin{aligned} P(k_i, m_i) = P(k_i|m_i, H_0) \int_{\mathcal{X}} \alpha(\text{loc})\, d\mu(\text{loc}) \\ + P(k_i|m_i, H_1; b) \int_{\mathcal{X}} (1 - \alpha(\text{loc}))\, d\mu(\text{loc}) \end{aligned} \tag{62}$$

Define the spatial average null probability:

$$\bar{\alpha} = \int_{\mathcal{X}} \alpha(\text{loc})\, d\mu(\text{loc}) \tag{63}$$

Since μ is a probability measure, $\int_{\mathcal{X}} d\mu(\text{loc}) = 1$, we have:

$$\int_{\mathcal{X}} (1 - \alpha(\text{loc}))\, d\mu(\text{loc}) = 1 - \bar{\alpha} \tag{64}$$

Therefore:

$$\begin{aligned} P(k_i, m_i) &= P(k_i|m_i, H_0) \cdot \bar{\alpha} + P(k_i|m_i, H_1; b) \cdot (1 - \bar{\alpha}) \\ &= \bar{\alpha} \cdot P(k_i|m_i, H_0) + (1 - \bar{\alpha}) \cdot P(k_i|m_i, H_1; b) \end{aligned} \tag{65}$$

This is exactly the form of a two-group mixture with global mixing proportion $\bar{\alpha}$, proving the theorem. □

C.3 Parameter Estimation Implementation Details

Subset Selection Criteria. We select hypotheses for parameter estimation based on:

1. Sample size threshold: $m_i \geq m_{\min}$ (typically $m_{\min} = 100$)
2. Upper quantile: Take the top 75% by sample size
3. Minimum subset size: At least $N_{\min} = 50$ hypotheses

This ensures the subset has sufficient statistical power while avoiding hypotheses with high measurement uncertainty.

C.4 Derivation of the Hessian

To ensure rigorous calculus, we must distinguish between the RKHS dual coefficient vector $\mathbf{c} \in \mathbb{R}^N$ and the resulting function evaluations (the null probabilities) $\boldsymbol{\alpha} \in \mathbb{R}^N$. By the representer theorem, the function evaluations are given by $\boldsymbol{\alpha} = K\mathbf{c}$, meaning $\alpha_i = (K\mathbf{c})_i$.

In the main text, the spatial penalty is centered: $\|\mathbf{c} - \hat{\bar{c}}\|^2_{\mathcal{H}_K}$. In this appendix we work with the uncentered penalty $\mathbf{c}^\top K\mathbf{c}$ via the constant shift; the Hessian is unchanged. The objective function with respect to the coefficients $\mathbf{c}$ is:

$$L(\mathbf{c}) = -\sum_{i=1}^{N} \log[\alpha_i P_{0,i} + (1 - \alpha_i) P_{1,i}] + \lambda \mathbf{c}^T K \mathbf{c} + \gamma \Lambda_{\text{bound}}(\boldsymbol{\alpha})$$

Data Term. Define the local loss at node i as $\ell_i(\alpha_i) = -\log[\alpha_i P_{0,i} + (1 - \alpha_i) P_{1,i}]$. The first derivative with respect to the output probability is:

$$\frac{\partial \ell_i}{\partial \alpha_i} = -\frac{P_{0,i} - P_{1,i}}{\alpha_i P_{0,i} + (1 - \alpha_i) P_{1,i}}$$

The second derivative gives the local observed Fisher information with respect to the probability:

$$v_i := \frac{\partial^2 \ell_i}{\partial \alpha_i^2} = \frac{(P_{0,i} - P_{1,i})^2}{[\alpha_i P_{0,i} + (1 - \alpha_i) P_{1,i}]^2}$$

To find the Hessian with respect to the coefficients $\mathbf{c}$, we apply the chain rule. Since $\alpha_i = \sum_m K_{im} c_m$, we have the Jacobian $\frac{\partial \alpha_i}{\partial c_j} = K_{ij}$. By the multivariate chain rule:

$$\frac{\partial^2}{\partial c_j \partial c_k} \left(-\sum_i \ell_i \right) = \sum_i v_i K_{ij} K_{ik} = [K^T \text{diag}(v_i) K]_{jk}$$

Regularization Term. The quadratic penalty yields:

$$\nabla^2_{\mathbf{c}}(\lambda \mathbf{c}^T K \mathbf{c}) = 2\lambda K$$

Boundary Penalty Term. The penalty $\Lambda_{\text{bound}}(\boldsymbol{\alpha}) = \sum_i [\max(0, \alpha_i - 1)^2 + \max(0, -\alpha_i)^2]$ contributes $\gamma K^T (\nabla^2_{\boldsymbol{\alpha}} \Lambda_{\text{bound}}) K$, where the inner matrix is diagonal and nonzero only when constraints are active.

Full Hessian. Combining all terms, the Hessian with respect to the RKHS coefficients is:

$$H(\mathbf{c}) = K^T \text{diag}(v_i) K + 2\lambda K + \gamma K^T (\nabla^2_{\boldsymbol{\alpha}} \Lambda_{\text{bound}}) K \tag{66}$$

Positive Definiteness: Since $v_i > 0$ for all i (the denominator is a valid probability mixture), and the kernel matrix K is positive definite, the Hessian $H(\mathbf{c}) \succ 0$, confirming strict convexity of the objective.

Interpretation of σ_i^2. We use $\sigma_i^2 = \mathbf{k}_i^\top H^{-1} \mathbf{k}_i$ as the working variance of $\hat{\alpha}(\text{loc}_i)$. This is the inverse penalized-information (Laplace) variance, not the M-estimator sandwich covariance; since $H = K^\top \text{diag}(v) K + 2\lambda K \succeq K^\top \text{diag}(v) K$, it *upper-bounds* the sandwich variance under the information equality. The induced intervals are therefore conservative, which is the safe direction for the stopping rule of App. C.5: it can only delay a decision, never trigger one prematurely.

C.5 Stopping rule and decision consistency

Proposition 1 (Stopping Rule Validity)**.** *Assume the budgets $m_i(t)$ at every round are fixed in advance of the counts. Let $\hat{\mathbf{c}}_t$ denote the MLE of the spatial coefficients at time t, and define:*

$$\alpha_{i,t} = (K \hat{\mathbf{c}}_t)_i, \quad \textit{lfdr}_{i,t} = \frac{\alpha_{i,t} P_{0,i,t}}{\alpha_{i,t} P_{0,i,t} + (1 - \alpha_{i,t}) P_{1,i,t}}$$

Let $[\textit{lfdr}_i^{\textit{low}}(t), \textit{lfdr}_i^{\textit{high}}(t)]$ be the union-bound $(1-\gamma)$ confidence interval for $\textit{lfdr}_i$ of Section 5.1.

***For the following** stopping rule, with threshold $\hat{\tau}_q$ (the level of Section 5.1), fixed across the K rounds—computed once and held, not re-estimated per round: At time t, we declare hypothesis i as:*

- ***Definitively rejected** if $\textit{lfdr}_i^{\textit{high}}(t) < \hat{\tau}_q$*
- ***Definitively accepted** if $\textit{lfdr}_i^{\textit{low}}(t) > \hat{\tau}_q$*
- ***Ambiguous** otherwise*

*Run the policy for K allocation rounds, evaluating the joint lfdr CI at each round. **With probability at least** $1 - K\gamma$ **asymptotically** (union bound over the K rounds), $\textit{lfdr}_i$ lies in its joint CI at every round simultaneously; on that event:*

1. *every definitive declaration is correct: whenever the CI lies entirely on one side of $\hat{\tau}_q$, the true $\textit{lfdr}_i$ lies on that side, so the reject/accept decision agrees with the truth;*
2. *a hypothesis declared definitively rejected at one round is never declared definitively accepted at another (a definitive label is never reversed to its opposite).*

Proof. At each round the joint lfdr interval of Section 5.1 is formed by union-bounding the Gaussian (Laplace) α-interval $\hat{\alpha}(\text{loc}_i) \pm z_{1-\gamma/4}\, \sigma_i$ (asymptotic coverage $1 - \gamma/2$ for $\alpha(\text{loc}_i)$) with the Beta–Binomial p-interval (coverage $1 - \gamma/2$), mapped through the lfdr by the monotonicity of Lemma 2 (corner evaluation). Hence at each round t,

$$P\big(\text{lfdr}_i \in [\text{lfdr}_i^{\text{low}}(t), \text{lfdr}_i^{\text{high}}(t)]\big) \geq 1 - \gamma.$$

These intervals are *not* nested across rounds: the α-interval is recentred at the updated $\hat{\mathbf{c}}_t$ and the p-interval at $(k_i + 1)/(m_i + 1)$, both of which move as data accrue, so a later interval need not lie

inside an earlier one. We therefore control the trajectory by a union bound rather than by nesting. Over K rounds,

$$P\Big(\bigcap_{t=1}^{K}\big\{\text{lfdr}_i \in [\text{lfdr}_i^{\text{low}}(t), \text{lfdr}_i^{\text{high}}(t)]\big\}\Big) \geq 1 - K\gamma.$$

On this event lfdr_i lies in the CI at every round. Consequence 1 is immediate: a CI entirely below (resp. above) $\hat{\tau}_q$ forces lfdr_i below (resp. above) $\hat{\tau}_q$. Consequence 2 follows because a definitive rejection at round t and a definitive acceptance at round t' would require $\text{lfdr}_i < \hat{\tau}_q$ and $\text{lfdr}_i > \hat{\tau}_q$ simultaneously, contradicting joint coverage. Running each round at level γ/K restores overall confidence $1 - \gamma$ across the run. □

Under the adaptive policy of Section 5.2 the budgets depend on the counts, the per-round coverage used above is not guaranteed, and the proposition does not apply (App. C.14.5).

C.6 Monotonicity of the Local False Discovery Rate

The combined confidence interval construction relies on the monotonicity of $\text{lfdr}(\alpha, p)$ in both arguments, enabling efficient computation via corner evaluation.

Lemma 2 (Monotonicity of lfdr)**.** *Let $f_0(p) = 1$ (uniform null) and $f_1(p) = p^{a-1}(1-p)^{b-1}/B(a,b)$ with $a \in (0,1)$ and $b > 1$ (the Beta(a,b) alternative of Section 3.1). The local false discovery rate*

$$\textit{lfdr}(\alpha, p) = \frac{\alpha f_0(p)}{\alpha f_0(p) + (1-\alpha) f_1(p)} = \frac{\alpha}{\alpha + (1-\alpha) f_1(p)} \tag{67}$$

is strictly increasing in both α and p for $\alpha \in (0,1)$ and $p \in (0,1)$.

Proof. We verify the partial derivatives are positive.

Monotonicity in α. Taking the partial derivative with respect to α:

$$\frac{\partial \text{lfdr}}{\partial \alpha} = \frac{[\alpha + (1-\alpha) f_1(p)] - \alpha[1 - f_1(p)]}{[\alpha + (1-\alpha) f_1(p)]^2} \tag{68}$$

$$= \frac{\alpha + (1-\alpha) f_1(p) - \alpha + \alpha f_1(p)}{[\alpha + (1-\alpha) f_1(p)]^2} \tag{69}$$

$$= \frac{f_1(p)}{[\alpha + (1-\alpha) f_1(p)]^2} \tag{70}$$

Since $f_1(p) > 0$ for $p \in (0,1)$, we have $\frac{\partial \text{lfdr}}{\partial \alpha} > 0$.

Monotonicity in p. Taking the partial derivative with respect to p:

$$\frac{\partial \text{lfdr}}{\partial p} = \frac{-\alpha(1-\alpha) f_1'(p)}{[\alpha + (1-\alpha) f_1(p)]^2} \tag{71}$$

For the Beta(a,b) alternative density,

$$\frac{f_1'(p)}{f_1(p)} = \frac{a-1}{p} - \frac{b-1}{1-p}. \tag{72}$$

With $a \in (0,1)$ and $b > 1$ both terms on the right are negative on $(0,1)$, so $f_1'(p) < 0$. Therefore

$$\frac{\partial \text{lfdr}}{\partial p} = \frac{-\alpha(1-\alpha)\, f_1'(p)}{[\alpha + (1-\alpha) f_1(p)]^2} > 0. \tag{73}$$

Both partial derivatives are strictly positive, establishing strict monotonicity. □

Corollary 1 (Corner Evaluation)**.** *For any rectangle $[\alpha^{low}, \alpha^{high}] \times [p^{low}, p^{high}]$:*

$$\min_{(\alpha,p)\in rect} \textit{lfdr}(\alpha, p) = \textit{lfdr}(\alpha^{low}, p^{low}) \tag{74}$$

$$\max_{(\alpha,p)\in rect} \textit{lfdr}(\alpha, p) = \textit{lfdr}(\alpha^{high}, p^{high}) \tag{75}$$

This reduces the computation of lfdr confidence bounds from a 2D optimization to two function evaluations.

C.7 Efficient Computation of Width Bounds

Conjugate Gradient for Quadratic Forms. The confidence interval width requires computing $q_i = \mathbf{k}_i^\top H^{-1}\mathbf{k}_i$ for each hypothesis i, where $\mathbf{k}_i$ is the i-th column of the kernel matrix K. Rather than explicitly forming and inverting H, we use Conjugate Gradient (CG) to solve $H\mathbf{x} = \mathbf{k}_i$, then compute $q_i = \mathbf{k}_i^\top \mathbf{x}$.

This approach is numerically stable because: (i) the Hessian $H = K^\top \text{diag}(v)K + 2\lambda K$ is symmetric positive definite, (ii) the regularization term $2\lambda K$ ensures good conditioning with $\kappa(H) = O(1/\lambda)$, and (iii) CG converges in $O(\sqrt{\kappa})$ iterations for well-conditioned systems.

Matrix-Vector Product. Computing $(KH^{-1}K^T)_{ii}$ for all i requires N CG solves. Each solve involves matrix-vector products with H, which decomposes as:

$$H\mathbf{x} = K^\top(\text{diag}(v)(K\mathbf{x})) + 2\lambda K\mathbf{x} \tag{76}$$

Each matrix-vector product costs $O(N^2)$ (dominated by $K\mathbf{x}$), and CG typically converges in $O(\sqrt{1/\lambda})$ iterations. Total cost for all N hypotheses: $O(N^3/\sqrt{\lambda})$ in the worst case, though early termination and warm-starting reduce this substantially in practice.

Sparse Kernel Optimization. For localized kernels (e.g., compact support or fast decay), we can exploit sparsity:

- Truncate kernel evaluations below threshold ϵ (e.g., 10^{-6})
- Use sparse matrix storage for K
- Cost reduces to $O(Nd)$ where d is average degree

For Matérn kernels with range parameter ℓ, approximately $d = O(1)$ hypotheses contribute significantly to each row.

C.8 Variance Decomposition: Local versus Spatial Components

The allocation policy decomposes uncertainty at each hypothesis into local and spatial components. This decomposition diagnoses whether progress is bottlenecked by the hypothesis itself or by its neighbors, guiding the helper sampling mechanism.

Proposition 2 (Variance Decomposition)**.** *Let $H = K^\top \text{diag}(v)K + 2\lambda K$ be the Hessian of the penalized likelihood, where $v_i = (P_{0,i} - P_{1,i})^2/f_i^2$ is the local Fisher information.The asymptotic variance of $\alpha_i = (K\mathbf{c})_i$ can be decomposed as:*

$$\textit{Var}[\alpha_i] = (KH^{-1}K^\top)_{ii} = \underbrace{K_{ii}^2 H_{ii}^{-1}}_{\textit{local}} + \underbrace{\sum_{(j,k)\neq(i,i)} K_{ij}K_{ik}H_{jk}^{-1}}_{\textit{spatial}} \tag{77}$$

Proof. By definition of the asymptotic covariance under the Laplace approximation:

$$\text{Cov}[\mathbf{c}] = H^{-1}, \quad \text{Cov}[\boldsymbol{\alpha}] = KH^{-1}K^\top \tag{78}$$

The variance of α_i is the (i,i) diagonal element:

$$\text{Var}[\alpha_i] = (KH^{-1}K^\top)_{ii} \tag{79}$$

$$= \sum_{j=1}^{N}\sum_{k=1}^{N} K_{ij}H_{jk}^{-1}K_{ki} \tag{80}$$

$$= \sum_{j=1}^{N}\sum_{k=1}^{N} K_{ij}K_{ik}H_{jk}^{-1} \tag{81}$$

where we used $K_{ki} = K_{ik}$ by symmetry of the kernel matrix. Separating the $j = i$ and $k = i$ terms:

$$\text{Var}[\alpha_i] = K_{ii}^2 H_{ii}^{-1} + \sum_{j \neq i} K_{ij} K_{ii} H_{ji}^{-1} + \sum_{k \neq i} K_{ii} K_{ik} H_{ik}^{-1} \tag{82}$$

$$+ \sum_{j \neq i} \sum_{k \neq i} K_{ij} K_{ik} H_{jk}^{-1} \tag{83}$$

For normalized kernels where $K_{ii} = 1$, the middle terms involve cross-covariances between hypothesis i and its neighbors. The proposition's decomposition is exact: the local term is $K_{ii}^2 H_{ii}^{-1}$ and the spatial term collects every remaining entry, $\sum_{(j,k) \neq (i,i)} K_{ij} K_{ik} H_{jk}^{-1}$, i.e. the double sum plus the two cross sums $2 \sum_{j \neq i} K_{ij} H_{ji}^{-1}$ (equal by symmetry of K and H^{-1}, with $K_{ii} = 1$). In the regularization-dominated regime (see Section C.9) the cross-covariances are small relative to the diagonal contributions, so the spatial term is dominated by the double sum:

$$\sum_{(j,k) \neq (i,i)} K_{ij} K_{ik} H_{jk}^{-1} \;\approx\; \sum_{j \neq i} \sum_{k \neq i} K_{ij} K_{ik} H_{jk}^{-1}. \tag{84}$$

The exact decomposition follows by defining the local term as $K_{ii}^2 H_{ii}^{-1}$ and absorbing all remaining terms into the spatial component.

□

Interpretation. The spatial variance fraction

$$r_i \;\coloneqq\; 1 - \frac{K_{ii}^2 H_{ii}^{-1}}{(K H^{-1} K^\top)_{ii}}. \tag{85}$$

quantifies how much of the uncertainty at hypothesis i originates from neighbors versus local data. When $r_i \approx 0$, the bottleneck is local: sampling hypothesis i directly will reduce its uncertainty. When $r_i \approx 1$, the bottleneck is spatial: the hypothesis has sufficient local information, but uncertainty propagates from poorly-characterized neighbors. In this case, sampling i yields diminishing returns, and the allocation policy should target the neighbors instead.

C.9 Regularization Dominance for Isolated Hypotheses

The Hessian-based variance $(K H^{-1} K^\top)_{ii}$ incorporates both local Fisher information and spatial regularization. However, for isolated hypotheses with weak kernel connectivity, the regularization term dominates, causing the variance to become unresponsive to local data accumulation. This section formalizes this phenomenon and motivates the Hessian-Fisher blending mechanism.

Hessian Structure. The Hessian of the penalized log-likelihood is:

$$H = K^\top \text{diag}(v) K + 2\lambda K \tag{86}$$

where $v_i = (P_{0,i} - P_{1,i})^2 / f_i^2$ is the local Fisher information at hypothesis i. The first term captures data-driven curvature; the second is the regularization penalty enforcing spatial smoothness.

Isolated Hypothesis Regime. Consider a hypothesis i with weak connectivity: $K_{ij} \approx 0$ for all $j \neq i$, but $K_{ii} = 1$ (normalized kernel). The kernel matrix has the approximate block structure:

$$K \approx \begin{pmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & K_{-i} \end{pmatrix} \tag{87}$$

where K_{-i} is the kernel matrix for all other hypotheses. In this limit, the Hessian contribution from hypothesis i is:

$$[K^\top \text{diag}(v) K]_{ii} = v_i K_{ii}^2 = v_i \tag{88}$$

and the regularization contribution is:

$$[2\lambda K]_{ii} = 2\lambda K_{ii} = 2\lambda \tag{89}$$

Variance Floor. The variance of α_i becomes:

$$\text{Var}[\alpha_i] = (KH^{-1}K^\top)_{ii} \approx K_{ii}^2 H_{ii}^{-1} = \frac{1}{v_i + 2\lambda} \tag{90}$$

As local data accumulates, v_i increases (more Fisher information). However, the variance is bounded below:

$$\text{Var}[\alpha_i] \geq \frac{1}{v_\infty + 2\lambda} \approx \frac{1}{2\lambda} \tag{91}$$

where $v_\infty = \lim_{m\to\infty} v_i$ is the limiting Fisher information (which is finite; see Section C.11).

The Problem. For well-connected hypotheses, the spatial term in the Hessian couples them to neighbors, allowing information to flow and uncertainty to shrink as the entire neighborhood is sampled. For isolated hypotheses, this coupling is absent. The regularization term $2\lambda K$ dominates the diagonal, creating a variance floor of approximately $1/(2\lambda)$ that cannot be reduced regardless of how much local data is collected.

This is incorrect behavior: an isolated hypothesis should still benefit from local sampling. The Fisher information v_i correctly captures this benefit, but it is overwhelmed by the regularization term in the Hessian.

Solution: Hessian-Fisher Blending. To address this, we blend the Hessian-based variance with the pure Fisher variance $1/v_i$ using an isolation-dependent weight. The blending mechanism, detailed in Section C.10, ensures that:

- Well-connected hypotheses use primarily Hessian-based variance (spatial coupling is informative)
- Isolated hypotheses use primarily Fisher-based variance (local data is all we have)
- The transition is smooth, based on kernel connectivity

C.10 Hessian-Fisher Blending Derivation

Building on the regularization dominance problem identified in Section C.9, we derive the blending mechanism that combines Hessian-based and Fisher-based variance estimates.

Design Criteria. The blended variance should satisfy:

1. **Early samples:** Use Hessian variance (trust the spatial prior when local data is sparse)
2. **Well-connected hypotheses:** Hessian variance remains relevant as it captures spatial information flow
3. **Isolated hypotheses:** Transition to Fisher variance as data accumulates (avoid regularization floor)
4. **Smooth transition:** No discontinuities as sample size or connectivity varies

Isolation Score. We quantify isolation via kernel connectivity. Define the connectivity score:

$$c_i = \sum_{j \neq i} K_{ij} \tag{92}$$

This measures the total kernel weight connecting hypothesis i to all others. The normalized isolation score is:

$$\text{isolation}_i = 1 - \frac{c_i}{\max_k c_k} \tag{93}$$

A hypothesis with $\text{isolation}_i \approx 0$ is well-connected (strong kernel coupling to neighbors). A hypothesis with $\text{isolation}_i \approx 1$ is isolated (minimal kernel coupling).

Blending Weight. The blending weight $w_i \in [0, 1]$ determines the contribution of Hessian-based variance:

$$w_i = \begin{cases} 1 & \text{if } m_i < m_{\text{start}} \\ \dfrac{m_0}{m_0 + m_i - m_{\text{start}}} \cdot (1 - \text{isolation}_i) & \text{otherwise} \end{cases} \tag{94}$$

where:

- m_{start}: Minimum samples before blending begins (ensures stable Fisher estimates)
- m_0: Decay rate parameter (controls transition speed)

Properties of the Blending Weight. Early samples ($m_i < m_{\text{start}}$): $w_i = 1$, so we use pure Hessian variance. This is appropriate because Fisher information estimates are unreliable with few samples, and the spatial prior provides useful regularization.

Well-connected hypotheses (isolation$_i \approx 0$): The weight decays as:

$$w_i \approx \frac{m_0}{m_0 + m_i - m_{\text{start}}} \tag{95}$$

This decreases with m_i but remains positive, allowing gradual incorporation of Fisher information while retaining spatial coupling benefits.

Isolated hypotheses (isolation$_i \approx 1$): The weight becomes:

$$w_i \approx 0 \tag{96}$$

regardless of m_i (once past m_{start}). This yields pure Fisher variance, correctly capturing that isolated hypotheses have no spatial information to exploit.

Blended Variance. The final variance estimate is:

$$\text{Var}_{\text{blend}}[\alpha_i] = w_i \cdot (KH^{-1}K^\top)_{ii} + (1 - w_i) \cdot \frac{1}{v_i} \tag{97}$$

This is a convex combination of the two variance sources, weighted by connectivity and sample size.

Confidence Interval Width. The blended α-interval is the Gaussian (Laplace) interval of Section 5.1, with the blended variance replacing $\sigma_i^2 = \mathbf{k}_i^\top \hat{H}^{-1}\mathbf{k}_i$:

$$[\alpha_i^{\text{low}}, \alpha_i^{\text{high}}] = \hat{\alpha}(\text{loc}_i) \pm z_{1-\gamma/4}\sqrt{\text{Var}_{\text{blend}}[\alpha_i]}, \qquad W_i^{\text{blend}} = 2\, z_{1-\gamma/4}\sqrt{\text{Var}_{\text{blend}}[\alpha_i]}. \tag{98}$$

This width correctly shrinks with local sampling for isolated hypotheses, while maintaining spatial coherence for well-connected ones.

Parameter Selection. In practice, we use:

- $m_{\text{start}} = 300$: Ensures Fisher information is estimated from sufficient data
- $m_0 = 100$: Provides gradual transition over approximately 100-500 additional samples

These values can be tuned based on the specific application and kernel characteristics.

C.11 Hopeless Detection: Terminal Bounds and Uninformative P-values

Some hypotheses are fundamentally unresolvable: regardless of sampling effort, their confidence intervals will never exclude the decision threshold $\hat{\tau}_q$. This section derives the terminal bounds that identify such "hopeless" cases and characterizes the uninformative p-value region where Fisher information vanishes.

Terminal Variance at $m \to \infty$. As the sample size $m_i \to \infty$, two things happen:

1. P-value uncertainty vanishes: $\text{Var}[p_i^*] = O(1/m_i) \to 0$
2. Fisher information saturates: $v_i \to v_\infty(\hat{p}_i)$

For isolated hypotheses with blending weight $w_i \to 0$, the terminal variance is:

$$\text{Var}_\infty[\alpha_i] = \frac{1}{v_\infty(\hat{p}_i)} \tag{99}$$

Limiting Fisher information and the uninformative p-value. As $m_i \to \infty$, the count-space Fisher information $v_i = (P_{0,i} - P_{1,i})^2 / f_i^2$ saturates at a limit $v_\infty(\hat{p}_i) \geq 0$ that vanishes exactly when the continuous densities coincide, $f_0(\hat{p}_i) = f_1(\hat{p}_i)$. Under the model of §3.1 ($f_0 \equiv 1$ on $(0,1)$ and $f_1 = \text{Beta}(a,b)$ strictly decreasing on $(0,1)$ by Lemma 2), this equation has a unique root $\hat{p}_{\text{uninf}} \in (0,1)$ (no closed form; found by 1D root search given (a,b)). Isolated hypotheses with $\hat{p}_i \approx \hat{p}_{\text{uninf}}$ have $v_\infty \to 0$, hence unbounded terminal variance and irreducible uncertainty about α_i.

Terminal Confidence Bounds. For an isolated hypothesis at the limit $m \to \infty$:

$$W_{i,\infty} = 2\, z_{1-\gamma/4} \sqrt{\text{Var}_\infty[\alpha_i]} = \frac{2\, z_{1-\gamma/4}}{\sqrt{v_\infty(\hat{p}_i)}} \tag{100}$$

The terminal bounds on α_i are:

$$\alpha_{i,\infty}^{\text{low}} = \hat{\alpha}_i - W_{i,\infty}/2 \tag{101}$$

$$\alpha_{i,\infty}^{\text{high}} = \hat{\alpha}_i + W_{i,\infty}/2 \tag{102}$$

Since p-value uncertainty vanishes, the terminal lfdr bounds are:

$$\text{lfdr}_{i,\infty}^{\text{low}} = \text{lfdr}(\alpha_{i,\infty}^{\text{low}}, \hat{p}_i) \tag{103}$$

$$\text{lfdr}_{i,\infty}^{\text{high}} = \text{lfdr}(\alpha_{i,\infty}^{\text{high}}, \hat{p}_i) \tag{104}$$

Hopeless Criterion. A hypothesis is deemed hopeless if its terminal confidence interval straddles the decision threshold:

$$\text{lfdr}_{i,\infty}^{\text{low}} < \hat{\tau}_q \quad \text{and} \quad \text{lfdr}_{i,\infty}^{\text{high}} > \hat{\tau}_q \tag{105}$$

Such hypotheses will never be decidable regardless of sampling effort and should be excluded from allocation.

Rescue by Extreme P-values. Importantly, even hypotheses with high isolation can be decidable if their p-value is sufficiently extreme. Consider two cases:

Small $\hat{p}_i$ (strong evidence against null): When $\hat{p}_i \ll \hat{p}_{\text{uninf}}$, the alternative density dominates: $f_1(\hat{p}_i) \gg 1$. This ensures:

$$\text{lfdr}_{i,\infty}^{\text{high}} = \frac{\alpha_{i,\infty}^{\text{high}}}{\alpha_{i,\infty}^{\text{high}} + (1 - \alpha_{i,\infty}^{\text{high}}) f_1(\hat{p}_i)} < \hat{\tau}_q \tag{106}$$

regardless of α uncertainty, guaranteeing rejection.

Large $\hat{p}_i$ (strong evidence for null): When $\hat{p}_i \gg \hat{p}_{\text{uninf}}$, the alternative density vanishes: $f_1(\hat{p}_i) \ll 1$. This ensures:

$$\text{lfdr}_{i,\infty}^{\text{low}} = \frac{\alpha_{i,\infty}^{\text{low}}}{\alpha_{i,\infty}^{\text{low}} + (1 - \alpha_{i,\infty}^{\text{low}}) f_1(\hat{p}_i)} > \hat{\tau}_q \tag{107}$$

guaranteeing acceptance.

Characterization of Hopeless Hypotheses. Combining these observations:

Proposition 3 (Hopeless Characterization)**.** *A hypothesis i is hopeless if and only if:*

1. *It is isolated: isolation$_i \approx 1$ (no spatial information flow)*
2. *Its p-value is uninformative: $\hat{p}_i \approx \hat{p}_{uninf}$ (null and alternative indistinguishable)*
3. *Its current $\hat{\alpha}_i$ places the terminal interval across $\hat{\tau}_q$*

Well-connected hypotheses are never hopeless (spatial borrowing provides information). Isolated hypotheses with extreme p-values are decidable (the p-value alone determines the outcome). Only isolated hypotheses with $\hat{p}_i \approx \hat{p}_{uninf}$ are truly hopeless.

This characterization guides the allocation policy: hopeless hypotheses are excluded from sampling, conserving budget for hypotheses where progress is possible.

C.12 Proof of Proposition: Spatial Information Propagation

Proposition (Spatial Information Propagation)**.** *Let $\Sigma_t = KH_t^{-1}K^T$ denote the asymptotic covariance of α at time t. Suppose we increase the sample size at location i from $m_{i,t}$ to $m_{i,t} + \Delta m_i$, causing the local Fisher information to increase by:*

$$\Delta v_i = v_{i,t+1} - v_{i,t} > 0$$

The reduction in variance at location j is given exactly by:

$$\mathit{Var}[\alpha_j]_t - \mathit{Var}[\alpha_j]_{t+1} = \frac{\Delta v_i}{1 + \Delta v_i [\Sigma_t]_{ii}} \cdot [\Sigma_t]_{ij}^2$$

Under the additional assumptions:

1. *Bounded kernel (Assumption 1): $\|K\|_{\mathrm{op}} = O(1)$ and $\|\mathbf{k}_i\|_2 = O(1)$ uniformly in i; set $\kappa := \max_i \|\mathbf{k}_i\|_2^2 = O(1)$*
2. *Well-conditioned Hessian: $\mu I \preceq H_t \preceq LI$ for $0 < \mu \leq L$*
3. *Small update: $\Delta v_i \leq \delta$ where $\delta \cdot L/\mu^2 < 1/2$*
4. *Regularization-dominated regime: $2\lambda\|K\| \geq \|K^T V_t K\|$*

The first-order approximation holds:

$$\left| \mathit{Var}[\alpha_j]_t - \mathit{Var}[\alpha_j]_{t+1} - \Delta v_i \cdot \left(\frac{K_{ij}}{2\lambda} \right)^2 \right| \leq C_1 \frac{\delta\kappa}{\lambda^3} + C_2 \frac{\delta^2 \kappa^3}{\mu^3}$$

where $C_1, C_2 = O(1)$ are absolute constants. The leading term is $O(\delta/\lambda^3)$, linear in the update δ (the regularization-dominated approximation of Σ); only the Sherman–Morrison remainder is $O(\delta^2)$. In the regularization-dominated regime $\mu = \Theta(\lambda)$, so both denominators may be written as λ^3.

Practical form: *For $\Delta v_i \propto \eta \Delta m_i$ and $W_i^2 \sim K_{ii}/(2\lambda)$:*

$$\mathit{Var}[\alpha_j]_t - \mathit{Var}[\alpha_j]_{t+1} = \gamma \cdot K(\mathit{loc}_i, \mathit{loc}_j)^2 \cdot W_i \cdot \Delta m_i + O(\Delta m_i^2)$$

where $\gamma = \eta/(4\lambda^2)$ depends on the problem parameters.

Proof. We establish the result in four steps: exact formula, first-order expansion, approximation of covariance, and error bounds.

The Hessian update is rank-one:

$$H_{t+1} = H_t + \Delta v_i \cdot \mathbf{k}_i \mathbf{k}_i^T$$

where $\mathbf{k}_i$ is the i-th column of K. By the Sherman-Morrison formula:

$$H_{t+1}^{-1} = H_t^{-1} - \frac{\Delta v_i}{1 + \Delta v_i \mathbf{k}_i^T H_t^{-1} \mathbf{k}_i} \cdot H_t^{-1} \mathbf{k}_i \mathbf{k}_i^T H_t^{-1}$$

Define the shrinkage factor:

$$\rho = \frac{\Delta v_i}{1 + \Delta v_i \mathbf{k}_i^T H_t^{-1} \mathbf{k}_i}$$

The variance at location j is $\text{Var}[\alpha_j]_t = \mathbf{k}_j^T H_t^{-1} \mathbf{k}_j$. The change is:

$$\begin{aligned} \text{Var}[\alpha_j]_t - \text{Var}[\alpha_j]_{t+1} &= \mathbf{k}_j^T (H_t^{-1} - H_{t+1}^{-1}) \mathbf{k}_j \\ &= \rho \cdot \mathbf{k}_j^T H_t^{-1} \mathbf{k}_i \mathbf{k}_i^T H_t^{-1} \mathbf{k}_j \\ &= \rho \cdot (\mathbf{k}_i^T H_t^{-1} \mathbf{k}_j)^2 \end{aligned} \tag{108}$$

Since $\mathbf{k}_i^T H_t^{-1} \mathbf{k}_j = [K H_t^{-1} K^T]_{ij} = [\Sigma_t]_{ij}$:

$$\text{Var}[\alpha_j]_t - \text{Var}[\alpha_j]_{t+1} = \frac{\Delta v_i}{1 + \Delta v_i [\Sigma_t]_{ii}} \cdot [\Sigma_t]_{ij}^2$$

This is exact with no approximation.

For small Δv_i, using $(1 + x)^{-1} = 1 - x + O(x^2)$:

$$\rho = \Delta v_i (1 - \Delta v_i [\Sigma_t]_{ii} + O(\Delta v_i^2)) = \Delta v_i + O(\Delta v_i^2)$$

Therefore:

$$\text{Var}[\alpha_j]_t - \text{Var}[\alpha_j]_{t+1} = \Delta v_i \cdot [\Sigma_t]_{ij}^2 + O(\Delta v_i^2)$$

The remainder term satisfies:

$$|O(\Delta v_i^2)| \leq \Delta v_i^2 \cdot [\Sigma_t]_{ii} \cdot [\Sigma_t]_{ij}^2 \leq \delta^2 \cdot \frac{\kappa}{\mu} \cdot \frac{\kappa^2}{\mu^2} = \frac{\kappa^3 \delta^2}{\mu^3}$$

using the bounds from assumptions 1-2.

The Hessian decomposes as $H_t = 2\lambda K + K^T V_t K$. Under assumption 4 (regularization-dominated), the data term is a small perturbation. Using the Neumann series for $(A + E)^{-1}$ with $A = 2\lambda K$ and $E = K^T V_t K$:

$$H_t^{-1} = (2\lambda K)^{-1} - (2\lambda K)^{-1} E (2\lambda K)^{-1} + O(\|E\|^2)$$

where $\|E\| \leq 2\lambda \|K\|$ by assumption 4. This gives:

$$H_t^{-1} = \frac{1}{2\lambda} K^{-1} + O(\lambda^{-2})$$

Therefore:

$$[\Sigma_t]_{ij} = [K H_t^{-1} K^T]_{ij} = \frac{1}{2\lambda} [K K^{-1} K^T]_{ij} + O(\lambda^{-2}) = \frac{K_{ij}}{2\lambda} + O(\lambda^{-2})$$

The error in $[\Sigma_t]_{ij}^2$ is:

$$\left| [\Sigma_t]_{ij}^2 - \left(\frac{K_{ij}}{2\lambda} \right)^2 \right| = 2 \frac{|K_{ij}|}{2\lambda} \cdot O(\lambda^{-2}) + O(\lambda^{-4}) = O(\kappa / \lambda^3)$$

Combining previous steps:

$$\begin{aligned} &\left| \text{Var}[\alpha_j]_t - \text{Var}[\alpha_j]_{t+1} - \Delta v_i \cdot \left(\frac{K_{ij}}{2\lambda} \right)^2 \right| \\ &\leq \Delta v_i \cdot O(\kappa / \lambda^3) + O(\Delta v_i^2) \\ &\leq \frac{\delta \kappa}{\lambda^3} + \frac{\kappa^3 \delta^2}{\mu^3} \end{aligned} \tag{109}$$

The first term is the $O(\delta)$ leading error (the regularization-dominated approximation $[\Sigma_t]_{ij} \approx K_{ij}/2\lambda$); the second is the genuinely second-order Sherman–Morrison remainder. We do *not* fold the linear term into the quadratic: that would require a "constant" $C = O(1/\delta)$, unbounded as $\delta \to 0$. In the regularization-dominated regime (assumption 4) with $\|K\|_{\text{op}} = O(1)$ one has $\mu = \Theta(\lambda)$, so both terms share the denominator scale λ^3 up to constants, giving the bound stated in the proposition.

For the Beta-Binomial likelihood, $\Delta v_i \approx \eta \Delta m_i$ where η depends on the current mixing proportions. The current uncertainty is $W_i^2 = [\Sigma_t]_{ii} \approx K_{ii}/(2\lambda)$. Substituting:

$$\begin{aligned}
\text{Var}[\alpha_j]_t - \text{Var}[\alpha_j]_{t+1} &\approx \eta \Delta m_i \cdot \frac{K_{ij}^2}{4\lambda^2} \\
&= \frac{\eta K_{ij}^2}{4\lambda^2 K_{ii}} \cdot K_{ii} \cdot \Delta m_i \\
&= \frac{\eta K_{ij}^2}{4\lambda^2 K_{ii}} \cdot 2\lambda W_i^2 \cdot \Delta m_i \\
&= \frac{\eta}{2\lambda K_{ii}} \cdot K_{ij}^2 \cdot W_i^2 \cdot \Delta m_i
\end{aligned} \tag{110}$$

Define $\gamma = \eta/(2\lambda K_{ii})$. For normalized kernels ($K_{ii} \approx 1$), the variance reduction at j is approximately:

$$\Delta \text{Var}[\alpha_j] \approx \gamma \cdot K(\text{loc}_i, \text{loc}_j)^2 \cdot W_i^2 \cdot \Delta m_i$$

Taking derivatives: $\Delta W_j \approx (\gamma/2W_j) K_{ij}^2 W_i^2 \Delta m_i$. For comparable uncertainties ($W_j \sim W_i$):

$$\Delta W_j \propto K(\text{loc}_i, \text{loc}_j)^2 \cdot W_i \cdot \Delta m_i$$

□

C.13 Computational Efficiency: Micro and Macro Updates

Micro-Update Derivation. Let $\hat{\alpha}_t$ be the optimal solution at time t, satisfying $\nabla L(\hat{\alpha}_t; \mathcal{F}_t) = 0$. After adding Δm_i samples at hypothesis i, the new optimal $\hat{\alpha}_{t+1}$ satisfies:

$$\nabla L(\hat{\alpha}_{t+1}; \mathcal{F}_{t+1}) = 0 \tag{111}$$

By Taylor expansion around $\hat{\alpha}_t$:

$$\begin{aligned}
0 = \nabla L(\hat{\alpha}_t; \mathcal{F}_{t+1}) + H(\hat{\alpha}_t)(\hat{\alpha}_{t+1} - \hat{\alpha}_t) \\
+ O(\|\hat{\alpha}_{t+1} - \hat{\alpha}_t\|^2)
\end{aligned} \tag{112}$$

The change in gradient is:

$$\nabla L(\hat{\alpha}_t; \mathcal{F}_{t+1}) - \nabla L(\hat{\alpha}_t; \mathcal{F}_t) = \nabla \Delta \ell_i \tag{113}$$

where only the i-th data term changes:

$$\begin{aligned}
\Delta \ell_i = \log\left[\alpha_i P(k_i + \Delta k_i | m_i + \Delta m_i, \cdot) + \cdots\right] \\
- \log\left[\alpha_i P(k_i | m_i, \cdot) + \cdots\right]
\end{aligned} \tag{114}$$

Ignoring second-order terms:

$$\hat{\alpha}_{t+1} - \hat{\alpha}_t \approx -H^{-1}(\hat{\alpha}_t) \nabla \Delta \ell_i \tag{115}$$

Sparse Gradient. The gradient $\nabla \Delta \ell_i$ has the form:

$$\frac{\partial \Delta \ell_i}{\partial \alpha_j} = \frac{\partial \Delta \ell_i}{\partial \alpha_i} \frac{\partial \alpha_i}{\partial \alpha_j} = \frac{\partial \Delta \ell_i}{\partial \alpha_i} K_{ij} \tag{116}$$

So $\nabla \Delta \ell_i = \frac{\partial \Delta \ell_i}{\partial \alpha_i} \mathbf{k}_i$, where $\mathbf{k}_i$ is the i-th column of K.

Update Formula. The natural gradient update is:

$$\Delta\hat{\alpha} = -\frac{\partial \Delta\ell_i}{\partial \alpha_i} H^{-1}\mathbf{k}_i \tag{117}$$

We maintain H^{-1} explicitly and update it via the Sherman-Morrison formula when Fisher information v_i changes:

$$H^{-1}_{\text{new}} = H^{-1}_{\text{old}} - \frac{\Delta v_i \cdot uu^\top}{1 + \Delta v_i \cdot k_i^\top u} \tag{118}$$

where $u = H^{-1}_{\text{old}}\mathbf{k}_i$. This rank-1 update costs $O(N^2)$ per hypothesis. The inverse Hessian is also needed for confidence interval computation ($\text{Var}[\alpha_i] = \mathbf{k}_i^\top H^{-1}\mathbf{k}_i$), making explicit maintenance advantageous. For very large problems ($N > 10^4$) where storing H^{-1} becomes prohibitive, the system $H\mathbf{u} = \mathbf{k}_i$ can be solved via Conjugate Gradient instead. CG is well-suited because the Hessian $H = K^\top \text{diag}(v)K + 2\lambda K$ is symmetric positive definite, with regularization $2\lambda K$ ensuring convergence typically within $O(\sqrt{\kappa})$ iterations where $\kappa = O(1/\lambda)$ is the condition number.

Numerical Stability Heuristics. For robustness in practice, we apply three safeguards: (i) when using Sherman-Morrison, numerically refresh H^{-1} periodically (every 100 updates) to prevent accumulation of floating-point errors, (ii) clip $\|\Delta\hat{\alpha}\|$ to at most 50% of $\|\hat{\alpha}\|$ per update to prevent catastrophic drift from ill-conditioned systems, and (iii) apply 0.8 damping to the Newton step to prevent overshooting. These heuristics are critical for stability when hypotheses approach decision boundaries where the likelihood curvature can vary dramatically.

Macro-Update Strategy. We trigger full re-optimization when:

1. After T micro-updates (e.g., $T = 10$)
2. Cumulative drift: $\sum_t \|\Delta\hat{\alpha}_t\| > \tau$ (e.g., $\tau = 0.1$)
3. Likelihood degradation: $L(\hat{\alpha}_{\text{micro}}) - L(\hat{\alpha}_{\text{macro}}) > \epsilon$

The warm-start initialization uses $\hat{\alpha}_{\text{micro}}$ instead of random or zero initialization, typically reducing iterations by 50-80%.

Lemma 3 (Micro-Update Approximation Error)**.** *Let $\hat{\alpha}_t$ denote the MLE at time t with data $\mathcal{F}_t$. After adding Δm_i samples at hypothesis i, let:*

- *$\hat{\alpha}^{true}_{t+1}$ denote the exact MLE with data $\mathcal{F}_{t+1}$*
- *$\hat{\alpha}^{micro}_{t+1} = \hat{\alpha}_t - H_t^{-1}\nabla\Delta\ell_i$ denote the micro-update approximation*

Define the approximation error:

$$\delta = \hat{\alpha}^{true}_{t+1} - \hat{\alpha}^{micro}_{t+1}$$

Under the assumptions:

1. *The Hessian is Lipschitz continuous: $\|H(\alpha) - H(\beta)\| \le L_H\|\alpha - \beta\|$*
2. *The Hessian is well-conditioned: $\mu I \preceq H \preceq LI$ for $0 < \mu \le L$*
3. *The update is small: $\epsilon := \|\nabla\Delta\ell_i\| \le \delta_0$ where $L_H\delta_0/\mu < 1/2$*

Then the error satisfies:

$$\|\delta\| \le C_1 \frac{\kappa\,\Delta v_i}{\mu^2}\,\epsilon \;+\; C_2 \frac{L_H}{\mu^2}\,\epsilon^2$$

where $\Delta v_i = v_i(\mathcal{F}_{t+1}) - v_i(\mathcal{F}_t)$ is the change in Fisher information, $\kappa = O(1)$ by Assumption 1, and $C_1, C_2 = O(1)$ are absolute constants. The leading term is linear in the gradient perturbation ϵ, weighted by the curvature change Δv_i; only the Lipschitz remainder is $O(\epsilon^2)$. This matches the practical bound below.

Practical bound: *For small relative updates $\Delta m_i/m_i \ll 1$:*

$$\|\delta\| = O\left(\frac{\Delta m_i}{m_i}\right)\cdot\epsilon + O(\epsilon^2)$$

Proof. By the implicit function theorem, the exact MLE satisfies:

$$\nabla L(\hat{\alpha}_{t+1}^{\text{true}}; \mathcal{F}_{t+1}) = 0$$

The micro-update uses the first-order approximation:

$$\hat{\alpha}_{t+1}^{\text{micro}} = \hat{\alpha}_t - H_t^{-1} \nabla L(\hat{\alpha}_t; \mathcal{F}_{t+1})$$

where $\nabla L(\hat{\alpha}_t; \mathcal{F}_{t+1}) = \nabla \Delta \ell_i$ since $\nabla L(\hat{\alpha}_t; \mathcal{F}_t) = 0$.

By Taylor expansion of ∇L around $\hat{\alpha}_t$:

$$\begin{aligned} \nabla L(\hat{\alpha}_{t+1}^{\text{micro}}; \mathcal{F}_{t+1}) = \nabla L(\hat{\alpha}_t; \mathcal{F}_{t+1}) + H_{t+1}(\hat{\alpha}_{t+1}^{\text{micro}} - \hat{\alpha}_t) \\ + O(\|\hat{\alpha}_{t+1}^{\text{micro}} - \hat{\alpha}_t\|^2) \end{aligned} \tag{119}$$

Substituting $\hat{\alpha}_{t+1}^{\text{micro}} - \hat{\alpha}_t = -H_t^{-1} \nabla \Delta \ell_i$ and $\nabla L(\hat{\alpha}_t; \mathcal{F}_{t+1}) = \nabla \Delta \ell_i$:

$$\nabla L(\hat{\alpha}_{t+1}^{\text{micro}}; \mathcal{F}_{t+1}) = \nabla \Delta \ell_i - H_{t+1} H_t^{-1} \nabla \Delta \ell_i + O(\epsilon^2)$$

Using $H_{t+1} = H_t + \Delta H_i$:

$$\begin{aligned} &= [I - (H_t + \Delta H_i) H_t^{-1}] \nabla \Delta \ell_i + O(\epsilon^2) \\ &= -\Delta H_i H_t^{-1} \nabla \Delta \ell_i + O(\epsilon^2) \end{aligned}$$

The exact solution satisfies $\nabla L(\hat{\alpha}_{t+1}^{\text{true}}; \mathcal{F}_{t+1}) = 0$. By Taylor expansion:

$$0 = \nabla L(\hat{\alpha}_{t+1}^{\text{micro}}; \mathcal{F}_{t+1}) + H_{t+1} \delta + O(\|\delta\|^2)$$

Rearranging:

$$\delta = -H_{t+1}^{-1} \nabla L(\hat{\alpha}_{t+1}^{\text{micro}}; \mathcal{F}_{t+1}) + O(\|\delta\|^2)$$

Substituting the residual from Step 1:

$$\begin{aligned} \delta &= -H_{t+1}^{-1} [-\Delta H_i H_t^{-1} \nabla \Delta \ell_i + O(\epsilon^2)] + O(\|\delta\|^2) \\ &= H_{t+1}^{-1} \Delta H_i H_t^{-1} \nabla \Delta \ell_i + O(\epsilon^2) \end{aligned}$$

where we absorbed $O(\|\delta\|^2)$ into $O(\epsilon^2)$ by noting that $\|\delta\| = O(\epsilon)$ self-consistently.

The Hessian change for location i is rank-one:

$$\Delta H_i = \Delta v_i \cdot \mathbf{k}_i \mathbf{k}_i^T$$

where Δv_i is the change in local Fisher information and $\mathbf{k}_i$ is the i-th column of the kernel matrix.

Therefore:

$$\|\delta\| \leq \|H_{t+1}^{-1}\| \cdot \|\Delta H_i\| \cdot \|H_t^{-1}\| \cdot \|\nabla \Delta \ell_i\| + O(\epsilon^2)$$

Using the bounds $\|H^{-1}\| \leq 1/\mu$, $\|\mathbf{k}_i\| \leq \sqrt{\kappa}$ (where $\kappa := \max_i \|\mathbf{k}_i\|_2^2 = O(1)$ by Assumption 1), and $\|\Delta H_i\| = \Delta v_i \|\mathbf{k}_i\|^2$:

$$\|\delta\| \leq \frac{\kappa \Delta v_i}{\mu^2} \epsilon + O(\epsilon^2)$$

The second-order remainder can be bounded using Lipschitz continuity:

$$\|O(\epsilon^2)\| \leq \frac{L_H}{2\mu^2} \epsilon^2$$

Combining:

$$\|\delta\| \leq \frac{\kappa \Delta v_i}{\mu^2} \, \epsilon + \frac{L_H}{2\mu^2} \, \epsilon^2$$

By Assumption 1, $\kappa = O(1)$, giving the stated bound. □

C.14 Allocation policy: closed-form expressions and tuning

This appendix details the implementation of the priority score S_i (9), the variance-channel-specific saturation and headroom factors, the β-mixing tuning rule, and the helper-sampling benefit comparison referenced in Section 5.2.

C.14.1 Per-channel saturation and headroom

Recall (App. C.8) that the per-hypothesis variance σ_i^2 decomposes into a local (Fisher) channel with diagonal contribution $K_{ii}^2 H_{ii}^{-1}$ and a spatial channel collecting the remaining off-diagonal contributions. Each channel has its own variance floor (App. C.9, App. C.11) and its own "how close to the floor" and "how much of the current variance comes from this channel" metrics.

Saturation ratios. Define the saturation ratios

$$\rho_{F,i} \;\coloneqq\; \frac{v_i}{v_\infty(\hat{p}_i)} + 1, \qquad \rho_{H,i} \;\coloneqq\; 1 + \frac{m_i - m_{\text{start}}}{m_0}\,\mathbb{I}[m_i > m_{\text{start}}], \tag{120}$$

where $v_\infty(\hat{p})$ is the limiting Fisher information of App. C.11, and (m_{start}, m_0) are the blending constants of App. C.10. Both ratios are ≥ 1, with $\rho_{F,i} \to 1$ as v_i saturates at its terminal value and $\rho_{H,i} \to 1$ when no fresh spatial information is expected to arrive (early in sampling, before m_{start}).

Saturation factors. The saturation factor of channel $c \in \{F, H\}$ is

$$\text{sat}_{c,i} \;\coloneqq\; \max\left\{0,\, \min\left\{1,\, \frac{\rho_{c,i} - 1}{\rho_{\text{stop}} - 1}\right\}\right\}, \tag{121}$$

with $\rho_{\text{stop}} = 1.1$ defining the saturation cutoff. $\text{sat}_{c,i} = 1$ when the channel still has substantial capacity for additional information; $\text{sat}_{c,i} = 0$ when the channel is at or beyond its capacity floor. The clamping prevents the priority score from being amplified beyond its natural range.

Headroom factors. The headroom factor measures how much of the current variance is attributable to the channel:

$$\kappa_{F,i} \;\coloneqq\; 1 - w_i, \tag{122}$$

$$\kappa_{H,i} \;\coloneqq\; \frac{m_0}{m_0 + \max(m_i - m_{\text{start}},\, 0)}, \tag{123}$$

where w_i is the Hessian–Fisher blending weight of App. C.10. $\kappa_{F,i}$ is the complement of the blending weight: when the local channel dominates ($w_i \to 0$, near the Fisher regime), local sampling is highly relevant ($\kappa_{F,i} \to 1$); when the spatial channel dominates ($w_i \to 1$), local sampling moves the variance only marginally ($\kappa_{F,i} \to 0$). $\kappa_{H,i}$ encodes the freshness of spatial information: a hypothesis sampled many times beyond m_{start} has small headroom for the spatial channel because most of the spatial coupling has already been accumulated.

Putting it together. Substituting (121)–(123) into (9) of the main text yields the explicit priority score evaluated in our experiments. Channels contribute multiplicatively: a hypothesis that has saturated its Fisher channel ($\text{sat}_{F,i} \to 0$) loses its local term automatically, regardless of how much CI width W_i remains; analogously for the spatial channel.

C.14.2 β-mixing parameter and burn-in adaptation

The mixing parameter $\beta \in [0, 1]$ in (9) controls how much weight the policy places on local Fisher information versus spatial borrowing across the kernel-coupled neighborhood. As discussed in the main text, setting β too high produces a cluster-size bias (alternatives in small clusters get undersampled); setting it too low over-relies on spatial borrowing in regimes where the spatial coupling is weak.

Adaptation rule. We use a short burn-in to pick β adaptively:

1. **Initialization:** $\beta \leftarrow 0.9$ (local-heavy, the safer default in early sampling when the spatial estimator is still coarse).

2. **Burn-in trigger:** once 50 hypotheses have accumulated at least 20 samples each (so the local Fisher information at those hypotheses is reliable), enter the adaptation step.
3. **Estimation:** estimate the global null fraction $\hat{\alpha}_0$ as the fraction of these 50 hypotheses with $\hat{p}_i > \tau$ (where τ is the FDR target).
4. **Update:** set $\beta \leftarrow \max(0.5, \hat{\alpha}_0)$.
5. **Frozen thereafter:** the resulting β is held constant for the remainder of allocation.

Intuition. When most hypotheses appear null ($\hat{\alpha}_0$ large), alternatives are likely sparse and clusters small; the policy biases toward more local sampling so as not to lose the few alternatives to spatial averaging. When alternatives are clustered ($\hat{\alpha}_0$ small), spatial borrowing is informative and a lower β gives appropriate weight to neighbors. The lower bound 0.5 ensures the local channel never disappears entirely.

C.14.3 Helper-sampling benefit comparison

Decided and hopeless hypotheses are removed from the priority queue in (9), but they may still carry residual local uncertainty whose reduction would propagate through the kernel to ambiguous neighbors. Helper sampling re-admits them when this spatial benefit exceeds direct sampling of the bottlenecked ambiguous neighbor.

Spatial variance fraction. For an ambiguous hypothesis i, define

$$r_i \;:=\; 1 - \frac{K_{ii}^2 H_{ii}^{-1}}{(K H^{-1} K^\top)_{ii}}. \tag{124}$$

$r_i \in [0,1]$ measures the fraction of i's uncertainty that originates from spatial coupling rather than from i's own local data. $r_i \to 0$ means the local Fisher channel is the bottleneck; $r_i \to 1$ means the bottleneck is spatial.

Local versus spatial benefit. Sampling hypothesis i directly yields a local benefit

$$\mathrm{LocalBenefit}_i \;:=\; W_i^2 \cdot (1 - r_i) \cdot \kappa_{F,i}, \tag{125}$$

weighted by the fraction of i's uncertainty actually addressable by local sampling. Sampling a (decided or hopeless) neighbor $j \in \mathcal{D}$ instead provides a spatial benefit to i

$$\mathrm{SpatialBenefit}_{i \leftarrow j} \;:=\; W_i^2 \cdot r_i \cdot K_{ij}^2 \cdot \kappa_{F,j}, \tag{126}$$

weighted by their kernel similarity squared (the rate at which information flows from j to i via the Hessian, by App. C.12) and the neighbor's remaining local headroom $\kappa_{F,j}$.

Substitution rule. For each ambiguous hypothesis i scheduled for sampling whose spatial fraction satisfies $r_i > r_{\mathrm{thresh}}$ (we use $r_{\mathrm{thresh}} = 0.5$), identify the best potential helper

$$j^*(i) \;:=\; \arg\max_{j \in \mathcal{D}} \; K_{ij}^2 \cdot \kappa_{F,j}. \tag{127}$$

If $\mathrm{SpatialBenefit}_{i \leftarrow j^*(i)} > \mathrm{LocalBenefit}_i$, substitute $j^*(i)$ for i in the next allocation batch. Note that this substitution does not reopen the rejection decision for $j^*(i)$: the additional sampling refines $j^*(i)$'s local Fisher information (and through the kernel, i's spatial estimate) without changing $j^*(i)$'s lfdr-CI position.

Why decided hypotheses can still help. A decided hypothesis j has its lfdr CI entirely on one side of $\hat{\tau}_q$, so its rejection outcome is fixed. But its $\hat{\alpha}(\mathrm{loc}_j)$ may still be far from the true $\alpha^*(\mathrm{loc}_j)$ in the sense that more samples would refine the spatial estimate at loc_j — and through the kernel, the spatial estimate at every loc_i within the kernel's range. Helper sampling exploits this "decided $\neq$ saturated" distinction.

C.14.4 Overall allocation algorithm

The full algorithm in pseudo-code:

Algorithm 2 Adaptive null-draw allocation

Require: Initial counts $\{(k_i, m_i)\}$, kernel K, FDR target q, batch size Δm, total budget M.

$\beta \leftarrow 0.9$
while total samples used $< M$ **do**
  Update $(\hat{\tilde{\alpha}}, \hat{b})$ via empirical Bayes (§3.1); refit $\hat{\boldsymbol{c}}$ via (6) if a macro-update is due (App. C.13); else apply Sherman–Morrison micro-update.
  Compute lfdr CIs $[\mathrm{lfdr}_i^{\mathrm{low}}, \mathrm{lfdr}_i^{\mathrm{high}}]$ (§5.1).
  Classify each hypothesis as *decided*, *ambiguous*, or *hopeless* (App. C.11, Fig. 1).
  if burn-in trigger met (50 hypotheses with $m_i \geq 20$) and not yet adapted **then**
    Estimate $\hat{\alpha}_0$ on burn-in subset; set $\beta \leftarrow \max(0.5, \hat{\alpha}_0)$. Mark adapted.
  end if
  Compute S_i for ambiguous hypotheses via (9) using (120)–(123).
  Pick top-priority ambiguous hypothesis i^*.
  if $r_{i^*} > r_{\mathrm{thresh}}$ **then**
    Compute $j^*(i^*)$ via (127).
    if $\mathrm{SpatialBenefit}_{i^* \leftarrow j^*(i^*)} > \mathrm{LocalBenefit}_{i^*}$ **then**
      Sample Δm additional null draws at $j^*(i^*)$.
    else
      Sample Δm additional null draws at i^*.
    end if
  else
    Sample Δm additional null draws at i^*.
  end if
  Update (k_{i^*}, m_{i^*}) or (k_{j^*}, m_{j^*}) accordingly.
end while
Apply Rule 1 or Rule 2 (§4) on the final state.

Constants used in our experiments. $m_{\mathrm{start}} = 300$; $m_0 = 100$; $\rho_{\mathrm{stop}} = 1.1$; $r_{\mathrm{thresh}} = 0.5$; $\Delta m = 10$; burn-in trigger: 50 hypotheses each with $m_i \geq 20$; CI coverage $1 - \gamma = 0.9$. The framework is robust to $\pm 50\%$ variation in $(m_{\mathrm{start}}, m_0)$.

Numerical safeguards. For the Sherman–Morrison micro-update, three safeguards are applied (also discussed in App. C.13): (i) periodic refresh of H^{-1} every 100 updates to prevent floating-point drift; (ii) clip $\|\Delta\hat{\boldsymbol{c}}\|$ to at most 50% of $\|\hat{\boldsymbol{c}}\|$ per update; (iii) damp the Newton step by 0.8 to prevent overshooting near decision boundaries where the likelihood curvature varies sharply.

C.14.5 Status of the guarantees under adaptive allocation

The policy decides each allocation from the data observed so far, before the corresponding draws, so the design factor is free of the parameters and drops out of the likelihood and the posterior. What fails is the fixed-m *sampling law*:

Component	Under adaptive m_i
Objective (6): convexity, representer form, natural-gradient solver	unaffected: the design factor drops out of the likelihood
Rule 1 conditioning identity $\mathbb{E}[V \mid \mathcal{D}] = \sum_{i \in \mathcal{R}_1} \mathrm{lfdr}_i^*$	unaffected, for the same reason
Marginal factorization (Theorem 5) behind $(\hat{\tilde{\alpha}}, \hat{b})$; the tower identity (129) that calibrates (PC); Theorem 1(b)	fail: the realized budgets depend on the counts
Discrete-uniform null (Assumption 4) behind Rule 2	fails under optional stopping; restored by folded allocation (Proposition 4)
Interval coverage (Section 5.1) and Proposition 1	fail: coverage is fixed-time

Proposition 4 (Folded allocation preserves the mirror)**.** *Suppose every allocation decision is a function of the folded count paths* $\breve{k}_{j,t} = \min(k_{j,t},\, t - k_{j,t})$, $j = 1, \ldots, N$, *and of independent*

randomization only. If $p_i^* \sim U(0,1)$ for every null i, then $(m_i, k_i) \overset{d}{=} (m_i, m_i - k_i)$ jointly with all other hypotheses' data, and Proposition 5 holds with the adaptively allocated budgets.

Proof. For a null i, reflect its latent p-value and each of its draw indicators, $p_i^* \mapsto 1 - p_i^*$ and $\mathbb{I}_{i,s} \mapsto 1 - \mathbb{I}_{i,s}$. Since $p_i^* \sim U(0,1)$ is symmetric about $1/2$ and the draws are Bernoulli(p_i^*), the reflection preserves the joint law of all data. It fixes every folded path, hence every allocation decision and m_i, maps k_i to $m_i - k_i$, and leaves all $j \neq i$ untouched. The orientations of the nulls are therefore fair coins independent of the folded data, which is what Step 3 and Proposition 5 use. □

The deployed policy reads the unfolded interval width W_i and is not of this form, so Proposition 4 describes a variant rather than a repair: it buys exactness under adaptive allocation at the cost of targeting on folded information only, $I^{\mathrm{f}}/I^{\mathrm{u}} \approx 0.44$ at $m = 6$ for the canonical alternative of (137). Two further repairs remain open: replacing the fixed-time intervals by time-uniform confidence sequences, and freezing $(\hat{\bar{\alpha}}, \hat{b})$ at the end of a burn-in whose depth is set by a criterion rather than by convention.

D Proofs for Decision Rules

This section collects the proofs of the three FDR-related theorems of Section 4: the FDR control of Rule 1 under condition (PC) (Theorem 1), the near-exact FDR control of Rule 2 ($\mathrm{FDR} \leq \tau + C_{\mathrm{BC}}(L\beta_N + L_b|\Delta\hat{b}|)$, Theorem 2), with exact control, $\mathrm{FDR} \leq \tau$, for its folded variant (Proposition 5), and the optimality of the spatial lfdr score under correct specification (Theorem 3).

D.1 Proof of Theorem 1 (Rule 1 FDR control)

The rule structurally inherits Sun–Cai's running-average rule [21]. The count-space-specific work is verifying that the marginal lfdr score $\mathrm{lfdr}_{\mathrm{marg}}(k_i, m_i)$ of (3) is the posterior null probability given the hypothesis's own counts under the finite-m Beta–Binomial model (Step 1), and that ordering by the data-dependent spatial score leaves the guarantee intact under (PC) (Step 2). Throughout, $\mathrm{lfdr}_i^* = \Pr(\theta_i = 0 \mid \text{all data})$ as in Section 4, and $(\bar{\alpha}, b)$ are at their true values.

Step 1: $\mathrm{lfdr}_{\mathrm{marg}}(k_i, m_i)$ is the Bayes posterior of $H_{0,i}$ under the marginal model. By Theorem 5, marginalizing the spatial mixture (5) over the location distribution μ recovers a non-spatial two-group mixture with global null proportion $\bar{\alpha} = \mathbb{E}_\mu[\alpha(\mathrm{loc})]$ and Beta-Binomial likelihoods $P_0(k, m), P_1(k, m; b)$ that are *themselves* the count-space marginal pmfs of the latent p_i^* (Section 3.1). By construction

$$\mathrm{lfdr}_{\mathrm{marg}}(k_i, m_i) \;=\; \frac{\bar{\alpha}\, P_{0,i}}{\bar{\alpha}\, P_{0,i} + (1 - \bar{\alpha})\, P_{1,i}} \;=\; \Pr(\theta_i = 0 \mid k_i, m_i) \tag{128}$$

under the marginal model, where the equality holds at every finite $m_i \geq 1$ exactly (not asymptotically): the Beta-Binomial conjugacy (2) integrates the latent-p^* mixture in closed form, with no approximation. Since (k_i, m_i) is part of the data, the tower property gives

$$\mathrm{lfdr}_{\mathrm{marg}}(k_i, m_i) \;=\; \mathbb{E}\big[\mathrm{lfdr}_i^* \mid k_i, m_i\big]. \tag{129}$$

This is the central finite-m point: the count-space substitution preserves the posterior property of the Sun–Cai score *without invoking the $m \to \infty$ limit*. The $O(1/m)$ rate at which $\mathrm{lfdr}_{\mathrm{marg}}(k_i, m_i)$ converges to the classical continuous lfdr (Theorem 4, App. C) governs how close the count-space rule comes to the continuous-data oracle, but does not enter the validity argument, which depends only on $\mathrm{lfdr}_{\mathrm{marg}}$ being the correct posterior under the model used to derive the rule, not on closeness to any continuous limit.

Step 2: Conditioning on the full data. Let $\mathcal{D} = \{(\mathrm{loc}_j, k_j, m_j)\}_{j=1}^N$. The rejection set $\mathcal{R}_1$ is a deterministic function of $\mathcal{D}$: the ordering, $\hat{\alpha}$ and λ are all computed from it. Hence, with $V = |\mathcal{R}_1 \cap \mathcal{H}_0|$,

$$\mathbb{E}[V \mid \mathcal{D}] \;=\; \sum_{i \in \mathcal{R}_1} \Pr(\theta_i = 0 \mid \mathcal{D}) \;=\; \sum_{i \in \mathcal{R}_1} \mathrm{lfdr}_i^*. \tag{130}$$

No independence between the ordering and the labels is required: conditioning on $\mathcal{D}$ fixes the ordering, $\hat{\alpha}$ and λ, however they were chosen. By construction, the rule enforces on every realization

$$\sum_{i\in\mathcal{R}_1} \mathrm{lfdr}_{\mathrm{marg}}(k_i, m_i) \;\le\; \tau\,|\mathcal{R}_1|. \tag{131}$$

Part (a). Since $|\mathcal{R}_1|$ is $\mathcal{D}$-measurable, (130) gives

$$\mathrm{FDR}(\mathcal{R}_1) = \mathbb{E}\Big[\tfrac{\sum_{i\in\mathcal{R}_1}\mathrm{lfdr}_{\mathrm{marg}}(k_i,m_i)}{|\mathcal{R}_1|\vee 1}\Big] + \mathbb{E}\Big[\tfrac{\sum_{i\in\mathcal{R}_1}(\mathrm{lfdr}_i^* - \mathrm{lfdr}_{\mathrm{marg}}(k_i,m_i))}{|\mathcal{R}_1|\vee 1}\Big] \;\le\; \tau,$$

the first term bounded by (131) and the second non-positive by (PC).

Part (b). Taking expectations in (130), $\mathbb{E}[V] = \mathbb{E}\big[\sum_{i\in\mathcal{R}_1}\mathrm{lfdr}_{\mathrm{marg}}(k_i,m_i)\big] + \sum_{i=1}^N \mathbb{E}\big[\mathbb{I}\{i \in \mathcal{R}_1\}(\mathrm{lfdr}_i^* - \mathrm{lfdr}_{\mathrm{marg}}(k_i,m_i))\big]$. By (129), $\mathbb{E}[\mathrm{lfdr}_i^* - \mathrm{lfdr}_{\mathrm{marg}}(k_i,m_i)] = 0$, so each summand is exactly $\mathrm{Cov}\big(\mathbb{I}\{i \in \mathcal{R}_1\}, \mathrm{lfdr}_i^* - \mathrm{lfdr}_{\mathrm{marg}}(k_i,m_i)\big) \le 0$. With (131), $\mathbb{E}[V] \le \tau\,\mathbb{E}|\mathcal{R}_1|$, i.e. $\mathrm{mFDR}(\mathcal{R}_1) \le \tau$. □

A random ordering satisfies (PC) **with equality.** Under the generative model with budgets m_j fixed in advance, the triples $(\mathrm{loc}_j, \theta_j, k_j)$ are independent across j, with $\mathrm{loc}_j \sim \mu$ (Theorem 5). Suppose the ordering is independent of the locations given the counts (in particular, a uniformly random ordering). Then $\mathcal{R}_1$ is a function of the counts and the ordering alone, and conditional on both, loc_i retains its law given (k_i, m_i), so $\mathbb{E}[\mathrm{lfdr}_i^* \mid \text{counts, ordering}] = \mathrm{lfdr}_{\mathrm{marg}}(k_i, m_i)$ by (129). Conditioning the left-hand side of (PC) on the counts and the ordering therefore makes it exactly zero. An informative ordering selects hypotheses whose true posterior lies below this conditional mean, and only an ordering that prefers hypotheses more likely null than their counts alone suggest can make it positive.

Why no perturbation slack arises. The accounting score $\mathrm{lfdr}_{\mathrm{marg}}(k_i, m_i)$ depends only on (k_i, m_i) and the global parameters, not on $\hat{\alpha}$. The spatial estimator enters Rule 1 only through the ordering, and hence only through (PC); the influence-function argument of Theorem 2, where a single flip moves the score itself, has no counterpart here. Given (PC), the bounds hold at every finite N with no perturbation term.

E Rule 2 Guarantees and Error control

Throughout, write $g_i(k) := \widehat{\mathrm{lfdr}}_{\mathrm{spatial}}(\hat{\alpha}(\mathrm{loc}_i), k, m_i)$ for the spatial lfdr score of hypothesis i as a function of its count, so $T_i = g_i(k_i)$ and $\tilde{T}_i = g_i(m_i - k_i)$; the score is increasing in the count, so the folded score is $\breve{T}_i = g_i(\breve{k}_i)$ with $\breve{k}_i = \min(k_i, m_i - k_i)$ and its mirror $\widetilde{\breve{T}}_i = g_i(m_i - \breve{k}_i) = \max(T_i, \tilde{T}_i)$.

E.1 Roadmap, contribution, and the three guarantees

As described in the main text, Rule 2 reuses each hypothesis's own count in the score it is then tested with., allowing for a better power (it keeps the local information that masking discards) and also the one place it departs from an exact mirror procedure. Here we account of the error control this buys in three registers, and presents the landscape of which guarantee rests on which assumptions.

Contribution.

(i) **Exact finite-sample FDR for a discrete mirror rule.** The *folded* construction (Section E.5) controls FDR at level τ with *no slack and no asymptotics*, at every multiplicity $m_i \ge 1$, using only the flat-null symmetry (Assumption 4) and conditional independence (Assumption 5). It requires *neither* that the fitted surface be correctly specified *nor* any threshold-regularity condition. This is the count-space analogue of Barber–Candès sign-symmetry, with the mirror supplied by the reflection $k \mapsto m - k$ and the fair coin by the discrete flat null. It is the rock the section stands on.

(ii) **A characterized, computable relaxation for the deployed plug-in.** The departure of the plug-in from exactness is governed by a single-flip influence bound $\delta = L\beta_N + L_b|\Delta\hat{b}| =$

$O(1/\lambda)+O(1/N)$ (Section E.6), with a spatial channel ($O(1/\lambda)$, the refit surface; Lemma 7) and a direct pooled-scalar channel ($O(1/N)$, the score's dependence on $\hat{b}$); at a cross-validated λ that does not grow with N, the first does not vanish as N grows. Notice, this is *not* an exact finite-sample FDR for the plug-in- we state the relaxation as a β_N-controlled quantity (Remark 3), computable on the data at hand.

(iii) **An averaged (mFDR) guarantee for the plug-in at the slack rate.** We prove $\mathrm{mFDR}(\mathcal{R}_2) \le \tau + O(\delta)$ (Theorem 7), *conditional* on a threshold-stability hypothesis (Assumption 8). The proof reduces the slack to a single sum over each null's *own* distance to the cutoff. Its role is to *isolate* the one genuine obstruction (the simultaneous-perturbation stability of the data-dependent cutoff) into a single named assumption, with the unconditional fallback (Theorem 8) is stated alongside, so the conditional nature is visible.

Aa a note, we see this delineation as part of the paper contribution: exact where exactness holds (the folded rule), characterized relaxation where it does not (the plug-in).

The three guarantees and their dependencies.

Register	Result	Guarantee	Assumptions
Folded ($\delta = 0$)	Prop. 5	FDR $\le \tau$ exact, all $m_i \ge 1$	4, 5
Plug-in, finite-sample FDR	Rem. 3	β_N-controlled relaxation	2 (characterization)
Plug-in, mFDR (conditional)	Thm. 7	$\mathrm{mFDR} \le \tau + O(\delta)$	4–8
Plug-in, mFDR (unconditional)	Thm. 8	$\tau + O(\underline{s})$ or $\tau + \tilde{O}(\lambda^{-d}\delta)$	4–7

Here, the $\delta = 0$ baseline (Section E.4) carries the only unconditional FDR claim, on the fewest assumptions. Turning on the plug-in (Section E.6) introduces δ and forfeits exact FDR for a *structural* reason (the surface is learned from the counts it scores), not a technical one. The averaged guarantee (Section E.8) recovers control at the slack rate, contingent on Assumption 8.

E.2 Setup and shared objects

We work in count space (k_i, m_i), reflecting that the latent p_i^* is never observed, on the active set $\mathcal{A} = \{i : m_i \ge 1\}$ with $N = |\mathcal{A}|$. The count-space null and alternative pmfs are $P_0(k \mid m) = 1/(m+1)$ and $P_1(k \mid m; b)$, a Beta–Binomial with $a \in (0,1)$, $b > 1$, concentrating its mass at small k. The surface $\hat{\alpha}$ is the penalized estimate of (6); the global scalars $(\hat{\tilde{\alpha}}, \hat{b})$ are estimated beforehand from the marginal counts (Storey's estimator and maximum likelihood, Section 3.1) and held fixed during the spatial fit. The step-up rule (7) sets

$$\hat{t}_q = \max\left\{t \in \{T_i\} \cup \{\tilde{T}_i\} : \tfrac{1+\#\{i:\tilde{T}_i \le t\}}{1 \vee \#\{i:T_i \le t\}} \le \tau\right\}, \qquad \mathcal{R}_2 = \{i : T_i \le \hat{t}_q\}.$$

Write $V = \#\{i \in \mathcal{H}_0 : T_i \le \hat{t}_q\}$ (false rejections), $\tilde{V} = \#\{i \in \mathcal{H}_0 : \tilde{T}_i \le \hat{t}_q\}$ (null mirror count), $R = |\mathcal{R}_2|$, $A = \#\{i : \tilde{T}_i \le \hat{t}_q\}$, and $\mathrm{FDR} = \mathbb{E}[V/(1 \vee R)]$, $\mathrm{mFDR} = \mathbb{E}[V]/\mathbb{E}[R]$. The *orientation* is $B_i = \mathbb{1}(k_i > m_i/2) = \mathbb{1}(T_i > \tilde{T}_i)$, the *active set* is $\{i \text{ active}\} = \{\breve{T}_i \le \hat{t}_q < \breve{\tilde{T}}_i\}$, and the *own-proximity* of i to the cutoff is $\eta_i = \min(|\breve{T}_i - \hat{t}_q|,\ |\breve{\tilde{T}}_i - \hat{t}_q|)$, with $N_\partial(w) = \#\{i \in \mathcal{H}_0 : \eta_i \le w\}$.

E.3 Assumptions

The first two assumptions are all the $\delta = 0$ register needs; the remainder enter only for the plug-in.

Assumption 4 (Discrete-uniform null)**.** *For $i \in \mathcal{H}_0$, $k_i \mid (H_{0,i}, m_i) \sim \mathrm{Uniform}\{0, \dots, m_i\}$; equivalently $P_0(k \mid m) = 1/(m+1)$ is the exact null law. When k_i is the rank of a statistic among m_i exchangeable null draws, this is exact by exchangeability, at every $m_i \ge 1$.*

Assumption 5 (Conditional independence across hypotheses)**.** *Given the label configuration $\{\mathbb{1}(j \in \mathcal{H}_0)\}$ and sizes $\{m_j\}$, the counts $\{k_j\}$ are mutually independent. (Spatial dependence resides in the latent label field. Simply put: given labels, each count depends only on its own hypothesis.)*

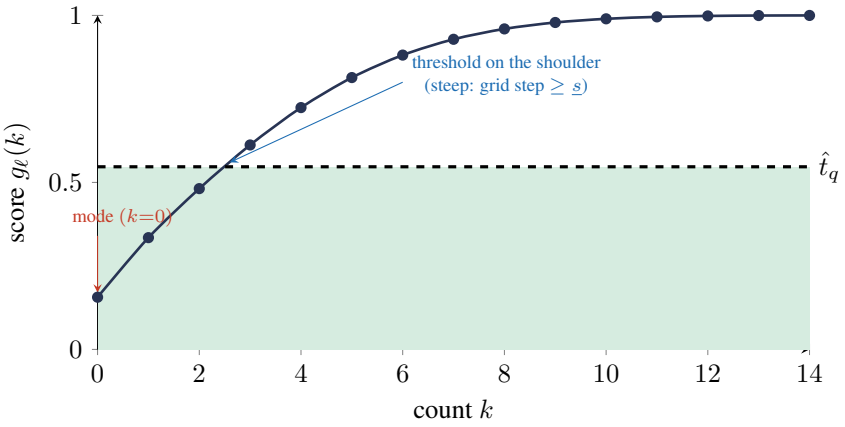


Figure 5: **Why the grid step is bounded below at the threshold.** The score $g_\ell(k)$ rises with the count (marks: true values for $a = 0.5$, $b = 6$, $m_\ell = 14$, $\alpha_\ell = 0.6$). The Beta–Binomial alternative concentrates mass at small k, so its mode sits at the low-score end; attainable scores are widely spaced there and through the threshold, crowding together only as $g_\ell \to 1$. The threshold $\hat{t}_q$ crosses on the steep rise, where the per-increment change is bounded below and the spacing of attainable scores is at least $\underline{s}$ (Assumption 6). This is the same quantity that gives the test power: a flat crossing would mean signal and null are indistinguishable there.

Assumption 6 (Discrimination at the threshold)**.** *There is $\underline{\sigma} > 0$ such that every null i with an attained score adjacent to $\hat{t}_q$ has per-increment log-likelihood-ratio change at the crossing $k_i^* = \arg\min_k |g_i(k) - \hat{t}_q|$ bounded below,*

$$\left|\Delta_k \log \mathrm{LR}_i(k_i^*)\right| = \left|\log \tfrac{(m_i - k_i^*)(k_i^* + a)}{(k_i^* + 1)(m_i - k_i^* - 1 + b)}\right| \geq \underline{\sigma},$$

equivalently the local score-grid step is bounded below, $s_i(\hat{t}_q) \geq \underline{s} := \hat{t}_q(1 - \hat{t}_q)\,\underline{\sigma}$.

This says the score crosses the threshold on a genuine slope, not a plateau. This condition is **not an extra regularity imposition on the model**: the count-space increment $\Delta_k \log \mathrm{LR}$ is precisely the local log-likelihood ratio between alternative and null, so the slope of the score at the cut *is* the local discriminability of signal from noise there. A flat crossing ($\underline{\sigma} = 0$) means P_1/P_0 is locally constant: adjacent counts carry identical evidence about null versus alternative, the cut separates nothing, and no thresholding procedure, ours or any other, has a quantity to control. Assumption 6 is therefore equivalent to the statement that the operating threshold sits where signal is detectable; it is implied by the existence of any distinguishable signal at the cut, and fails only in the degenerate regime where there is nothing to detect. For the Beta–Binomial alternative with $a \in (0, 1), b > 1$ the increment is bounded away from zero on the entire attainable interior, its only zero at the mode $k = 0$ (deep at the low-score end of the rejection region, not at the boundary), so any threshold in the informative region inherits a uniform $\underline{\sigma}$ automatically: qualitatively the bound is detectability, quantitatively it is supplied by the model rather than assumed against it. This sets Assumption 6 apart from the stability hypotheses introduced below (Assumptions 8 and 9), which constrain how the procedure responds to a single-coin perturbation.

A single coin flip perturbs the fitted surface, and hence every score, by a controlled amount $\delta_i := L\,\beta_N^{(i)} + L_b\,|\Delta\hat{b}|$, where $\beta_N^{(i)} = \|\hat{\boldsymbol{\alpha}}^{(i)} - \hat{\boldsymbol{\alpha}}\|_\infty$ is the sup-norm surface change from refitting with i's count flipped (Section E.6) and L, L_b are the Lipschitz constants of $\widehat{\mathrm{lfdr}}_{\mathrm{spatial}}$ in α, b. Counts live on a grid of $\sim m_i$ values, so the attainable scores are discrete, and near the threshold their spacing is bounded below by $\underline{s}$ (Assumption 6); when the drift δ_i is less than a quarter of that spacing it cannot push any score across the threshold, a *dead zone* in which the perturbation alters no decision, the discrete analogue of the structural invariance masking imposes. Unlike masking, we do not posit this as an assumption: both sides are characterized in this section, the drift by the influence operator and the gap by the Beta–Binomial geometry, and Lemma 9 (Section E.6) shows the sub-grid bound $\delta_i < \frac{1}{4}\underline{s}$ holds throughout the regime this framework targets and is checkable on the fitted surface.

Assumption 7 (Bounded boundary density)**.** *The score pairs do not pile up at the cutoff: there is $\rho < \infty$ with $\mathbb{E}\, N_\partial(w) \le \rho\, w\, N$ for all $w \le \frac{1}{2}\underline{s}$. Equivalently, $\mathbb{E}\,\#\{i \in \mathcal{H}_0 : |\breve{T}_i - \hat{t}_q| \le \delta_i\} = O(|\mathcal{R}_2|)$.*

This is a condition on the *population* of scores across the m_0 nulls. The Beta–Binomial structure makes it mild: each null's mass is spread over its $\sim m_i$ grid values (step bounded below by Assumption 6), so any single null lands within δ_i of the threshold with probability at most $1/(m_i+1)$, and the expected boundary count is at most $\sum_i 1/(m_i+1)$. What the assumption adds is that the realized surfaces $\hat{\alpha}(\mathrm{loc}_i)$ do not themselves cluster many distinct nulls at the cut, a cross-null condition the per-null grid cannot control (See Figure 5).

Assumption 8 (Threshold-stability floor; promotes App. E.9)**.** *A single flip $k_i \mapsto m_i - k_i$ moves the realized cutoff by $O(\delta)$: $|\hat{t}_q^{(i)} - \hat{t}_q| \le C_0\,\delta$, uniformly over $i \in \mathcal{H}_0$. Equivalently, the empirical ratio $r(t) = (1 + A(t))/(1 \vee R(t))$ crosses τ transversally, with slope $s_{\min} = \Theta(1)$ at the crossing; the two forms are* linked, not equivalent*, by $C_0 = \rho/s_{\min}$ (with ρ from Assumption 7). See Figure 8.*

This is the paper's hard floor asserted, not a routine regularity. Appendix E.9 identifies the simultaneous-perturbation stability of the global functional $\hat{t}_q$ as a hard floor and *declines* to bound it in closed form; Assumption 8 assumes precisely that bound. It is *not* implied by Assumptions 6–7: those control the *expected* count near the cutoff but not the realized flatness of r at τ, and on the discrete score lattice a flat stretch at τ has positive probability, so the crossing can lurch a full grid step under an arbitrarily small perturbation (Figure 8, right). We therefore treat it as a flagged hypothesis: Theorem 7 is conditional on it, and Theorem 8 gives the unconditional fallback that holds without it. Section E.9 discusses why the model structure favors it and where it is fragile.

The remaining regularity conditions (Assumption 2: residuals/observed informations uniformly bounded; Assumption 3: $\lambda \ge \lambda_0 > 0$, $\lambda_{\min}(\boldsymbol{K}) > 0$; Assumption 1: symmetric kernel with $\|\boldsymbol{K}\|_{\mathrm{op}} = O(1)$ and bounded effective degree ν) are exactly as in Section 3.2 and feed only the influence bound of Section E.6.

E.4 The $\delta = 0$ baseline (Steps 1–3): exact control for a flip-invariant surface

We first establish exact control for *any* surface that does not depend on the orientations $\{B_i\}$. This reproduces the Barber–Candès guarantee in count space and is the baseline the plug-in perturbs. Two facts are used repeatedly.

Lemma 4 (Deterministic ratio bound; exact, engine-free)**.** *At the realized cutoff, $1 + A \le \tau(1 \vee R)$, hence $A \le \tau R$ pointwise, and since $\tilde{V} \le A$, $\mathbb{E}[\tilde{V}] \le \mathbb{E}[A] \le \tau\,\mathbb{E}[R]$.*

Proof. $\hat{t}_q$ lies in the finite candidate set $\{T_i\} \cup \{\tilde{T}_i\}$, so the maximum in (7) is attained and $1 + A \le \tau(1 \vee R)$ holds at $\hat{t}_q$. If $R \ge 1$ this gives $A \le \tau R - 1 < \tau R$; if $R = 0$ then $1 + A \le \tau \le 1$ forces $A = 0 = \tau R$. Take expectations and use $\tilde{V} \le A$. No martingale or independence is used. □

Lemma 5 (Measure-preserving flip of a null)**.** *Fix $i \in \mathcal{H}_0$. Under Assumptions 4–5, the joint law of the counts is invariant under $k_i \mapsto m_i - k_i$ (all other counts fixed); hence for any measurable functional Φ, $\mathbb{E}[\Phi(\textit{counts})] = \mathbb{E}[\Phi(\textit{flip}_i\ \textit{counts})]$.*

Proof. Condition on labels and sizes. Given these, k_i is independent of $\{k_j\}_{j \ne i}$ (Assumption 5) and uniform (Assumption 4), hence invariant under $k \mapsto m_i - k$; the conditional joint law is unchanged, and averaging over labels gives the claim. □

Step 1: Discrete null symmetry. By Assumption 4, for every $j \in \{0, \dots, m_i\}$,

$$P(k_i = j \mid H_{0,i}) = P(k_i = m_i - j \mid H_{0,i}) = \tfrac{1}{m_i+1}. \tag{132}$$

Defining $B_i = \mathbb{1}(k_i > m_i/2)$ (ties at $k_i = m_i/2$ broken by an independent fair coin), (132) gives that, conditional on the fold $\breve{k}_i$, B_i is $\mathrm{Bernoulli}(1/2)$, and by Assumption 5 the $\{B_i\}_{i \in \mathcal{H}_0}$ are mutually independent. Consequently, for any surface α° *not depending on* $\{B_i\}_{i \in \mathcal{H}_0}$, the rejection-mirror pair is exchangeable for nulls,

$$(T_i, \tilde{T}_i) \mid (\alpha^\circ, H_{0,i}) \overset{d}{=} (\tilde{T}_i, T_i) \mid (\alpha^\circ, H_{0,i}). \tag{133}$$

This is the count-lattice replacement for the continuous mirror $p_i \leftrightarrow 1 - p_i$; it is exact for a surface external to the flips and only approximate for the plug-in (Section E.6).

Step 2: FDP decomposition. With $V, \tilde{V}$ as above,

$$\mathrm{FDP}(\mathcal{R}_2) = \frac{V}{1 \vee R} \leq \underbrace{\frac{1+\tilde{V}}{1 \vee R}}_{\leq \tau \text{ by (7)}} \cdot \frac{V}{1+\tilde{V}}, \tag{134}$$

the first factor $\leq \tau$ by construction of $\hat{t}_q$ (restricting its numerator to nulls only tightens it). It therefore suffices, for exact control, to show $\mathbb{E}[V/(1+\tilde{V})] \leq 1$.

Step 3: Fair-coin reduction. Order the nulls by folded score $\breve{T}_{(1)} \leq \cdots \leq \breve{T}_{(m_0)}$ and let J be the number with $\breve{T}_{(i)} \leq \hat{t}_q$. For a flip-invariant surface, $\breve{T}_i$ is a function of $(\alpha^\circ, \breve{k}_i)$ and $\breve{k}_i \perp B_i$ by (132), so the folded ordering is *independent of the orientations* $\{B_i\}$, and conditional on it the $B_{(i)}$ are i.i.d. $\mathrm{Bernoulli}(1/2)$. With $V = \sum_{i \leq J}(1 - B_{(i)})$, $\tilde{V} = \sum_{i \leq J} B_{(i)}$ and J a stopping time on the backward filtration $\mathcal{F}_j = \sigma(\sum_{i \leq j} B_{(i)}, B_{(j+1)}, \ldots, B_{(m_0)})$, the Barber–Candès reverse-martingale [4] gives

$$\mathbb{E}\left[\frac{1+J}{1+\sum_{i \leq J} B_{(i)}}\right] \leq 2, \qquad \text{hence} \qquad \mathbb{E}\left[\frac{V}{1+\tilde{V}}\right] \leq 1. \tag{135}$$

With (134) this yields $\mathrm{FDR}(\mathcal{R}_2) \leq \tau$ *exactly*, for any flip-invariant surface. *The sign-independence of the folded ordering is exactly what the reverse-martingale needs, and exactly what the deployed plug-in forfeits* (Section E.6).

E.5 The folded construction: a concrete flip-invariant surface with exact control

The reflection $k \leftrightarrow m - k$ fixes the folded count $\breve{k}_i$. Refitting the surface on the folded data realizes the $\delta = 0$ baseline concretely. Replace the count-space likelihoods by their flip-symmetrized forms

$$P_0^{\mathrm{f}}(\breve{k}) = \tfrac{2}{m+1}\mathbb{1}[\breve{k} \neq m/2] + \tfrac{1}{m+1}\mathbb{1}[\breve{k} = m/2], \quad P_1^{\mathrm{f}}(\breve{k}) = P_1(\breve{k}) + P_1(m-\breve{k})\mathbb{1}[\breve{k} \neq m/2], \tag{136}$$

and solve (6) with $(P_{0,i}, P_{1,i})$ replaced by $(P_{0,i}^{\mathrm{f}}, P_{1,i}^{\mathrm{f}})$ at $\breve{k}_i$, with the shape also fit on folded counts, $\hat{b}^{\mathrm{f}} = \hat{b}(\{\breve{k}_j\})$. Write $\hat{\alpha}^{\mathrm{f}}$ for the result and $\hat{\alpha}^{\mathrm{u}} = \hat{\alpha}$ for the plug-in. Folding is a deterministic many-to-one map; the pushforward of the mixture is again a two-group mixture with the *same* mixing weight $\alpha(\mathrm{loc}_i)$ and components (136), so the folded model is correctly specified with the same estimand (identifiable because the one-sided alternative keeps its mass at $\breve{k} \approx 0$, distinct from the flat folded null).

Lemma 6 (Flip-invariance of the folded rule)**.** *Under* (136)*, the flip $k_i \leftrightarrow m_i - k_i$ leaves every folded count $\breve{k}_j$ unchanged (it fixes $\breve{k}_i$ and does not touch $j \neq i$), so $\hat{\alpha}^{\mathrm{f}}$ and $\hat{b}^{\mathrm{f}}$ are functions of $\{\breve{k}_j\}_{j=1}^N$ only, hence invariant under the flip of any null's count. The flip merely swaps $T_i \leftrightarrow \tilde{T}_i$ and fixes all $(T_j, \tilde{T}_j)$, $j \neq i$.*

Proposition 5 (Exact control of the folded rule)**.** *Under Assumptions 4–5, Rule 2 run with the folded score controls FDR with no slack and non-asymptotically: $\mathrm{FDR}(\mathcal{R}_2^{\mathrm{f}}) \leq \tau$, at every $m_i \geq 1$. Validity requires* only *flip-invariance (Lemma 6),* not *correct specification of the folded model.*

Proof. By Lemma 6 the folded surface is external to the orientations, so Step 3 applies verbatim: the folded ordering is independent of $\{B_i\}$, the reverse-martingale gives $\mathbb{E}[V/(1+\tilde{V})] \leq 1$, and (134) closes $\mathrm{FDR}(\mathcal{R}_2^{\mathrm{f}}) \leq \tau$. At $m_i = 1$ the construction is valid: $k_i \in \{0, 1\}$ each w.p. $1/2$, the mirror $0 \leftrightarrow 1$ preserves the null, and a self-mirror count $k_i = m_i/2$ (m_i even, $T_i = \tilde{T}_i$) is inert in the ratio (7) and oriented by the independent tie-break coin, leaving the fair-coin structure intact. Nowhere is it used that $\hat{\alpha}^{\mathrm{f}}$ is consistent or correctly specified; an arbitrarily misspecified flip-invariant surface still gives $\mathrm{FDR} \leq \tau$, exactly as Rule 1's ordering may be misspecified without affecting its validity under (PC). □

This is the unconditional rock of the section. It needs no discrimination, resolution, boundary-density, or threshold-stability condition: only the discrete null symmetry and conditional independence. The remaining registers ask what is gained, and at what cost, **by *not* folding.**

Why the folded rule is not the deployed method ? Here we analyze the justification for not using the folded procedure (which will give exact guarantees). Folding is a parameter-free coarsening, so by the data-processing inequality it cannot increase Fisher information about α. With $d(k) = P_0(k) - P_1(k)$ and f the mixture pmf,

$$I^{\mathrm{u}} - I^{\mathrm{f}} = \sum_{k<m/2} \left[\tfrac{d(k)^2}{f(k)} + \tfrac{d(m-k)^2}{f(m-k)} - \tfrac{(d(k)+d(m-k))^2}{f(k)+f(m-k)} \right] \geq 0, \tag{137}$$

each summand ≥ 0 by Cauchy–Schwarz, concentrated in the small-k discovery cells and not vanishing with m. For the canonical alternative ($a = 0.5$, $b = 6$, $\alpha = 0.5$):

m	1	2	4	6	10	20	50
$I^{\mathrm{f}}/I^{\mathrm{u}}$	0.00	0.20	0.37	0.44	0.51	0.57	0.62

At $m = 1$ the folded count is constant and $I^{\mathrm{f}} = 0$: the folded surface carries no local information and reverts to $\bar{\alpha}$. At the small m this framework targets, folding discards more than half the per-hypothesis information and the ratio never approaches 1. The exact construction is therefore information-starved precisely in the regime that motivates the method, which is why we deploy the plug-in and characterize its slack (Sections E.6–E.8) rather than fold. The surface gap $\hat{\alpha}^{\mathrm{f}} - \hat{\alpha}^{\mathrm{u}} = (2\lambda \boldsymbol{I} + \boldsymbol{K}\boldsymbol{V}^{\mathrm{f}})^{-1}\boldsymbol{K}\mathbf{g}$ (with $g_i = s_i^{\mathrm{f}} - s_i^{\mathrm{u}}$ the change in null-evidence score) is computable from one extra fit and serves as a per-dataset diagnostic: a small gap licenses the plug-in; a large gap flags reliance on Rule 1.

E.6 Turning on the plug-in: the perturbation δ

The deployed surface is fit on the same counts it scores, so the independence Step 3 requires fails. We quantify the failure by a single-flip influence bound, then state precisely why exact FDR does not survive and what we claim instead.

Step 4: the single-flip influence bound.

Lemma 7 (Uniform stability of the plug-in surface)**.** *Fix $i \in \mathcal{H}_0$. Write the fitted surface as $\hat{\alpha} = \hat{\bar{\alpha}} + \hat{f}$, with $\hat{f} \in \mathcal{H}_K$ the minimizer of* (6)*, and let $\hat{\alpha}^{(i)} = \hat{\bar{\alpha}}^{(i)} + \hat{f}^{(i)}$ be the full refit with k_i replaced by $m_i - k_i$, the global scalars $(\hat{\bar{\alpha}}, \hat{b})$ refit as well. For each j, let $g_j(u; \bar{\alpha}, b) = -\log\left[(\bar{\alpha}+u)P_{0,j} + (1-\bar{\alpha}-u)P_{1,j}(b)\right] + \gamma\,\Lambda_{[0,1]}(\bar{\alpha}+u)$ be its per-hypothesis term, convex in u. Assume (i) Assumption 1, so $\kappa^2 := \max_j K_{jj} = 1$; (ii) $|\partial_u g_j| \leq L_\ell$ between the two fits (on $[0,1]$, where the boundary term vanishes, $\partial_u g_j$ is the residual weight w_j, so Assumption 2 gives $L_\ell \leq c_2$); (iii) $\partial_u g_j$ is C_2-Lipschitz in $(\bar{\alpha}, b)$ on the same set; (iv) pooled-scalar stability, $|\Delta\hat{\bar{\alpha}}| + |\Delta\hat{b}| \leq C_s/N$. Then, for every N and every $\lambda > 0$,*

$$\beta_N^{(i)} = \|\hat{\boldsymbol{\alpha}}^{(i)} - \hat{\boldsymbol{\alpha}}\|_\infty \;\leq\; \frac{\kappa^2\,(L_\ell + C_2 C_s/2)}{\lambda} + \frac{C_s}{N}. \tag{138}$$

Proof. This is the uniform-stability argument of Bousquet and Elisseeff [8] for regularized risk minimization in an RKHS, applied to the summed objective. Let J and $J^{(i)}$ be the objectives before and after the flip as functions of f, and $\Delta = \hat{f}^{(i)} - \hat{f}$. Both are convex data terms plus $\lambda\|f\|_{\mathcal{H}_K}^2$, hence 2λ-strongly convex, so at their minimizers $J(\hat{f}^{(i)}) - J(\hat{f}) \geq \lambda\|\Delta\|_{\mathcal{H}_K}^2$ and $J^{(i)}(\hat{f}) - J^{(i)}(\hat{f}^{(i)}) \geq \lambda\|\Delta\|_{\mathcal{H}_K}^2$. Adding the two, the RKHS penalty cancels and, with $h_j = g_j - g_j^{(i)}$,

$$2\lambda\|\Delta\|_{\mathcal{H}_K}^2 \;\leq\; \sum_j \left[h_j(\hat{f}^{(i)}(\mathrm{loc}_j)) - h_j(\hat{f}(\mathrm{loc}_j))\right] \;\leq\; \kappa\|\Delta\|_{\mathcal{H}_K} \sum_j \sup|h_j'|,$$

by the reproducing property $|\Delta(\mathrm{loc})| \leq \kappa\|\Delta\|_{\mathcal{H}_K}$. Hence $\|\Delta\|_\infty \leq \kappa\|\Delta\|_{\mathcal{H}_K} \leq \frac{\kappa^2}{2\lambda}\sum_j \sup|h_j'|$. The flipped term contributes $\sup|h_i'| \leq 2L_\ell$ by (ii); each of the other $N-1$ terms changes only through the refit scalars, so $\sup|h_j'| \leq C_2(|\Delta\hat{\bar{\alpha}}| + |\Delta\hat{b}|) \leq C_2 C_s/N$ by (iii)–(iv). Summing gives $\|\Delta\|_\infty \leq \kappa^2(L_\ell + C_2C_s/2)/\lambda$, and $\beta_N^{(i)} \leq |\Delta\hat{\bar{\alpha}}| + \|\Delta\|_\infty$ gives (138). □

Remark 2 (Normalization and the rate in N)**.** *Objective* (6) *sums the N data terms, whereas Bousquet and Elisseeff [8] average them; their regularization $\tilde{\lambda}$ therefore corresponds to our $\lambda = N\tilde{\lambda}$, and their $O(1/N)$ rate is our $O(1/\lambda)$ with λ growing linearly in N. In our experiments λ is tuned by cross-validation to a value that does not grow with N, so the spatial channel is $O(1/\lambda)$ and does not vanish as N grows; it is fixed by λ and computable on the data by a single refit (Lemma 9). Only the direct dependence of the score on $\hat{b}$ is $O(1/N)$.*

Lemma 7 bounds the drift uniformly; the first-order expansion below shows where it lands. Fix $i \in \mathcal{H}_0$; let $\hat{\boldsymbol{c}}, \hat{\boldsymbol{c}}^{(i)}$ be the kernel coefficients before/after flipping i's count. Because P_0 is flip-invariant, only the alternative-likelihood term changes, and a first-order Newton step gives

$$\hat{\boldsymbol{c}}^{(i)} - \hat{\boldsymbol{c}} = -\boldsymbol{H}^{-1}\boldsymbol{K}_{\cdot,i}\,\Delta w_i + O(\|\cdot\|^2), \qquad \boldsymbol{H} = \boldsymbol{K}\boldsymbol{V}\boldsymbol{K} + 2\lambda\boldsymbol{K}, \quad \boldsymbol{V} = \mathrm{diag}(\hat{v}), \tag{139}$$

with $|\Delta w_i| \le c_2$ (Assumption 2). The induced surface change at j is governed by the influence operator

$$\boldsymbol{M} := \boldsymbol{K}\boldsymbol{H}^{-1}\boldsymbol{K} = \boldsymbol{K}(\boldsymbol{V}\boldsymbol{K} + 2\lambda\boldsymbol{I})^{-1} \succeq 0, \qquad \|\boldsymbol{M}\|_{\mathrm{op}} \le \frac{\|\boldsymbol{K}\|_{\mathrm{op}}}{2\lambda}, \tag{140}$$

the identity following from $\boldsymbol{H} = \boldsymbol{K}(\boldsymbol{V}\boldsymbol{K}+2\lambda\boldsymbol{I})$ and the norm bound from $\boldsymbol{V}\boldsymbol{K} \succeq 0$ (its eigenvalues are those of the PSD $\boldsymbol{V}^{1/2}\boldsymbol{K}\boldsymbol{V}^{1/2}$). To first order, $\beta_N \approx \max_j |(\boldsymbol{M})_{ji}\Delta w_i|$: the drift at j is set by the (j,i) entry of the influence operator, largest at and near the flipped hypothesis and larger where i has few neighbors, consistent with the $O(1/\lambda)$ worst case of Lemma 7. The global scalars pool all N counts, so a single flip moves them by $O(1/N)$. The score is Lipschitz in (α, b), so every score moves by at most

$$|T_j^{(i)} - T_j| \le L\beta_N + L_b|\Delta\hat{b}| =: \delta = O(1/\lambda) + O(1/N), \qquad |\Delta\hat{b}| = O(1/N), \tag{141}$$

the spatial channel contributing $O(1/\lambda)$ (RKHS coefficients) and the pooled-scalar channel $O(1/N)$.

Why exact FDR does not survive (a structural obstruction). A flip $k_i \mapsto m_i - k_i$ has two effects (Figure 6). For the flipped hypothesis i, its score jumps from T_i to $\tilde{T}_i$, the benign rejection/mirror swap the Barber–Candès argument is built on. For *every other* hypothesis ℓ, the count k_ℓ is unchanged but refitting the surface drifts $g_\ell(k_\ell)$ by at most δ. This collateral drift couples the folded ordering to the orientations $\{B_i\}$: the independence Step 3 relies on is gone, the reverse-martingale from B-C no longer applies. Conditioning each null on its score magnitude and all other scores (the leave-one-out route for non-exchangeable statistics) degenerates here, because the conditioning vector is itself a function of the orientation; the conditional symmetry index is not well-posed for deterministic, discrete plug-in scores. We therefore do *not* claim exact finite-sample FDR for the plug-in. The obstruction is intrinsic to using the data both to learn the score and to test with it.

Remark 3 (Plug-in finite-sample FDR)**.** *The plug-in relaxes the exact control of Proposition 5 by an amount governed by the single-flip stability $\delta = O(1/\lambda) + O(1/N)$ (eq.* (141)*), and this section controls its FDR at three strengths, each tied to an explicit hypothesis. Exact and unconditional: the folded rule attains $\mathrm{FDR} \le \tau$ with no stability hypothesis (Proposition 5), at the cost of the "information tax". Per-realization, conditional on ranking stability: for the deployed plug-in, $\mathrm{FDR} \le \tau + C_{\mathrm{BC}}\delta$ holds under Assumption 9 (Theorem 6), the rank-space sibling of the threshold-stability floor. Averaged, conditional on threshold stability: the marginal FDR satisfies $\mathrm{mFDR} \le \tau + O(\delta)$ under Assumption 8 (Theorem 7), with an unconditional fallback (Theorem 8). Absent either stability hypothesis, what survives is the unconditional grid-scale bound $\mathrm{mFDR} \le \tau + O(\underline{s})$ (Theorem 8); the drift δ itself remains computable on the data by a single refit. Trading the provably-exact but lower-power folded rule for the near-exact, higher-power plug-in is the design choice Rule 2 makes deliberately; the two stability hypotheses are what convert "near-exact" into a quantified FDR bound, per-realization and averaged respectively.*

The dead zone makes the collateral drift inert for all but a thin boundary set, which is what the mFDR proof exploits.

Lemma 8 (Score-grid step from the Beta–Binomial)**.** *For a null i with $g_i(k) = (1 + \rho_i(k))^{-1}$, $\rho_i(k) = \frac{1-\alpha_i}{\alpha_i}\mathrm{LR}_i(k)$, the spacing of attainable scores is $|g_i(k+1) - g_i(k)| = g_i(k)(1 - g_i(k))\,|\Delta_k \log \mathrm{LR}_i(k)|(1 + o(1))$, with $\Delta_k \log \mathrm{LR}_i(k) = \log \frac{(m_i-k)(k+a)}{(k+1)(m_i-k-1+b)}$. For $a \in (0,1), b > 1$ this is bounded away from 0 on the attainable interior (the only zero is at the mode $k = 0$), so at an interior threshold $s_i(\hat{t}_q) \ge \hat{t}_q(1 - \hat{t}_q)\underline{\sigma} = \underline{s}$, recovering Assumption 6.*

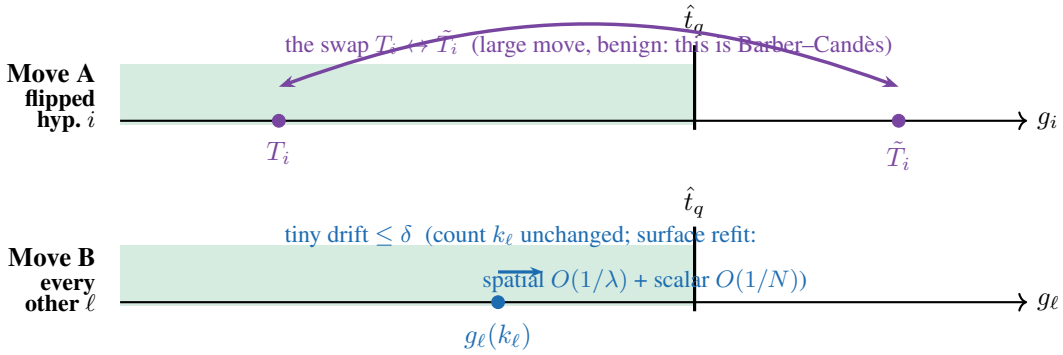


Figure 6: **The two effects of a single flip.** Flipping i's count does two different things. *Move A* (top): for the flipped hypothesis, its score jumps the full distance $T_i \to \tilde{T}_i$, a large move, but exactly the rejection/mirror swap the Barber–Candès argument is built on. *Move B* (bottom): for every *other* hypothesis ℓ, the count is unchanged but refitting the surface drifts its score by at most $\delta = L\beta_N + L_b|\Delta\hat{b}|$, through a spatial channel ($O(1/\lambda)$) and a pooled-scalar channel ($O(1/N)$). Move B couples the ranking to the orientations and is why the reverse-martingale of Step 3 fails for the plug-in; its effect on any one decision is governed by the dead zone (Figure 7).

Proof. $g(k{+}1){-}g(k) = \frac{\rho(k){-}\rho(k{+}1)}{(1{+}\rho(k))(1{+}\rho(k{+}1))}$, and to first order $\rho(k{+}1){-}\rho(k) = \rho(k)(e^{\Delta_k \log \mathrm{LR}}{-}1)$; using $g = (1+\rho)^{-1}$, $1-g = \rho/(1+\rho)$ gives the product form as a two-sided bound. The Beta–Binomial pmf is log-concave with a single mode; for $a \in (0,1), b > 1$ the mode is at $k = 0$, so the log-LR increment has no interior zero and is bounded away from 0 across the attainable interior. □

The dead zone of the preceding paragraph needs one quantitative fact: that a single flip moves the scores by less than the room available before the nearest attainable score reaches the threshold. There are two competing scales. The room is the score-grid spacing at the cut, $\underline{s}$, a property of the model that does not shrink with the sample size. The flip's reach is the surface drift δ_i, which does shrink: one count out of N can only move a regularized surface a little. The lemma is just the statement that the second is eventually smaller than the first, with both sides quantified, so the dead zone is not assumed but earned, and can moreover be checked directly on a fitted surface by flipping a count and measuring how far the fit moves.

Lemma 9 (Resolution: the single flip is sub-grid)**.** *The single-flip drift is small relative to the score-grid spacing at the threshold:*

$$\frac{\delta_i}{\underline{s}} = O(1/\lambda) + O(1/N), \qquad \textit{so} \qquad \delta_i < \tfrac{1}{4}\underline{s}$$

for every near-threshold null i, once λ exceeds a constant fixed by $\underline{s}$, L and the constants of Lemma 7, and N one fixed by $\underline{s}$, L_b and C_s (in particular under Assumption 3 with λ_0 large enough relative to $\underline{s}$). At a cross-validated λ that does not grow with N, this is a condition on the chosen λ rather than an asymptotic one, checked on the data by the refit described below. Here $\underline{s} = \hat{t}_q(1{-}\hat{t}_q)\underline{\sigma} = \Theta(1)$ is the grid gap at the threshold (Lemma 8), and the drift $\delta_i = L\,\beta_N^{(i)} + L_b\,|\Delta\hat{b}|$ splits into an independent spatial channel ($\beta_N^{(i)} = O(1/\lambda)$, Lemma 7) and pooled-scalar channel ($|\Delta\hat{b}| = O(1/N)$); since larger λ stiffens the surface and shrinks $\beta_N^{(i)}$, the bound is monotone in the regularization. The drift δ_i is computable by refitting with i's count flipped and measuring $\|\hat{\boldsymbol{\alpha}}^{(i)} - \hat{\boldsymbol{\alpha}}\|_\infty$, so the condition is verifiable on the data rather than posited, which distinguishes it from the threshold- and ranking-stability hypotheses (Assumptions 8, 9), not checkable in closed form and able to fail even asymptotically.

Proof. The grid gap $\underline{s} = \Theta(1)$ is Lemma 8. The drift bound $\delta_i = L\beta_N^{(i)} + L_b|\Delta\hat{b}| = O(1/\lambda) + O(1/N)$, with $\beta_N^{(i)} = O(1/\lambda)$ from Lemma 7 and $|\Delta\hat{b}| = O(1/N)$ from the pooled scalars, is the single-flip influence bound eq. (141). Dividing by $\underline{s} = \Theta(1)$ gives $\delta_i/\underline{s} = O(1/\lambda) + O(1/N)$, and choosing each channel constant below $\frac{1}{8}\underline{s}$ yields $\delta_i < \frac{1}{4}\underline{s}$. □

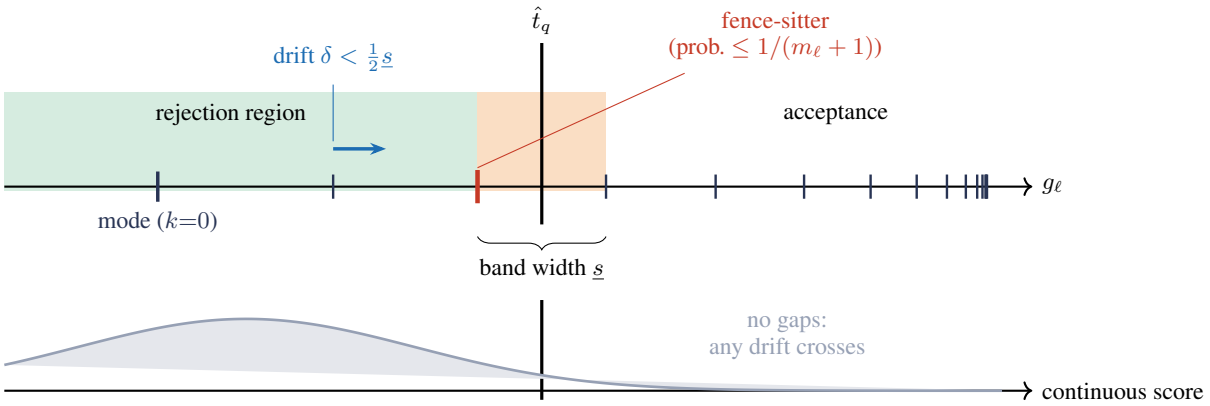


Figure 7: **The discrete dead zone.** Top: hypothesis ℓ's score takes only the $m_\ell + 1$ attainable values (navy ticks; $a = 0.5$, $b = 6$, $m_\ell = 14$, $\alpha_\ell = 0.6$), widest-spaced in the rejection region and bunching only as $g_\ell \to 1$; $\hat{t}_q$ falls in the well-spaced region (Lemma 8). A drift $\delta < \frac{1}{2}\underline{s}$ (blue) cannot move any score across the threshold; the only crossable score is the lone attainable value in the width-$\underline{s}$ band (red). Since k_ℓ is uniform and at most one value lands in the band, the chance ℓ is caught on the fence is $\leq 1/(m_\ell + 1)$ (Lemma 10). Bottom (*schematic*): continuous scores have no gaps, so any perturbation can cross, that the margin discreteness provides, is absent.

Lemma 10 (Discrete dead zone)**.** *If a perturbation shifts every score by at most η and the threshold by at most η' with $\eta + \eta' < \frac{1}{2}\underline{s}$ (Figure 7), then under Assumption 6 each null ℓ's decision $\mathbb{1}[g_\ell(k_\ell) \leq \hat{t}_q]$ is unchanged unless $g_\ell(k_\ell)$ is the unique attained value of ℓ within $\frac{1}{2}\underline{s}$ of $\hat{t}_q$. Hence under Assumption 4,* $\Pr(\ell$ *changes decision*$) \leq 1/(m_\ell + 1)$ *and* $\mathbb{E}\,\#\{$*decision changes from one flip*$\} \leq \sum_{\ell \in \mathcal{H}_0} 1/(m_\ell + 1)$.

Proof. By Assumption 6 (equivalently Lemma 8) the attained scores near $\hat{t}_q$ are $\underline{s}$-separated and $k \mapsto g_\ell(k)$ is strictly monotone there, so at most one attained value $v^\dagger$ lies in $(\hat{t}_q - \frac{1}{2}\underline{s}, \hat{t}_q + \frac{1}{2}\underline{s})$. A decision flips only if $g_\ell(k_\ell) - \hat{t}_q$ changes sign; the relative shift is $\leq \eta + \eta' < \frac{1}{2}\underline{s}$, so this requires $g_\ell(k_\ell) = v^\dagger$. By strict monotonicity a single count realizes $v^\dagger$, and $k_\ell \sim \text{Uniform}\{0, \ldots, m_\ell\}$ gives $\Pr \leq 1/(m_\ell + 1)$; sum over nulls. □

Remark 4 (Threshold attainment)**.** *$\hat{t}_q$ is always an attained score. For a perturbation of magnitude $t \in [0, \eta]$, let $t^* = \inf\{t :$ some decision changes$\}$; for $t < t^*$ the rejection set is the unperturbed one, so $\hat{t}_q$ tracks a single attained value and has moved by $\leq t < \eta$, whence the total relative shift is $< 2\eta < \frac{1}{2}\underline{s}$ and the grid separation forbids any crossing at t^*, a contradiction. So no decision changes for $t \leq \eta$ whenever $\eta < \frac{1}{4}\underline{s}$, which is the sub-grid bound of Lemma 9 ($\eta = \delta_i$).*

E.7 A per-realization guarantee: plug-in FDR under ranking stability

The averaged mFDR we prove next (Theorem 7, Section E.8) controls a *ratio of expectations*: $\mathbb{E}[V]/\mathbb{E}[R]$. This is the natural object when the slack is allowed to average over datasets, and it closes because the per-null proximity bound there linearizes null-by-null and the global denominator $\mathbb{E}[R] = \Theta(N)$ absorbs the boundary count. It does *not* by itself control the realized false discovery proportion on a single dataset, $\mathbb{E}[V/(1 + \tilde{V})]$, because that is an expectation *of* a ratio and the denominator $1 + \tilde{V}$ couples to the numerator through the shared cutoff.

This subsection supplies the per-realization statement. It costs one further hypothesis (Assumption 9 below), which is to the *ranking* what Assumption 8 is to the *threshold*: both assert that a single null's flip perturbs a global feature of the procedure by $O(\delta)$ rather than by an $O(1)$ lurch, and both are inherited from the same elementary fact (a flip moves the fitted surface, and hence every folded score, by at most δ). We state it, prove the resulting bound, and (as with Assumption 8) flag exactly what it carries.

Where the plug-in breaks the exact argument, and where it does not. Order the nulls by their folded scores and write $B_{(j)}$ for the orientation at rank j, $S_j = \sum_{i \le j} B_{(i)}$, and $M_j = (1+j)/(1+S_j)$. The exact ($\delta = 0$) proof of Step 3 runs the backward supermartingale: with the fold-invariant ranking the orientations are exchangeable given the partial sum, so the conditional boundary probability equals the running average,

$$\Pr\big(B_{(j)} = 1 \mid \mathcal{F}_j\big) = \frac{S_j}{j}, \qquad \mathcal{F}_j = \sigma\big(\alpha^\circ,\, S_j,\, B_{(j+1)}, \ldots, B_{(m_0)}\big), \tag{142}$$

M_j is a martingale, and optional stopping at the step-up index J gives $\mathbb{E}[M_J] \le 2$. The plug-in breaks (142) for one specific reason: the folded score $\breve{T}_i$ is computed from a surface fit on the same counts that determine B_i, so a null's rank and its orientation are functions of shared data and are correlated. Under that coupling the ones need not be uniformly placed among the bottom j ranks, and $\Pr(B_{(j)} = 1 \mid \mathcal{F}_j)$ can exceed S_j/j.

The coupling, however, can only *reorder* a null across the cutoff if that null's folded score sits within δ of $\hat{t}_q$: outside this band the surface drift δ moves the score but cannot change which side of $\hat{t}_q$ it falls on, so the null's sign and its above/below-threshold status are decoupled regardless of the drift. The defect in (142) is therefore carried entirely by the $O(\delta N)$ boundary-band nulls, the same dead-zone object as Figure 7 and Lemma 10, now read in the martingale's language. This is the precise sense in which "most nulls are far from the threshold" resolves the obstruction: the far nulls contribute to the count S_j but never to the boundary anomaly.

The exact one-step defect. A direct two-case expansion of M_{j-1} ($B_{(j)} = 1$ gives $M_{j-1} = j/S_j$; $B_{(j)} = 0$ gives $M_{j-1} = j/(1+S_j)$) yields, with $\epsilon_j := \Pr(B_{(j)} = 1 \mid \mathcal{F}_j) - S_j/j$, the identity

$$\mathbb{E}[M_{j-1} \mid \mathcal{F}_j] - M_j \;=\; \frac{j\,\epsilon_j}{S_j\,(1+S_j)} \;=:\; D_j. \tag{143}$$

So $\epsilon_j \le 0$ makes step j a supermartingale step, $\epsilon_j > 0$ leaks, and the masked/exchangeable case is $\epsilon_j \equiv 0$. Two structural facts about (143) do the work. First, the *baseline is exactly fair even under the coupling*: since $B_i = \mathbb{1}(k_i > m_i/2)$ depends only on k_i, each B_i is exactly $\mathrm{Bernoulli}(1/2)$ and independent across nulls (Assumptions 4–5), so the terminal sum $S_{m_0} \sim \mathrm{Bin}(m_0, 1/2)$ is order-independent and

$$\mathbb{E}[M_{m_0}] = \frac{1 - 2^{-(m_0+1)}}{1/2} \le 2 \tag{144}$$

holds for the plug-in verbatim: the coupling reorders which B_i sits where, but cannot move the total. All of the plug-in's cost is in the rank-resolved defects D_j, which are second order. Second, the denominator $S_j(1+S_j)$ supplies a $1/j$ weight: each boundary-band rank is charged once in the backward sum with weight $j/[S_j(1+S_j)] \approx 4/j$, rather than the boundary band being recharged at every visited rank. This is the mechanism by which the naive union-bound over-count, which would multiply the per-null $O(\delta)$ defect by the full $O(N)$ rank range, is avoided; the surviving factor is the harmonic sum $\sum 1/j$, logarithmic rather than linear in the range.

Assumption 9 (Ranking stability; rank-space sibling of Assumption 8)**.** *There is a constant $\rho_0 < \infty$ such that, at every rank j visited by the step-up,*

$$\big|\epsilon_j\big| = \Big|\Pr\big(B_{(j)} = 1 \mid \mathcal{F}_j\big) - \tfrac{S_j}{j}\Big| \;\le\; \rho_0\,\delta. \tag{145}$$

Theorem 6 (Per-realization FDR of the plug-in; conditional on Assumption 9)**.** *Under Assumptions 4–5 and Assumption 9, on the overwhelming-probability event $\{S_j = \Theta(j)$ at visited ranks$\}$,*

$$\mathrm{FDR}(\mathcal{R}_2) \;\le\; \tau + C_{\mathrm{BC}}\,\delta \;=\; \tau + O(1/\lambda) + O(1/N), \qquad C_{\mathrm{BC}} = 4\rho_0\,\overline{\ln(m_0/J)},$$

where J is the step-up index and $\overline{\ln(m_0/J)}$ its expectation. With a constant null fraction below threshold ($J = \Theta(m_0)$) the logarithmic factor is $O(1)$ and the fair-coin constant $C_{\mathrm{BC}} = 4$ is recovered as $\rho_0, \overline{\ln(m_0/J)} \to O(1)$.

Proof. By optional stopping applied to (143) and the exact baseline (144),

$$\mathbb{E}[M_J] = \mathbb{E}[M_{m_0}] + \mathbb{E}\Big[\textstyle\sum_{j=J+1}^{m_0} D_j\Big] \le 2 + \mathbb{E}\Big[\textstyle\sum_{j=J+1}^{m_0} (D_j)_+\Big].$$

Under Assumption 9, $(D_j)_+ \le j\rho_0\delta/[S_j(1+S_j)]$, and on the event $S_j = \Theta(j)$ (its complement has probability $\le m_0 e^{-cN} = o(\delta)$ and contributes $o(\delta)$ to the bound since $M_J \le 1 + m_0$ always),

$$\sum_{j=J+1}^{m_0} (D_j)_+ \;\le\; \rho_0\delta \sum_{j=J+1}^{m_0} \frac{j}{S_j(1+S_j)} \;\lesssim\; \rho_0\delta \sum_{j=J+1}^{m_0} \frac{4}{j} \;=\; 4\rho_0\delta\,\ln\frac{m_0}{J}.$$

Hence $\mathbb{E}[M_J] \le 2 + 4\rho_0\,\overline{\ln(m_0/J)}\,\delta = 2 + C_{\mathrm{BC}}\delta$. Since $V/(1+\tilde{V}) = M_J - 1$ and the Step-2 decomposition (134) gives $\mathrm{FDP} \le \tau \cdot V/(1+\tilde{V})$,

$$\mathrm{FDR}(\mathcal{R}_2) = \mathbb{E}[\mathrm{FDP}] \le \tau\,\mathbb{E}[M_J - 1] \le \tau(1 + C_{\mathrm{BC}}\delta) \le \tau + C_{\mathrm{BC}}\delta. \qquad \square$$

Remark 5 (What Assumption 9 carries, in plain terms)**.** *In words, Assumption 9 says:* flipping one null's coin does not suddenly reshuffle the order of hypotheses at the decision boundary. *The fitted surface depends on all the counts, so flipping a single null nudges every folded score, but only by at most δ. If two nulls sitting next to the cutoff are separated in score by more than δ (the discrimination of Assumption 6), that nudge cannot swap their order or carry either across $\hat{t}_q$, so the rank at the boundary behaves as if the surface had not read the flipped coin, and the boundary orientation stays fair up to $O(\delta)$.* ***The assumption fails*** *only where folded scores* pile up within δ of one another*: there a single flip can reorder nulls and the boundary anomaly ϵ_j can jump to $O(1)$. The discrimination (Assumption 6) and bounded boundary density (Assumption 7) confine that dense regime to the $O(\delta N)$ band, where the martingale's $1/j$ weight already discounts it, so the assumption is a transversality (no-coincidence) condition, mild for the same model reasons that make Assumption 8 mild (Beta–Binomial separation, Fisher rigidity of the fit), and fragile in the same corner: isolated hypotheses with weak signal (m_i small), where a flip can move a score across a sparse neighbourhood.*

Three qualifications distinguish this register from the mFDR one: (i) *Assumptions 8 and 9 are similar but not the same: the former controls the cutoff's response to a flip, the latter the conditional boundary sign-probability's response. Neither implies the other and neither is derivable from Assumptions 6–7 in the sure form* (145)*; both are flagged hypotheses.* (ii) *The constant carries a $\overline{\ln(m_0/J)}$ factor absent from the mFDR bound. Under a strong signal ($J \ll m_0$) this grows, but only as $O(\delta \ln m_0)$, a genuine but only logarithmic cost.* (iii) *Like Assumption 8, Assumption 9 admits a probabilistic relaxation, bounding $\Pr(|\epsilon_j| > \rho_0\delta)$ at the boundary rather than the sure bound, which suffices for the same conclusion up to an additive $O(\Pr(\cdot)\cdot\mathbb{E}[M_J])$ term and is the form we would verify empirically.*

E.8 The averaged guarantee: plug-in mFDR

We now control mFDR for the plug-in at the slack rate. The argument is elementary: a slack identity, a per-null bound by own-proximity (using the measure-preserving flip and Assumption 8), and the deterministic ratio bound. It uses none of the machinery the per-realization FDP route required.

Lemma 11 (Slack identity; exact)**.** $\mathbb{E}[V] - \mathbb{E}[\tilde{V}] = \sum_{i\in\mathcal{H}_0} \mathbb{E}\big[\mathbb{1}\!\!\not\!\!\leq(i \textit{ active})(1-2B_i)\big]$.

Proof. For each null, $\mathbb{1}\!\!\not\!\!\leq(T_i \le \hat{t}_q) - \mathbb{1}\!\!\not\!\!\leq(\tilde{T}_i \le \hat{t}_q)$ is $+1$ on $\{T_i \le \hat{t}_q < \tilde{T}_i\}$, -1 on $\{\tilde{T}_i \le \hat{t}_q < T_i\}$, 0 otherwise; both nonzero cases require i active, with sign $+1$ iff the observed side is the smaller one ($B_i = 0$). So the integrand is $\mathbb{1}\!\!\not\!\!\leq(i \text{ active})(1-2B_i)$; sum and take expectations. $\square$

Lemma 12 (Each null's slack term is controlled by its own proximity)**.** *Under Assumptions 4–6 and 8, for every $i \in \mathcal{H}_0$,*

$$\big|\mathbb{E}[\mathbb{1}\!\!\not\!\!\leq(i \textit{ active})(1-2B_i)]\big| \le \tfrac{1}{2}\Pr(\eta_i \le c\,\delta), \qquad c = 1 + C_0.$$

Proof. Let $a_i = \mathbb{E}[\mathbb{1}\!\!\not\!\!\leq(i \text{ active})(1-2B_i)]$ and apply Lemma 5 with $\Phi = \mathbb{1}\!\!\not\!\!\leq(i \text{ active})(1-2B_i)$. On the flipped data the orientation flips ($B_i \mapsto 1 - B_i$, so $1 - 2B_i \mapsto -(1-2B_i)$) and the active indicator becomes $\{i \text{ active}\}^{(i)}$; thus $a_i = -\mathbb{E}[\mathbb{1}\!\!\not\!\!\leq(\{i \text{ active}\}^{(i)})(1-2B_i)]$, and averaging,

$$a_i = \tfrac{1}{2}\mathbb{E}\big[(\mathbb{1}\!\!\not\!\!\leq(i \text{ active}) - \mathbb{1}\!\!\not\!\!\leq(\{i \text{ active}\}^{(i)}))(1-2B_i)\big], \qquad |a_i| \le \tfrac{1}{2}\Pr\big(\{i \text{ active}\} \ne \{i \text{ active}\}^{(i)}\big).$$

The active interval $[\breve{T}_i, \widetilde{\breve{T}}_i)$ has endpoints that are i's own *unordered* pair $\{\min, \max\}$ of scores; the flip moves these by at most δ (eq. (141) applied to the 1-Lipschitz min/max, which is the *surface drift* of the pair, not the observed swap $T_i \leftrightarrow \tilde{T}_i$, which is an $O(1)$ reordering within the fixed pair and is invisible to the active indicator). By Assumption 8 the cutoff moves by at most $C_0\delta$. For the two indicators to differ, $\hat{t}_q$ (or $\hat{t}_q^{(i)}$) must lie within $\delta + C_0\delta = c\delta$ of an endpoint, i.e. $\eta_i \le c\delta$. □

Notice, the slack (Lemma 11) is a sum over hypotheses of a signed indicator, **not** a sum over flips of cross-crossings. Lemma 12 bounds each term by i's *own* distance to the cutoff, because the active interval's endpoints are i's own scores and the cutoff moves $O(\delta)$ *by Assumption 8*. No other hypothesis enters, and no influence-operator locality is invoked, so the spatial channel's decay rate, and with it any λ^{-d}, never appears. The pooled scalars enter only through the $O(1/N)$ part of δ; **their non-local action on the global cutoff is contained by Assumption 8, not by the composition of δ.**

Theorem 7 (mFDR control of the plug-in rule; conditional on Assumption 8)**.** *Under Assumptions 4–8 and $\mathbb{E}[R] = \Theta(N)$,*

$$\mathrm{mFDR}(\mathcal{R}_2) \le \tau + \frac{c\rho}{2}\,\delta = \tau + O(\delta) = \tau + O(1/\lambda) + O(1/N), \qquad c = 1 + C_0 = 1 + \rho/s_{\min}.$$

The pooled-scalar channel's share of the slack is $O(1/N)$; a flip-invariant surface with fold-fit scalars gives $\delta = 0$ and $\mathrm{mFDR} \le \tau$.

Proof. By Lemma 11, then Lemma 12, then summing and using Assumption 7,

$$\mathbb{E}[V] - \mathbb{E}[\tilde{V}] \le \sum_{i \in \mathcal{H}_0} \tfrac{1}{2}\Pr(\eta_i \le c\delta) = \tfrac{1}{2}\mathbb{E}[N_\partial(c\delta)] \le \tfrac{1}{2}\rho c\,\delta N.$$

With Lemma 4 ($\mathbb{E}[\tilde{V}] \le \tau\mathbb{E}[R]$), $\mathbb{E}[V] \le \tau\mathbb{E}[R] + \frac{1}{2}\rho c\,\delta N$, so $\mathrm{mFDR} = \mathbb{E}[V]/\mathbb{E}[R] \le \tau + \frac{\rho c}{2}\delta \cdot N/\mathbb{E}[R] = \tau + O(\delta)$ by $\mathbb{E}[R] = \Theta(N)$. □

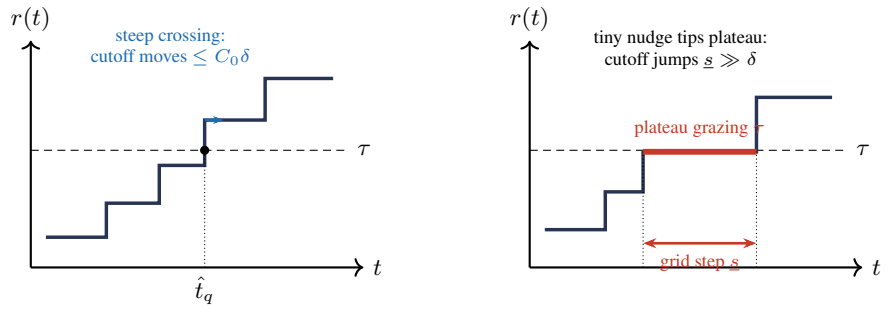
Theorem 8 (Unconditional fallback; without Assumption 8)**.** *Under Assumptions 4–7 and $\mathbb{E}[R] = \Theta(N)$* but not *Assumption 8, the cutoff move under a single flip is bounded only by the grid scale, $|\hat{t}_q^{(i)} - \hat{t}_q| \le \underline{s}$ ($\hat{t}_q$ is an attained value), and Lemma 12 holds with the window $c\delta$ widened to $O(\underline{s})$, giving*

$$\mathrm{mFDR}(\mathcal{R}_2) \le \tau + O(\underline{s}),$$

the paper's existing grid-scale hard-floor bound. Alternatively, if the kernel decays exponentially (or with spatial-decay exponent $> d$), the lurch frequency can be bounded by resolvent locality of $\boldsymbol{V}\boldsymbol{K} + 2\lambda\boldsymbol{I}$, whose localization length scales as $\|\boldsymbol{V}\boldsymbol{K}\|_{\mathrm{op}}/(2\lambda)$, yielding $\mathrm{mFDR}(\mathcal{R}_2) \le \tau + \tilde{O}(\lambda^{-d}\delta)$, which degrades as $\lambda \to 0$. These are the only unconditional bounds; they are weaker than Theorem 7 by exactly the factor Assumption 8 buys.

E.9 Assumption 8: Rationale, Validity, and Open Problems

The hard floor. Assumption 8 is the count-space form of a recognized difficulty: the stability of a data-dependent step-up threshold under simultaneous one-coordinate perturbations. Because $\hat{t}_q$ is a global functional of all scores, a single flip (which moves every score by $\le \delta$) can move $\hat{t}_q$ and thereby re-cross the entire near-threshold layer at once; the disturbances from distinct flips are funnelled through the shared cutoff and concentrate on the same thin set. This is the obstruction that leads [5] to a leave-one-out construction and [12] to an explicit conditional calibration. The mFDR route does *not* remove it; it *isolates* it: the per-null accounting of Lemma 12 clears away the reverse-martingale's independence requirement, the conditional symmetry index, and influence-operator locality, and what remains, alone, is the cutoff-stability bound, which we state as Assumption 8. The gain of Theorem 7 over the unconditional Theorem 8 is *exactly* what it buys, and is real iff it holds. It is the only assumption masking would remove outright; the exact route (Proposition 5: folded, or LOO, Rem. 10) removes it by making every flip act through masking, at the masking cost in power Rule 2 is designed to avoid.



(A6 holds) **(A6 fails: lurch)**

Figure 8: **The threshold-stability floor (Assumption 8).** The cutoff $\hat{t}_q$ is where the empirical FDR ratio $r(t)$ crosses τ. *Left:* a transversal crossing on a steep step: a single flip perturbs r by $O(\delta)$ and moves the crossing by $\leq C_0\delta$, so the slack stays $O(\delta)$ (Theorem 7). *Right:* a flat stretch grazing τ (positive-probability on the discrete score lattice) : an arbitrarily small nudge tips the whole plateau across, and the crossing lurches by a full grid step $\underline{s} \gg \delta$, re-crossing every near-cutoff null at once. A6 assumes the left case; without it only the grid-scale fallback (Theorem 8) holds. This is the count-space form of the hard floor (App. E.9); the folded construction (Prop. 5) avoids it entirely by handling the cutoff as a stopping time in the reverse-martingale.

Why the model structure favors it? The relaxation is mild for two reasons we can make precise, and fragile in one regime. (i) *The two global functions are rigid.* $(\hat{\bar{\alpha}}, \hat{b})$ are each pooled over all N counts, so a single flip moves them by $O(1/N)$, a leave-one-out/influence statement, not a hope; pooling over N points is what makes them stable. (ii) *Separation steepens the crossing.* The flat null and the small-k-concentrated Beta–Binomial alternative are well-separated, so attainable scores have log-LR increments bounded away from zero (Lemma 8), with no bunching, and the cutoff sits in the informative region where the alternative density is steep. Both steepen the empirical staircase $r(t)$ and disfavor the flat crossings that cause a lurch (Figure 8). Equivalently, Fisher information (high where signal and null separate) is what makes estimates rigid against a single perturbation. (iii) *Where it is fragile.* The folded likelihood pays an information tax (eq. (137): $I^{\mathrm{f}}/I^{\mathrm{u}} \approx 0.44$ at $m = 6$, $\to 0$ at $m = 1$), so small-m hypotheses carry little information and are least rigid; the condition is most secure for large m and strong separation, most fragile at $m = 1$ or weak signal.

The unconditional floor, and what remains open. Two of the section's three guarantees hold with no appeal to Assumption 8 at all: the folded rule is exactly valid unconditionally (Proposition 5), and the plug-in keeps the grid-scale bound $\mathrm{mFDR} \leq \tau + O(\underline{s})$ without it (Theorem 8). The assumption is therefore not load-bearing for validity; it is the single input beyond Assumptions 4–7 that *sharpens* this floor to the slack rate $O(\delta)$, and since the mFDR proof has already reduced every other ingredient away, it is the only hypothesis left to discharge. The averaged setting is moreover forgiving: because mFDR is an expectation, the uniform-over-flips form is suffices to bound the lurch *frequency*, replacing $O(\delta)$ by $O(\delta) + \Pr(\text{lurch}) \cdot \underline{s}$, the natural target for a probabilistic strengthening. **We remain explicit about the gap this leaves**: this sure form does not hold at every configuration. It holds when $r(t)$ crosses τ transversally, **the regime that the Beta–Binomial separation and the pooled-scalar rigidity make typical and most secure at large m and strong signal** (Figure 8); it can fail where r grazes τ on a flat stretch, which has positive probability on the lattice. The cost of a failure is bounded and explicit: the cutoff then lurches by a grid step $\underline{s}$ in place of $O(\delta)$, and the guarantee falls back to the $O(\underline{s})$ of Theorem 8. Recovering the sharper averaged rate from that fallback is the open step, and bounding $\Pr(\text{lurch})$ to do so routes through the same density and locality analysis rather than escaping it. We therefore present Theorem 7 as a conditional sharpening on a single named hypothesis, with the unconditional Theorem 8 and the exact folded guarantee (Proposition 5) as the floor that holds regardless.

Remark 6 (Reading the three guarantees together)**.** *Proposition 5, Theorem 7, and Theorem 6 form a ladder of decreasing strength of conclusion bought with increasing strength of method or hypothesis. The folded rule gives* unconditional, exact *FDR control but pays the information tax* (137)*; it is the rock. The plug-in mFDR theorem gives $\tau + O(\delta)$ control of the* averaged *ratio under Assumption 8; it is the high-power deployment with the weaker error notion. The present theorem upgrades that to the*

per-realization *FDR under the additional ranking-stability Assumption 9; it is the strongest statement available for the unmodified plug-in, and it is exactly as conditional as its two siblings. We present all three rather than collapse them: the folded guarantee is what we can prove with no transversality hypothesis, and the plug-in guarantees are what we can prove, in averaged and per-realization form, when the flagged stability conditions hold, which the Beta–Binomial geometry makes typical but not automatic.*

E.10 Consistency and evidence

Remark 7 (Non-spatial limit recovers $O(1/N)$)**.** *When $\alpha(\mathrm{loc}) \equiv \bar{\alpha}$ the spatial channel vanishes and only the two pooled scalars remain, so $\delta = L_b|\Delta\hat{b}| = O(1/N)$ and Theorem 7 reads* $\mathrm{mFDR} \le \tau + O(1/N)$*, matching the paper's independently-derived non-spatial guarantee. A purely-global $O(1)$ slack would contradict this limit; the resolution is that the $O(1)$ absolute asymmetry of a fixed-dimensional scalar perturbation becomes an $O(1/N)$* rate *once divided by $\mathbb{E}[R] = \Theta(N)$.*

Remark 8 (Consistency with the exact-null FDR statement)**.** *Under exact Assumption 4, the $\tau + O(\delta)$ rate matches the mirror-exact regime; under the mirror-conservative weakening of [5] ($P(k_i \le j) \le P(k_i \ge m_i - j) + O(1/m_i)$), the deviation propagates to an additional $O(1/m_{\min})$, zero under exact Assumption 4. In our experiments the exact-uniform null is met by construction.*

Remark 9 (Empirical scope and caveats)**.** *The simulations bear on the numerator-vs-rate point of Remark 7, not on Assumption 8. In the surface-off regime (scalars fit on raw counts), the data-reuse slack* $\mathrm{mFDR}_{\mathrm{raw}} - \mathrm{mFDR}_{\mathrm{oracle}}$ *fluctuated about* 0 *within Monte-Carlo error across $N = 150$–1200, with raw, folded, and oracle agreeing to three decimals; the product slack $\times N$ was too noisy to separate $O(1/N)$ from a small constant, so it is consistent with, not independent confirmation of, the analytics. This regime is exactly where the spatial channel and most lurch risk are absent, so it is silent on Assumption 8; a direct probe would measure the realized cutoff move $|\hat{t}_q^{(i)} - \hat{t}_q|$ under single flips in the* spatial *regime and its scaling against δ vs. $\underline{s}$, which we leave to future work.*

Remark 10 (Exact control by flip-invariant scoring)**.** *Beyond the folded rule, the leave-one-out estimate $\hat{\alpha}^{(-i)}(\mathrm{loc}_i)$ (computed without i) is also exactly orientation-invariant and gives* $\mathrm{FDR} \le \tau$ *via Step 3, at the cost of information (it reverts isolated hypotheses to $\bar{\alpha}$) and of N separate fits. We deploy the plug-in with its characterized slack as the default, and report the folded rule as the unconditional guarantee.*

E.11 Proof of Theorem 3 (Optimality of the spatial lfdr score)

We restate the theorem for convenience.

Theorem (Optimality of the spatial lfdr score; restated)**.** *Assume the compositional model of Section 3 holds with true null-probability function $\alpha^*(\mathrm{loc})$, and that $(\hat{\alpha}, \hat{b}, a) = (\alpha^*, b^*, a^*)$ are correctly specified. For a level τ, let $\Pi_T(\tau)$ be the expected number of true discoveries of the sublevel set of T_i with mFDR τ, and $\Pi_{\mathrm{marg}}(\tau)$ that of any rule measurable with respect to the counts $\{(k_i, m_i)\}_{i=1}^N$ alone, in particular any rule ranking by* $\mathrm{lfdr}_{\mathrm{marg}}$*, with mFDR $\le \tau$. Then $\Pi_T(\tau) \ge \Pi_{\mathrm{marg}}(\tau)$ for every attainable τ, with strict inequality at every τ at which, with positive probability, the location-free rule rejects a hypothesis while retaining one with a smaller true posterior null probability.*

Proof. The argument compares power functions of the mFDR level and never requires two procedures to realize the same level. Step 1 identifies the optimal power frontier $\Pi^*(\cdot)$ and shows it is traced by sublevel sets of the true posterior. Step 2 shows that under correct specification the sublevel sets of T_i are exactly these sets. Step 3 shows that any location-free rule lies on or below the frontier. Step 4 makes the gap strict by an exchange argument.

Step 1: the optimal power frontier and the mFDR identity. For a level $\ell \in (0, 1)$, let

$$\Pi^*(\ell) \;=\; \sup\left\{ \mathbb{E}\Big[\sum_i \theta_i\, \delta_i\Big] \;:\; \delta \text{ data-measurable},\ \mathrm{mFDR}(\delta) \le \ell \right\}$$

be the largest expected number of true discoveries attainable at mFDR $\le \ell$, where $\theta_i = 1$ for the alternative. By [21], the supremum is attained by rejecting a sublevel set $\{i : \mathrm{lfdr}_i^* \le t\}$ of the true posterior null probability $\mathrm{lfdr}_i^* = \Pr(\theta_i = 0 \mid \text{all data})$; under the model the triples $(\mathrm{loc}_i, \theta_i, k_i)$ are

independent across i, so $\mathrm{lfdr}_i^* = \Pr(\theta_i = 0 \mid \mathrm{loc}_i, k_i, m_i)$. Because $\mathbb{E}[1-\theta_i \mid \text{data}] = \mathrm{lfdr}_i^*$, the mFDR of any such sublevel set equals the average posterior inside it,

$$\mathrm{mFDR}\big(\{\mathrm{lfdr}_i^* \le t\}\big) \;=\; \frac{\mathbb{E}\sum_i \mathrm{lfdr}_i^*\,\mathbb{I}[\mathrm{lfdr}_i^* \le t]}{\mathbb{E}\sum_i \mathbb{I}[\mathrm{lfdr}_i^* \le t]} \;=:\; \rho(t). \tag{146}$$

Thus a cutoff t realizes level $\ell = \rho(t)$, and as t sweeps $(0,1)$ the sublevel sets trace the frontier Π^*.

Step 2: sublevel sets of T_i attain the frontier. Under correct specification,

$$T_i \;=\; \widehat{\mathrm{lfdr}}_{\mathrm{spatial}}(\hat\alpha(\mathrm{loc}_i), k_i, m_i) \;=\; \frac{\hat\alpha(\mathrm{loc}_i)\,P_0(k_i,m_i)}{\hat\alpha(\mathrm{loc}_i)\,P_0(k_i,m_i) + (1-\hat\alpha(\mathrm{loc}_i))\,P_1(k_i,m_i;\hat b)} \tag{147}$$

is Bayes' theorem read off the mixture with parameters $(\hat\alpha, \hat b, a)$; correct specification sets them to the truth, so $T_i = \mathrm{lfdr}_i^*$. (The dependence on $\hat b$ through $P_1(\cdot;\hat b)$ is why correct specification must bundle all three parameters; see Remark (b).) Hence every region $\{i : T_i \le t\}$ is a posterior sublevel set and attains Π^* at its realized level $\rho(t)$:

$$\Pi_T(\tau) \;=\; \Pi^*(\tau) \qquad \text{for every attainable } \tau. \tag{148}$$

Step 3: a location-free rule lies on or below the frontier. Let δ^{marg} be measurable with respect to the counts alone, for instance the running-average rule applied in order of $\mathrm{lfdr}_{\mathrm{marg}}$. It is one data-measurable rule with mFDR $\le \tau$, so by the definition of Π^* and (148),

$$\Pi_{\mathrm{marg}}(\tau) \;\le\; \Pi^*(\tau) \;=\; \Pi_T(\tau). \tag{149}$$

The loss has a concrete source. Two hypotheses with identical counts (k,m) but $\alpha^*(\mathrm{loc}_i) < \alpha^*(\mathrm{loc}_j)$ satisfy $T_i < T_j$, since the lfdr is strictly increasing in its prior argument (the map $\alpha \mapsto \alpha P_0/(\alpha P_0 + (1-\alpha)P_1)$ has derivative $P_0P_1/(\cdot)^2 > 0$); yet a location-free rule assigns them the same score and cannot prefer i over j. More generally, it cannot trade a hypothesis with slightly stronger counts in a null-rich region against one with slightly weaker counts in a signal-rich region.

Step 4: strictness by exchange. Suppose that, with positive probability at level τ, δ^{marg} rejects some i while retaining some j with $\mathrm{lfdr}_j^* < \mathrm{lfdr}_i^*$. Form the modified rule that rejects j in place of i on this event. The number of rejections is unchanged, so the denominator of (146) is fixed; the numerator changes by $\mathrm{lfdr}_j^* - \mathrm{lfdr}_i^* < 0$, so the mFDR decreases and the modified rule remains feasible at τ. The expected number of true discoveries changes by $(1-\mathrm{lfdr}_j^*) - (1-\mathrm{lfdr}_i^*) = \mathrm{lfdr}_i^* - \mathrm{lfdr}_j^* > 0$. Hence $\Pi_{\mathrm{marg}}(\tau) < \Pi^*(\tau) = \Pi_T(\tau)$. Such inversions require α^* to be non-constant: if $\alpha^* \equiv \bar\alpha$, then $\mathrm{lfdr}_i^* = \mathrm{lfdr}_{\mathrm{marg}}(k_i, m_i)$ is a function of the counts alone. □

The deployed rules. Both decision rules rank by T_i: Rule 2 rejects $\{i : T_i \le \hat t_q\}$, and Rule 1 rejects a prefix of the T-ordering. Neither is a location-free rule, so the theorem separates their shared ranking from location-free ranking, not the two rules from each other. They differ only in their data-dependent cutoffs. For Rule 1, the identity in the proof of Theorem 1(b) gives $\mathbb{E}[V] = \mathbb{E}\big[\sum_{i\in\mathcal{R}_1} \mathrm{lfdr}_{\mathrm{marg}}(k_i,m_i)\big] + \sum_i \mathrm{Cov}\big(\mathbb{I}\{i\in\mathcal{R}_1\}, \mathrm{lfdr}_i^* - \mathrm{lfdr}_{\mathrm{marg}}(k_i,m_i)\big)$, with the first term at most $\tau\,\mathbb{E}|\mathcal{R}_1|$; whenever the covariance sum is strictly negative, as for an informative ordering, Rule 1's realized mFDR is strictly below τ. Rule 2's realized level is set by the Barber–Candès estimate and is not characterized here.

Remarks. *(a)* The theorem compares scores at a common mFDR level. Exact level matching need not be achievable at finite N, since the attainable mFDR values are discrete; the statement is then a $\le$-dominance of the power functions. It does *not* compare the deployed Rule 1 and Rule 2 at their operating points.

(b) If $\hat\alpha$ or $\hat b$ is misspecified, T_i is no longer the true posterior and Step 2 fails.

E.12 Discreteness of the threshold (Remark 11)

Remark 11 (Discreteness affects power, not validity)**.** *The score T_i depends on k_i only through $(P_{0,i}, P_{1,i}(k_i,m_i))$, so T_i takes values on a lattice of size at most m_i+1. The candidate threshold set*

$\{T_i\} \cup \{\tilde{T}_i\}$ has size at most $\sum_{i=1}^{N} 2(m_i+1)$, and $\hat{t}_q$ in (7) is forced to land on one of these values. This costs power when $m_{\min}$ is small, since the procedure cannot fine-tune the threshold between adjacent lattice values. However, FDR validity is unaffected: the mirror symmetry $k_i \overset{d}{=} m_i - k_i$ holds exactly *on the lattice under $H_{0,i}$ (Step 1 of the proof of Theorem 2), and the BC argument requires no continuity in T_i. Validity is governed by N (through the influence-function slack); finite m_i is purely a power consideration.*

F ADBench benchmark: full setup

Data	10 ADBench datasets, $N = 1{,}000$ points subsampled per dataset; locations are the feature vectors
Labels	clusters define null and alternative regions; unclustered points are background nulls; 20% of labels flipped at random
Generator: latent p-values	nulls $p_i^* \sim U(0,1)$; alternatives $p_i^* \sim \mathrm{Beta}(0.3, 3)$
Generator: latent field	smoothness $\nu \in \{d/2 + k : k = 1, \ldots, 5\}$
Fit: alternative	$\mathrm{Beta}(a, \hat{b})$, a held fixed at each of 8 values in $[0.1, 0.8]$; $\hat{b}$ estimated from the counts (Section 3.1)
Fit: kernel	Matérn of fixed smoothness; λ by cross-validation
Null budget	$m \in \{5, 10, \ldots, 100\}$
Target	$\tau = 0.1$
Oracle	RegFDR [18] on the true $\{p_i^*\}$
Baselines	RegFDR and Benjamini–Hochberg on point-estimated p-values

G LLM-as-judge benchmark: full setup

This appendix details the AlpacaEval 2.0 evaluation referenced in Section 6.1 of the main text. Our objective is to demonstrate the count-level model in a regime with *no spatial structure*, isolating the contribution of the finite-m machinery from any benefit due to spatial regularization.

G.1 Benchmark and judge

We use the AlpacaEval 2.0 benchmark which consists of $N = 805$ instruction-following prompts paired with model-generated responses. For each prompt the AlpacaEval length-controlled judge produces a binary preference between a candidate model and a fixed baseline (the GPT-4 response shipped with the benchmark). The prompts span a broad distribution of instruction types: writing, summarization, reasoning, coding, factual recall, and conversational tasks. We make no modifications to the benchmark, the judge, or the evaluation rubric.

G.2 Challenger pool and group split

We evaluate 12 challenger LLMs (open-weights and API models spanning 2024–2025 releases) against the same fixed baseline. To prevent data leakage between the test statistic and the ground-truth label, the 12 challengers are partitioned into two disjoint groups of size $m = 6$:

- **Group A** (test statistic) supplies the count k_i: the number of Group A challengers whose response on prompt i fails to defeat the baseline (i.e., the judge prefers the baseline). By construction $k_i \in \{0, 1, \ldots, 6\}$ and $m_i = 6$ uniformly for all prompts.
- **Group B** (ground truth) provides the ground-truth label for prompt i: the prompt is labeled an *alternative* (genuinely distinguishes model quality) if the majority of Group B challengers defeat the baseline, and *null* otherwise.

The split is performed once at random; we do not search across splits.

G.3 Hypothesis structure

For each prompt i, the null hypothesis is:

$H_{0,i}$: the prompt does not systematically distinguish model quality, so any preference over the baseline is random.

Under $H_{0,i}$, the latent probability p_i^* that a Group-A challenger fails to defeat the baseline is $1/2$ in expectation (a random preference), so $k_i \mid m_i, H_{0,i}$ has the discrete-uniform distribution on $\{0, 1, \ldots, 6\}$ implied by the Beta–Binomial$(m, 1, 1)$ model of the main text.

For framework configuration, we use the identity kernel ($K = I$), which sets all spatial coupling to zero. Each $\hat{\alpha}_i$ is then estimated from the count $(k_i, m_i = 6)$ alone, with the global prior $\hat{\bar{\alpha}}$ providing the only source of pooling.

G.4 Baseline

The natural baseline of Benjamini–Hochberg on standard discrete p-values $p_i = (k_i + 1)/(m_i + 1) \in \{1/7, 2/7, \ldots, 7/7\}$ *cannot* produce any rejection at $m = 6$: the smallest attainable p-value is $1/7 \approx 0.143$, which exceeds any conventional FDR target ($\tau \in \{0.05, 0.1\}$). To give BH a fair chance, we compare instead against an uncalibrated heuristic:

$$\tilde{p}_i \;\coloneqq\; 1 - \frac{\mathrm{wins}_i}{m},$$

where $\mathrm{wins}_i = m - k_i$ is the number of Group-A challengers defeating the baseline on prompt i. This $\tilde{p}_i$ takes values on the same lattice $\{0, 1/6, 2/6, \ldots, 1\}$ but is not a calibrated p-value (it lacks the $+1$ Laplace smoothing); we use it as the most charitable BH baseline available at $m = 6$.

G.5 Configuration

- Target FDR: $\tau = 0.1$.
- Beta-shape parameter: $a = 0.8$, held fixed (Section 3.1).
- Global null fraction $\hat{\bar{\alpha}}$: Storey's estimator of Section 3.1 on the marginal counts; we obtain $\hat{\bar{\alpha}} \approx 0.79$.
- Alternative-shape parameter b: maximum likelihood given $\hat{\bar{\alpha}}$ (Section 3.1); we obtain $\hat{b} \approx 1.6$.
- No spatial structure: $K = I$, λ irrelevant.
- Both Rule 1 (spatially ordered marginal lfdr) and Rule 2 (count-space mirror) are applied to the same data.
- The β-mixing and adaptive-allocation policy of Section 5.2 are not used here (this is a non-allocation experiment).

G.6 Evaluation metrics

- **Discoveries:** the number of prompts rejected.
- **FDR:** the empirical false-discovery proportion, $\mathrm{FDP} = V/(R \vee 1)$, where V is the number of rejections labeled null by Group B's majority vote and R is the total number of rejections.
- **Power:** the empirical true-positive rate, $T/(T + F)$, where T is the number of rejections labeled alternative by Group B and F is the total number of Group-B-labeled alternatives.

We report point estimates without confidence intervals because the benchmark is run once on the full $N = 805$ prompts; there is no stochastic component to average over once the group split is fixed.

G.7 Why no spatial structure here

The AlpacaEval prompts do not have a meaningful spatial geometry: any embedding (e.g., sentence-encoder representations of the prompts) would impose an artificial structure. We therefore evaluate the framework as a non-spatial procedure here, which is also why Rule 1 and Rule 2 are expected to produce nearly identical outputs — with $K = I$, the spatial score carries no information beyond each hypothesis's own counts, so both rules rank by the same count-based order and differ only in their cutoffs.

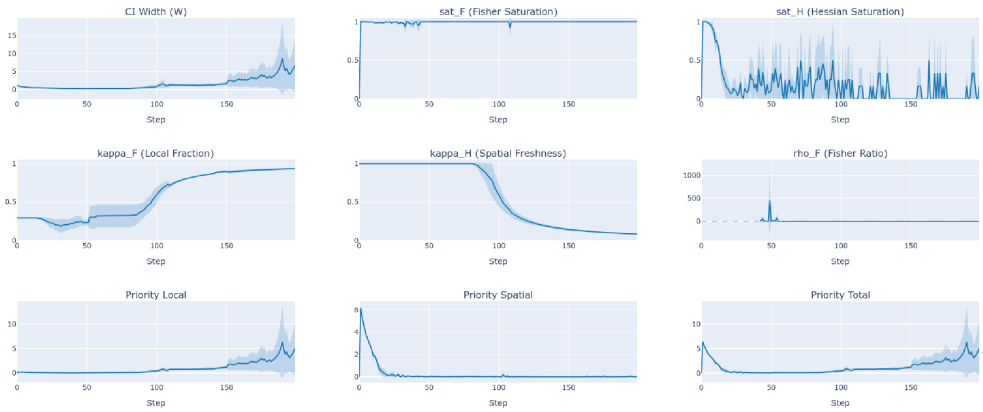


Figure 9: **Priority Score Components Over Sampling Steps.** Mean priority score components across hypotheses and datasets and random seeds (shaded regions: ± 1 std). Top row: confidence interval width (W), Fisher saturation factor (sat_F), and Hessian saturation factor (sat_H). Middle row: local freshness ($\kappa_F = 1 - w_i$), spatial freshness (κ_H), and Fisher saturation ratio (ρ_F). Bottom row: decomposition of total priority S_i into local component ($\beta \cdot \text{sat}_F \cdot \kappa_F \cdot W^2$) and spatial component ($(1-\beta) \cdot \text{sat}_H \cdot \kappa_H \cdot \sum_j K_{ij}^2 W_j^2$). As sampling progresses, saturation factors decrease from 1 toward 0, reflecting diminishing marginal information gain per sample.

G.8 Reproducibility notes

The benchmark queries are public (AlpacaEval 2.0 official release). Challenger model identities, the random Group A/Group B split, and the per-prompt judge outputs are released alongside the paper's code repository, allowing exact replication of Table 1.

H Additional Evaluations Figures

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state the contributions (a count-space likelihood under finite null draws, an RKHS-structured prior, per-hypothesis uncertainty and an allocation policy) together with their scope: Theorem 1 holds under condition (PC), Theorem 2 up to a characterized slack under stated stability assumptions, and the allocation policy of Section 5.2 is presented as a policy without a formal guarantee (see also the Limitations paragraph).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

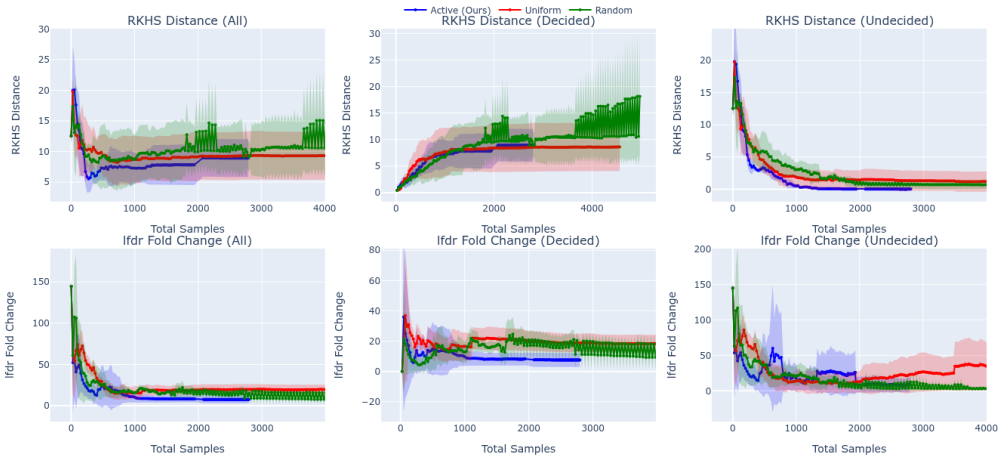


Figure 10: **Convergence to the RegFDR oracle [18].** RKHS distance $\|\alpha - \alpha_{\text{oracle}}\|_K$ (top) and mean lfdr fold change $\langle |\text{lfdr}_i - \text{lfdr}_i^*| / \text{lfdr}_i^* \rangle$ (bottom), evaluated on all hypotheses (left), decided only (center), and undecided only (right). For *undecided* hypotheses, while all three strategies converge to the oracle (right column), the active allocation converges to better α faster, and mostly for the lfdr as well (notice active decides faster, hence per step, remain with more difficult hypotheses). For *decided* hypotheses (center column), the active allocation does not seem to gain advantage for early steps (although the variance is big), but finally converge to better results. **Active allocation reaches comparable** undecided convergence with fewer total samples than Uniform, while Random wastes budget on already-decided hypotheses, producing higher variance and slower overall convergence.

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: A dedicated Limitations paragraph in the Conclusions covers the uniform-null and mirror-symmetry assumptions, the variance-floor regime, and the failure of fixed-m guarantees under adaptive allocation; App. C.14.5 tabulates which guarantees survive adaptive allocation, and App. E states the stability assumptions Rule 2 requires and that they can fail.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Every theorem states its assumptions (the standing assumptions of App. A and the Rule 2 assumptions of App. E) and is proved in the appendices, with proof sketches in the main text for Theorems 1–3.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 6, App. F (ADBench setup) and App. G (AlpacaEval configuration) give the data construction, generator and fitted parameters, budgets, baselines and metrics, and Section 6.1 gives the HIGGS setup; the estimators and the allocation algorithm are fully specified (Section 3, Algorithm 1, App. C.14).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code will be released upon publication. All datasets (ADBench, AlpacaEval 2.0, the HIGGS dataset) are public, and the procedures are fully specified in Sections 3, 5 and App. C.14.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: App. F (ADBench: data construction, generator $\mathrm{Beta}(0.3, 3)$, fitted-a grid, ν grid, budgets $m \in \{5, \ldots, 100\}$, τ), App. G (AlpacaEval: $a = 0.8$, τ, λ, $\hat{\tilde{\alpha}}$, $\hat{b}$, metrics), Section 6.1 (HIGGS: subsample, binning, $m \approx 20$, τ) and App. C.14 (allocation hyperparameters, β tuning, micro/macro schedule).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Figure 2 shows 0.95 quantile envelopes over the 10 datasets and the (a, ν) sweep; Figure 3 aggregates 4 seeds $\times$ 10 datasets; Table 2 gives Wilson 95% intervals for the FDP given the reported counts; Table 1 reports point estimates for one fixed group split, as explained in App. G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments ran on one workstation with a single NVIDIA RTX 3090 GPU; the full ADBench sweep takes a few hours, and the AlpacaEval and HIGGS experiments minutes. No large-scale compute was used.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics: public benchmark data only, no human subjects, and no models or data with misuse potential.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This is foundational statistical methodology for multiple testing; the Societal Impact section states that we see no specific societal consequences beyond more reliable large-scale discovery, and there is no direct path to negative applications.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: No data or models with a high risk of misuse are released.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All assets are cited: ADBench [14], AlpacaEval 2.0 [19] and the HIGGS dataset of Chen et al. [10], all publicly released research benchmarks used under their published licenses.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: No new datasets or models are introduced; the code will be released upon publication with documentation.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No crowdsourcing or research with human subjects; the AlpacaEval judge outcomes are an existing public benchmark.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a component of the method. The AlpacaEval 2.0 LLM-as-judge benchmark is used only as evaluation data (App. G).

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.