CONTROLLING FALSE DISCOVERY IN ARBITRARILY STRUCTURED HYPOTHESIS SPACES VIA REPRODUCING KERNELS.

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale hypothesis testing is central to science, where controlling the False Discovery Rate (FDR) is the standard approach to managing false positives across many simultaneous tests. Hypotheses rarely exist in isolation, they often exhibit structure through proximity, connectivity, or hierarchy. This structure represents both a challenge and an opportunity: while classical methods treat this structure as an obstacle requiring conservative correction, leveraging it can substantially increase power. Here, we reframe structured FDR control as a regularized learning problem. By optimizing within a suitable Reproducing Kernel Hilbert Space (RKHS), we introduce a framework that unifies continuous domains, graphs, or hierarchies under a single algorithm through kernel choice alone. This formulation enables smooth solutions in place of the piecewise-constant fits of prior methods, principled likelihood-based hyperparameter selection, and inference at unobserved locations which in turn supports sample-efficient experimental design. Building on this estimator, we provide two decision rules whose guarantees do not require the spatial model to be correct: one controls the FDR unless the ordering it induces is systematically inverted, and the other controls the marginal FDR up to a slack set by the stability of the fit, with an exact masked variant. We validate our method on both synthetic and real-world benchmarks.

1 INTRODUCTION

The challenge of multiple hypothesis testing is one of the most fundamental in science, addressing our ability to control the probability of false discoveries that arise purely from random chance when evaluating a large number of hypotheses. In practice, this often translates to controlling the False Discovery Rate (FDR). In most domains, it is more the rule than the exception to encounter multiple hypothesis testing problems where hypotheses are not independent, forming a structured hypothesis space: nearby brain regions activate together, genes in the same pathway co-regulate, and spatially proximate cells share molecular signatures. In practice, this structure significantly affects FDR results (Schwartzman and Lin, 2011). We frame our work under the **structured** hypothesis space setup, which can be seen as a variant of the *Covariate-Dependent Two-Group Models*, i.e., with geometric or topological organization induced by **relations** between hypotheses (any measurable association, from spatial proximity to hierarchical nesting, that indicates hypotheses "belong together" in some sense). This differs from approaches that require explicit **dependencies**, i.e., formal probabilistic relationships specified through joint or conditional distributions (e.g., positive regression dependency, correlation structures), which need not be present for relations or structure to be meaningful. Our framework also somewhat resemble the *regularization-based* approaches which also operates on relational information with the leading example, SmoothFDR (Tansey et al., 2016), encodes relations through graph connectivity and uses Total Variation regularization to enforce piecewise-constant significance patterns over connected components. This relaxes the probabilistic requirement but introduces new limitations: the graph structure must be specified a priori, solutions are constrained to piecewise-constant functions poorly suited to smoothly varying phenomena, and the framework lacks principled mechanisms for hyperparameter selection or interpolation to unobserved locations.

In this paper, we address the general problem by assuming a continuous hypothesis space defined over a d-dimensional domain, subsequently demonstrating that almost any discrete problem can be

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

approximated under this framework. By modeling the spatially varying structure within a Reproducing Kernel Hilbert Space (RKHS) (Aronszajn, 1950), diverse structures, from spatial grids to phylogenetic trees to protein interaction networks, can be incorporated through kernel choice alone. Moreover, we move beyond the restrictive piecewise-constant assumptions of current L_1-based methods, enabling flexible smoothness regularization and continuous maps of the prior null probability, estimated at the observed locations and interpolated, with uncertainty, to unobserved ones. Our framework is computationally efficient, scaling as $O(N^2)$ via natural gradient optimization, and distinguishes itself by supporting valid likelihood-based cross-validation.

The remainder of this paper is organized as follows. Section 2 provides preliminary definitions and positions the approach within the literature. Section 3 establishes the problem formulation and introduces the spatially-varying mixture model. Section 4 details our point-wise optimization framework in RKHS. Section 5 translates the resulting estimates into two decision rules with formal FDR guarantees and characterizes when the spatial ranking is optimal. Section 6 extends inference to the entire domain and addresses optimal experimental design for untested locations. Section A discusses practical implementation, including likelihood-based hyperparameter selection and the application of our framework to discrete structures via graph kernels. Finally, Section 7 evaluates the method on semi-synthetic benchmarks and on two real-world tasks: HIGGS particle physics and TCGA gene expression on the STRING protein-protein interaction network.

2 Preliminaries, Related Works, and Positioning

p-value is the probability under the null of obtaining a test statistic at least as extreme as the one observed. **False Discovery Rate (FDR)** is the expected proportion of false positives in the multiple hypotheses scenario: $\mathrm{FDR} = \mathbb{E}\left[\frac{V}{R} \,\middle|\, R > 0\right] P(R > 0)$ where V is the number of false positives and R is the total number of rejections. FDR control literature distinguishes between global and **local FDR** ($lfdr$): $lfdr$ is a Bayesian approach providing per-hypothesis significance by estimating the posterior null probability (Efron, 2004) using a two-group mixture with marginal density: $f(z) = \alpha_0 f_0(z) + (1 - \alpha_0) f_1(z)$. Here, α_0 and $(1 - \alpha_0)$ are the prior probabilities of the null and alternative hypotheses, respectively, $f_0(z)$ is the null distribution (theoretically uniform for p-values, though often empirically estimated), and $f_1(z)$ is the alternative distribution (typically unknown but assumed to be skewed toward zero). Given this mixture model, the local-fdr for test statistic z is defined as the posterior probability: $\mathrm{lfdr}(z) = P(H_0|Z = z) = \frac{\alpha_0 f_0(z)}{\alpha_0 f_0(z) + (1-\alpha_0) f_1(z)}$.

2.1 Related Works and Positioning

Controlling the FDR for related hypotheses is a long-standing challenge. Classical procedures, such as the Benjamini-Hochberg (B-H) method (Benjamini and Hochberg, 2018), control the FDR under independence or positive regression dependency, while arbitrary dependence are treated as a nuisance requires more conservative methods (Benjamini and Yekutieli, 2001). Recent work leverages relations for power gains, yielding structured FDR methods that fall into several categories: **Explicit Dependency Modeling** models dependencies as proper statistical relations (e.g., joint or conditional distributions), and have shown that this can lead to optimal *"oracle"* procedures that significantly boost power (Cai et al., 2011). However, those methods (Cai et al., 2011; Heller and Rosset, 2020) typically require: (1) the dependencies must represent actual probabilistic relationships with dependency graph given explicitly; this lacks mechanisms for handling continuous domains, (2) these dependencies must be sparse as noted by (Heller and Rosset, 2020). **Adaptive P-value Methods.** Methods such as LAWS (Cai et al., 2021) and STRAW (Wang et al., 2023) addresses spatial dependencies through adaptive p-value weighting, constructing local weights from spatial neighbors using discrete windows or local weighted averaging to up-weight or down-weight p-values before applying standard FDR correction. Importantly, these approaches lacking an explicit model of the spatial structure with no clear way to enforce prior beliefs or a principled mechanism for hyperparameter selection. **Covariate-Dependent Methods.** AdaFDR (Zhang et al., 2019) and NeuralFDR (Xia et al., 2017) learn a discovery threshold as a function of the features associated with each hypothesis. FDR regression (Scott et al., 2014), Black Box FDR (Tansey et al., 2018), AdaPT (Lei and Fithian, 2018), and CAMT (Zhang and Chen, 2020) let the null prior, and in AdaPT and CAMT also the alternative density, depend on a covariate vector and reject by a local-fdr statistic. AdaPT controls FDR in finite samples by fitting on masked p-values; CAMT proves asymptotic FDR control under misspecification of the working model and

weak dependence. **Regularization-Based Approaches** focuses on regularization-based approaches that treat structure as a smoothness constraint, with the *SmoothFDR* framework (Tansey et al., 2016) represents as leading example. SmoothFDR utilize an Empirical Bayes two-group mixture model where the prior null is constrained by a Total Variation (TV) penalty over a graph structure, enforcing piecewise-constant solutions that effectively cluster significant hypotheses into discrete spatial blocks. The optimization is performed via an Expectation-Maximization (EM) algorithm, where the M-step requires solving a non-differentiable graph-weighted fused lasso problem, necessitates the use of the Alternating Direction Method of Multipliers (ADMM), involving repeated solution of linear systems, scaling with the number of edges E and imposes substantial computational costs.

3 Problem Setup

3.1 Problem Formulation and the Spatially Varying Mixture Model

Let the Hypothesis Index Space (HIS) be a continuous domain $\mathcal{H} \subset \mathbb{R}^D$. Our goal is to estimate a spatially varying local False Discovery Rate (lfdr) over $\mathcal{H}$. Let $\mathcal{P}$ denote the set of candidate data-generating distributions. For each location $\text{loc} \in \mathcal{H}$, we define a null hypothesis $H_{\text{loc}} \subset \mathcal{P}$. For any true distribution $P \in \mathcal{P}$, the set of true nulls is the subset of locations where the null hypothesis holds (assumed measurable): $\mathcal{H}_0(P) := \{\text{loc} \in \mathcal{H} \mid P \in H_{\text{loc}}\}$. We assume our observed data, consisting of p-values $\{p_i\}_{i=1}^N$ at discrete spatial coordinates $\{\text{loc}_i\}_{i=1}^N \subset \mathcal{H}$, represent a finite sample from an underlying continuous p-value process, denoted $(p_{\text{loc}}(X))_{\text{loc}\in\mathcal{H}}$. Following Perone Pacifico et al. (2004) and Blanchard et al. (2014), this **conceptual** process must satisfy the *joint measurability* condition, which states that the mapping $(\omega, \text{loc}) \mapsto p_{\text{loc}}(X(\omega))$ is jointly measurable. We treat the triples $(\text{loc}_i, \theta_i, p_i)$, where $\theta_i = 0$ indicates that H_{loc_i} holds, as independent draws from the model in Eq. equation 1 below. The relations between hypotheses are carried by the prior $\alpha(\text{loc})$; they do not enter the joint law of the p-values. This is the sampling assumption used by both decision rules (Appendices C and F). The structure is modeled implicitly through the choice of a reproducing kernel and smoothness regularization. Following (Efron, 2004), we model the p-value process by the conditional PDF where $\alpha(\text{loc})$ is the *spatially varying prior probability* of H_{loc}:

$$f(p \mid \text{loc}) = \alpha(\text{loc}) f_0(p) + (1 - \alpha(\text{loc})) f_1(p), \tag{1}$$

Because $\alpha(\text{loc})$ represents a mixing probability, a critical constraint is that $\alpha(\text{loc}) \in [0, 1]$ for all $\text{loc} \in \mathcal{H}$. Here, $f_0(p)$ is the PDF for p-values drawn from a location where the null hypothesis is true. While validity assumptions typically require $f_0(p)$ to be uniform, we follow (Efron, 2004) in allowing for more flexible, empirically estimated null distributions. Conversely, $f_1(p)$ is the PDF for p-values drawn from locations where **alternative hypotheses** are true.

3.2 Estimating the Component Densities.

A key advantage of the formulation is the separation between spatial structure and component densities (Proposition 3.1). Spatial dependence enters only through the mixing proportion $\alpha(\text{loc})$, while the null and alternative densities $f_0(z), f_1(z)$ remain spatially invariant. Consequently, the marginal distribution of test statistics, pooled across all locations, follows the classical lFDR. Hence, we can estimate f_0 and f_1 using any methods for non-spatial lFDR (e.g central matching or empirical null fitting (Efron, 2004)), treating spatial coordinates as irrelevant to the marginal estimation.

Proposition 3.1 (Marginal Density Independence)**.** *Under the spatially-varying mixture model (Eq. 1), the marginal density of test statistics follows the standard two-group mixture:*

$$f(z) = \bar{\alpha} f_0(z) + (1 - \bar{\alpha}) f_1(z) \tag{2}$$

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

Proof in Supplementary Section B; the estimation procedure used in Section 7 is given in Supplementary Section H.

4 Point-wise Solution

We frame our problem as a Tikhonov-regularized maximum likelihood problem within a Reproducing Kernel Hilbert Space (RKHS), $\mathcal{H}_K$. Let $\|\cdot\|_{\mathcal{H}_K}$ denote the norm induced by the positive definite

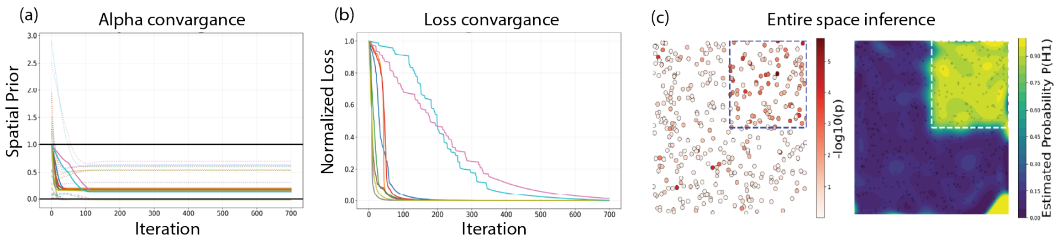

Figure 1: (a) Alpha convergence to [0,1] across all datasets- broken lines indicates min/max values, full line the mean. (b) Loss convergence. (c) Learning on entire space- Left: observed p-values. Right: α over the entire space.

kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Our goal is to estimate $\alpha(\text{loc}_i)$ at the observed locations $\{\text{loc}_i\}_{i=1}^N$ where we have test statistics $\{z_i\}_{i=1}^N$ (or equivalently, p-values $\{p_i\}_{i=1}^N$). Notice Section 6 will address the case of estimating on the entire domain. We regularize toward the global prior $\bar{\alpha}$ rather than toward zero, so hypotheses with weak spatial information default to the marginal estimate and the $\lambda_{\text{reg}} \to \infty$ limit recovers the classical non-spatial two-group mixture. Let $\mathbf{c}_{\bar{\alpha}}$ denote the precomputed coefficient vector satisfying $K\mathbf{c}_{\bar{\alpha}} = \bar{\alpha}\,\mathbf{1}$, obtained once via conjugate gradient before optimization begins; the centering adds no per-iteration cost and preserves the kernel-cancellation property established below. Our objective is then to minimize the negative log-likelihood:

$$\min_{\alpha \in \mathcal{H}_K} \quad \mathcal{J}(\alpha) = -\sum_{i=1}^{N} \log\big(\alpha(\text{loc}_i) f_0(p_i) + (1 - \alpha(\text{loc}_i)) f_1(p_i)\big) + \lambda_{\text{reg}} \|\alpha - \bar{\alpha}\|^2_{\mathcal{H}_K} \tag{3}$$

with $\lambda_{\text{reg}} > 0$ controlling the regularization strength. Since the empirical risk term depends on α solely through its evaluations at the finite set of points $\{\text{loc}_i\}_{i=1}^N$, the Generalized Representer Theorem (Schölkopf et al., 2001) guarantees that the minimizer lies in the finite-dimensional subspace spanned by the kernel sections centered at the data: $\alpha(\cdot) = \sum_{i=1}^N c_i K(\cdot, \text{loc}_i)$. This reduces the variational problem to optimization over the coefficient vector $\mathbf{c} \in \mathbb{R}^N$. The core challenge in this formulation is restricting $\alpha(\text{loc}_i) \in [0, 1]$ at all observed locations while maintaining a tractable convex optimization problem. We address this challenge here under the point-wise setting, with Section 6 extending the approach to the entire domain. Rather than imposing hard box constraints $\alpha \in [0, 1]$, which would necessitate constrained optimization methods (see below), we introduce a *soft boundary penalty* that penalizes violations per unit interval:

$$\Lambda_{\text{bound}}(\alpha) = \sum_{i=1}^{N} \left[\max(0, \alpha_i - 1)^2 + \max(0, -\alpha_i)^2\right] \tag{4}$$

This formulation maintains convexity (each term is a composition of the convex squared hinge loss with linear function values), is differentiable almost everywhere enabling efficient gradient-based optimization, and is inactive when constraints are satisfied ($\alpha_i \in [0, 1]$), ensuring it does not interfere with well-behaved solutions. Alternative approaches include hard projection methods or squashing functions (e.g., logistic transformations); however, as detailed in Supplementary Section K, we found the penalty-based formulation to be both elegant and sufficient in practice. Figure 1(a) validates that α remains well-bounded within $[0, 1]$ across all datasets evaluated in Section 7, with 1(b) shows this happens before convergence. Our complete point-wise objective then becomes:

$$\min_{\alpha \in \mathcal{H}_K} \mathcal{L}(\alpha) = -\sum_{i=1}^{N} \log h_i(\alpha) + \lambda_{\text{reg}} \|\alpha - \bar{\alpha}\|^2_{\mathcal{H}_K} + \lambda_{\text{bound}} \Lambda_{\text{bound}}(\alpha) \tag{5}$$

where $h_i(\alpha) = \alpha(\text{loc}_i) f_0(p_i) + (1 - \alpha(\text{loc}_i)) f_1(p_i)$.

Optimization via Natural Gradient Descent While convex in $\mathbf{c} \in \mathbb{R}^N$, standard optimizers operating in Euclidean geometry struggle with RKHS objectives: for smooth kernels the Gram matrix

K has rapidly decaying eigenvalues, creating ill-conditioned optimization landscapes with narrow valleys. In our experiments, even sophisticated adaptive methods fail to overcome this. A more principled approach is the *natural gradient* (Munkres, 2018), which accounts for the geometry of the parameter space, and proved beneficial in similar setups (Kozdoba et al., 2024). For RKHS optimization, the natural gradient is defined with respect to the metric tensor K, $\tilde{\nabla}_\alpha \mathcal{L} = K^{-1}\nabla_{\mathbf{c}}\mathcal{L}$. Naively, this appears to trade one ill-conditioned operation for another, since K^{-1} inherits the conditioning of K. Fortunately, the structure of our objective leads to an exact cancellation:

Proposition 4.1 (Kernel Cancellation in Natural Gradient)**.** *For the point-wise objective in Equation 5, the natural gradient admits the simple form:*

$$\tilde{\nabla}_\alpha \mathcal{L} = \mathbf{w} + 2\lambda_{reg}\left(\mathbf{c} - \mathbf{c}_{\bar{\alpha}}\right) + \lambda_{bound}\nabla_\alpha \Lambda_{bound}, \quad w_i = -\frac{f_0(p_i) - f_1(p_i)}{\alpha_i f_0(p_i) + (1-\alpha_i) f_1(p_i)} \tag{6}$$

With $[\nabla_\alpha \Lambda_{bound}]_i = 2(\alpha_i - 1)$ *if* $\alpha_i > 1$, $2\alpha_i$ *if* $\alpha_i < 0$, *and* 0 *otherwise. And* $\mathbf{c}_{\bar{\alpha}}$ *precomputed once.*

Proof in Supplementary Section C. The update rule depends only on residuals $\mathbf{w}$ and current parameters $\mathbf{c}$, eliminating K from the gradient step entirely. The kernel matrix is required only for the forward evaluation $\boldsymbol{\alpha} = K\mathbf{c}$, which is numerically stable and yields significantly faster convergence.

5 Decision Rules and FDR Control

Here we address the translation of $\hat{\alpha}(\mathrm{loc})$ into formal rejection rules. The validity of FDR guarantees is intrinsically predicated on having calibrated p-values, reflecting a different philosophical stance than the one outlined in the introduction, where p-values may be erroneous and side information is a source of evidence no less important. Nonetheless, formal guarantees under the calibration assumption remain important. We present two complementary rules, neither requires the structure to be correct: Rule 1 controls FDR provided the ordering it induces is not systematically inverted (condition (PC)), and Rule 2 controls mFDR up to a slack set by the stability of the fit, with an exact masked variant. Under correct specification, ranking by the spatial lfdr is optimal (Theorem G.1).

Rule 1 uses the spatial model to order the hypotheses with the error accounting been marginal. The ordering may depend on the data in any way, but it must not be anti-informative: an ordering that front-loads a null-rich region defeats the budget (Remark C.3), which is what (PC) excludes. Estimate $\hat{\alpha}(\mathrm{loc}_i)$ via Stage 1 (Section 4), and separately estimate $f_0, f_1, \bar{\alpha}$ from the marginal two-group mixture (Proposition 3.1). For every hypothesis compute the spatial and the marginal local fdr,

$$\mathrm{lfdr}_{\mathrm{spatial}}(p_i, \mathrm{loc}_i) = \frac{\hat{\alpha}(\mathrm{loc}_i)\, f_0(p_i)}{\hat{\alpha}(\mathrm{loc}_i)\, f_0(p_i) + (1 - \hat{\alpha}(\mathrm{loc}_i))\, f_1(p_i)}, \tag{7}$$

$$\mathrm{lfdr}_{\mathrm{marg}}(p_i) = \frac{\bar{\alpha}\, f_0(p_i)}{\bar{\alpha}\, f_0(p_i) + (1 - \bar{\alpha})\, f_1(p_i)}. \tag{8}$$

Order all N hypotheses by increasing $\mathrm{lfdr}_{\mathrm{spatial}}$, giving $\pi(1), \ldots, \pi(N)$, and reject the longest prefix whose running average of the marginal lfdr stays within the level:

$$R_1 = \{\pi(1), \ldots, \pi(r^*)\}, \qquad r^* = \max\left\{r : \tfrac{1}{r}\textstyle\sum_{j \le r} \mathrm{lfdr}_{\mathrm{marg}}(p_{\pi(j)}) \le q\right\}, \tag{9}$$

with $R_1 = \emptyset$ if no r qualifies. This is the running-average rule of (Sun and Cai, 2007), applied along the spatial ordering instead of the marginal one. Because the ordering is not by $\mathrm{lfdr}_{\mathrm{marg}}$, the running average need not be monotone; r^* is the largest qualifying prefix, not the first crossing. Rule 1 never rejects more hypotheses than the marginal running-average rule at the same level; its gain is in which hypotheses it rejects (Remark C.5). Each rejected hypothesis adds to the expected number of false rejections its posterior null probability given all the data, $\mathrm{lfdr}_i^* = P(\theta_i = 0 \mid \mathcal{D})$, $\mathcal{D} = \{(\mathrm{loc}_j, p_j)\}_{j=1}^N$, whereas Rule 1 budgets with $\mathrm{lfdr}_{\mathrm{marg}}$. Hence $\mathrm{FDR}(R_1) \le q$ whenever, on the rejected set, the marginal lfdr does not underestimate lfdr^* on average:

$$\mathbb{E}\Big[\tfrac{1}{|R_1| \vee 1}\textstyle\sum_{i \in R_1}\left(\mathrm{lfdr}_i^* - \mathrm{lfdr}_{\mathrm{marg}}(p_i)\right)\Big] \le 0. \tag{PC}$$

An ordering that ignores the locations satisfies (PC) with equality; an ordering that front-loads hypotheses whose locations lower their posterior null probability makes it negative; it fails only if the ordering systematically prefers hypotheses that are more likely null than their p-values alone indicate. Theorem C.1 states the guarantee, with $\mathrm{lfdr}_{\mathrm{marg}}$ at the marginal parameters; it requires neither a correct $\hat{\alpha}$ nor sample splitting, only (PC).

The Spatial FDR Control via Mirror Statistics decision rule leverage the mirror-statistic framework (Barber and Candès, 2015). Define the *rejection score* and *mirror score*:

$$T_i = \frac{\hat{\alpha}(\text{loc}_i)\, f_0(p_i)}{\hat{\alpha}(\text{loc}_i)\, f_0(p_i) + (1-\hat{\alpha}(\text{loc}_i))\, f_1(p_i)}, \tilde{T}_i = \frac{\hat{\alpha}(\text{loc}_i)\, f_0(1-p_i)}{\hat{\alpha}(\text{loc}_i)\, f_0(1-p_i) + (1-\hat{\alpha}(\text{loc}_i))\, f_1(1-p_i)} \tag{10}$$

The mirror score replaces p_i with its complement $1-p_i$. The procedure computes T_i and $\tilde{T}_i$ for all hypotheses, then sets the threshold via the Barber–Candès step-up rule:

$$\hat{t} = \max\left\{ t : \frac{1 + \#\{i : \tilde{T}_i \le t\}}{1 \vee \#\{i : T_i \le t\}} \le q \right\}, \quad \mathcal{R} = \{i : T_i \le \hat{t}\}. \tag{11}$$

Rule 2 rests on the symmetry of null p-values under $p_i \mapsto 1-p_i$. A null p-value is as likely to fall above $1/2$ as below it, so a null is as likely to fall below the cutoff through its mirror score $\tilde{T}_i$ as through its own score T_i, and the mirror count estimates the number of false rejections. Exact control requires that everything the rule uses to rank the hypotheses be blind to the side of $1/2$ on which each null fell. Fitting $\hat{\alpha}$, $\bar{\alpha}$ and f_1 on the masked values $\min(p_i, 1-p_i)$ achieves this and the knockoff+ argument of Barber and Candès (2015) gives $\text{FDR} \le q$ for every N (Proposition F.7). However, masking discards information, hence in this work the deployed rule uses a variant of Barber and Candès (2015) which leverages the tractability of the RKHS formulation. Fits $\hat{\alpha}$ on the p-values themselves and controls the mFDR up to a slack: the proof compares the data with a copy in which one null p-value is flipped. The copy has the same distribution, so on average a null is counted through its own score as often as through its mirror score, and the step-up rule keeps the mirror count below q times the number of rejections. The error comes from the flip changing the fit: it nudges every other score by at most (Lemma F.3), and it can move the cutoff. A nudge can change a decision only for nulls whose scores lie within the nudge, plus the cutoff's own move, of the cutoff. Theorem F.5 bounds the mFDR by q plus half the expected number of nulls in this band, divided by $\mathbb{E}|R|$. The bound uses no stability assumption and is informative when the band is thin. When a single flip moves the cutoff by $O(\delta)$, scores do not pile up at it, and $\mathbb{E}|R|$ grows linearly in N, this gives $\text{mFDR} \le q + O(1/\lambda_{\text{reg}}) + O(1/N)$ (Corollary F.6) at a fixed λ_{reg}, as selected by cross-validation in our experiments. Appendix F gives the full statements, assumptions, and proofs.

Theorem G.1 compares ranking in general, giving light to the differences between the two rules. It shows that under correct specification a sublevel set of $\text{lfdr}_{\text{spatial}}$ attains, at every mFDR level, at least as many true discoveries in expectation as any rule that uses the p-values alone, and strictly more whenever such a rule would, with positive probability, reject a hypothesis while retaining one with a smaller posterior null probability. Both rules reject sublevel sets of $\text{lfdr}_{\text{spatial}}$ with data-dependent cutoffs; which of the two stops closer to its level is an empirical question (Section 8).

6 Solution Over the Entire Domain

While the point-wise approach yields reliable estimates at observed coordinates, extending inference to the entire continuous domain $\mathcal{H}$ requires ensuring that $\alpha(\text{loc}) \in [0,1]$ holds globally. This introduces a significant challenge: enforcing the bounds $\forall \text{loc} \in \mathcal{H}$ generates an infinite number of constraints. Formally, this is known as a *Semi-Infinite Programming (SIP)* problem (Hettich and Kortanek, 1993; López and Still, 2007). In Supplementary Section K, we provide a detailed analysis of classical SIP solvers (such as exchange methods and barrier functions) and demonstrate why they are ill-suited. Here we favor a robust two-stage approach that separates the problem into two stages: **The first stage** is the described point-wise estimation which obtain estimates $\{\hat{\alpha}_i\}_{i=1}^N$ at the observed locations. **The second stage** treats these point-wise estimates as target labels and learns a globally valid function $\alpha(\text{loc}) \in [0,1]$ that closely approximates them across the entire domain. This is fundamentally a problem of *learning a probability-valued function*: from an information-theoretic perspective, we seek a function $\alpha(\text{loc})$ that best approximates the target distribution $\hat{\alpha}(\text{loc})$ over the spatial domain. Conceptualizing the true spatial function $\alpha^*(\text{loc})$ as existing over the entire domain $\mathcal{H}$, with the point-wise estimates $\{\hat{\alpha}_i\}$ as noisy observations at discrete locations, we seek to minimize the expected KL divergence over the spatial domain plus a smoothness regularization:

$$\min_{\alpha \in \mathcal{H}_K} \int_{\mathcal{H}} D_{\text{KL}}(\alpha^*(\text{loc}) \| \alpha(\text{loc}))\, p(\text{loc})\, d\text{loc} + \lambda_{\text{global}} \|\alpha\|^2_{\mathcal{H}_K} \tag{12}$$

where $p(\text{loc})$ measures over the domain and $\lambda_{\text{global}} > 0$ controls smoothness. Since each $\alpha(\text{loc})$ represents a Bernoulli parameter (the probability of the null hypothesis): $D_{\text{KL}}(\hat{\alpha}\|\alpha) = \hat{\alpha}\log\frac{\hat{\alpha}}{\alpha} + (1-\hat{\alpha})\log\frac{1-\hat{\alpha}}{1-\alpha}$. For the data fidelity term, minimizing KL is equivalent to minimizing the cross-entropy, since $\hat{\alpha}$ entropy is constant with respect to α. Hence, we can approximate the integral using our finite observations $\{\hat{\alpha}_i, \text{loc}_i\}_{i=1}^N$, treating $p(\text{loc})$ as an empirical measure concentrated at the data points, yielding the *logistic loss*:

$$\mathcal{L}_{\text{CE}}(\alpha) = -\sum_{i=1}^{N} [\hat{\alpha}_i \log \alpha(\text{loc}_i) + (1-\hat{\alpha}_i)\log(1-\alpha(\text{loc}_i))] \tag{13}$$

Crucially, the logistic loss is a *proper scoring rule* (Gneiting and Raftery, 2007), minimizing expectation when $\alpha = \hat{\alpha}$, ensuring that our learned function faithfully approximates the point-wise estimates without systematic bias. To ensure $\alpha(\text{loc}) \in [0,1]$ everywhere while minimizing this discrepancy, we use a squashing function. The natural choice is the logistic function $\sigma(z) = \frac{1}{1+e^{-z}}$, which maps the entire real line to $[0,1]$. This leads us to parameterize $\alpha(\text{loc}) = \sigma(g(\text{loc}))$ where $g \in \mathcal{H}_K$ is unconstrained. The resulting optimization is then the *kernel logistic regression*:

$$\mathbf{c}^* = \arg\min_{\mathbf{c}\in\mathbb{R}^N} -\sum_{i=1}^{N} [\hat{\alpha}_i \log\sigma(g_i) + (1-\hat{\alpha}_i)\log(1-\sigma(g_i))] + \lambda_{\text{global}}\mathbf{c}^T K\mathbf{c} \tag{14}$$

In practice, we typically set $\lambda_{\text{global}} \leq \lambda_{\text{reg}}$ for closely follow the point-wise estimates while maintaining spatial coherence. Another key practical consideration is how to initialize Stage 2. The coefficients $\mathbf{c}_{\text{point}}$ from Stage 1 provide a natural warm start, but correspond to a different parameterization (direct α vs. squashed g). We use the inverse logistic transformation $\mathbf{c}^{(0)} = K^{-1}\cdot\text{logit}(\hat{\boldsymbol{\alpha}})$ where $\text{logit}(\alpha) = \log(\alpha/(1-\alpha))$ and we clip $\hat{\alpha}_i$ to $[\epsilon, 1-\epsilon]$ for $\epsilon > 0$ (e.g., $\epsilon = 0.01$) to avoid numerical issues. Crucially, the logistic loss formulation yields a convex optimization problem and by applying natural gradients as in Section 4 we obtain a simple update rule :

$$\tilde{\nabla}_{\mathbf{c}}\mathcal{L}_{\text{logistic}} = (\boldsymbol{\sigma} - \hat{\boldsymbol{\alpha}}) + 2\lambda_{\text{global}}\mathbf{c} \tag{15}$$

Figure 1(c) illustrates the alpha inference on a synthetic 2-dimensional dataset. The left panel shows the observed locations with p-values encoded by color, while the right panel presents the inferred alpha function. Notice that while the p-values are relatively noisy, our method was able to correctly infer the regions of high alpha.

Efficient Testing via Optimal Experimental Design. The ability to infer $\alpha(\text{loc})$ at unseen locations enables a data-efficient testing protocol: strategically select locations to test, then leverage spatial structure for decisions at remaining locations. This raises two questions: (1) which locations should we test when designing the experiment? (2) how do we quantify uncertainty at untested locations to determine if predictions are reliable? For Question 1, we select N_0 locations using density-aware A-optimal design (Supplementary Section N), which minimizes average prediction variance across the domain. A parameter γ controls the cluster-isolation tradeoff: from dense regions ($\gamma > 0$) to isolated points ($\gamma < 0$), with $\gamma = 0$ yields uniform weighting. For Question 2, we quantify uncertainty at untested locations via Laplace approximation (Supplementary Section M.3), which combines spatial uncertainty (distance from observed locations) and statistical uncertainty (Fisher information). We evaluate this framework empirically in Section 7.

7 EVALUATIONS

Evaluating the proposed framework requires both known relations between hypotheses and ground-truth significance labels, hence we employ two complementary benchmarks: (i) a **semi-synthetic protocol** over 10 high-dimensional real-world datasets, where the spatial structure is real but p-values are injected with known labels; and (ii) two **real-world tasks** (HIGGS particle physics and TCGA differential expression) where both p-values and ground-truth labels are genuine- but the labels reflects the inherent labeling error of any learning problem. Across all evaluations we report both decision rules. Baselines include Benjamini-Hochberg (BH), a non-spatial two-group model, SmoothFDR (Tansey et al., 2016), and AdaPT (Lei and Fithian, 2018) with three learners (logistic regression, random forest, XGBoost) which is considered the best method in practice.

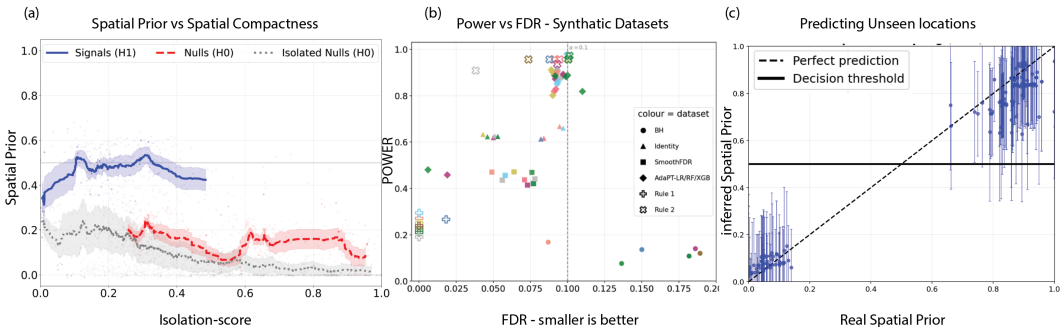

Figure 2: (a) Inferred spatial prior $(1-\alpha)$ versus geometric isolation score across semi-synthetic datasets. (b) Power vs. FDR on the 10 semi-synthetic datasets across all baselines at $\alpha = 0.10$. (c) Predicted spatial prior at held-out locations versus ground truth computed from full dataset. Error lines are the CI for 95% confidence.

Simulation Study - Semi-Synthetic Benchmarks. Here we aim to evaluate the framework's ability to detect a signal under the scenario that motivates this paper: data arising from a genuinely spatial and localized signal, observed through noisy p-values. We utilize 10 high-dimensional real-world datasets onto which we synthetically inject p-values. For each dataset, we first cluster the data to identify stable clusters, then select two clusters mutually distant in kernel space: one serves as the signal-enriched cluster (C_{signal}), the other as the null-enriched cluster (C_{null}). Random null locations with no specific spatial structure are then added. A cluster corruption parameter ($\gamma = 0.1$) randomly flips 10% of the labels, creating outlier nulls and signals. P-values are generated from these labels, with nulls drawn from $\mathcal{U}[0,1]$ and alternatives from Beta$(0.3, 3)$. Our method fits the local FDR using a Matérn kernel (distinct from the one used for clustering), with hyperparameters selected via cross-validation and boundary penalty fixed at $\lambda_{\text{bound}} = 500$. For the baselines, SmoothFDR converts p-values to z-scores via the probit transform and runs the graph fused lasso with a Gaussian two-group model; AdaPT uses 50 threshold iterations with the spatial coordinates as covariates. Figure 2(a) plots $(1-\alpha)$ against geometric isolation, defined as inverted normalized kernel similarity to the k nearest "H_1" neighbors, reflecting distance from signal-enriched regions. Red points belong to the H_1 cluster, blue to H_0, and gray to isolated locations. The method achieves clear separation between labels across isolation scores. While H_0 cluster maintain appropriate prior, they generally trend up because the null is distributed according to $U[0,1]$, so after spatial smoothing the low tail pulls the local mean, an effect less pronounced in isolated locations. Figure 2(b) reports FDR and power at $\alpha = 0.10$ across all datasets. While almost all methods control the FDR, `BH`, `Two-group`, and `SmoothFDR` forfeit substantial power. As the figure show, **Rule 1** is able to gain more power for lower FDR compared with BH, following 5. **Rule 2** dominates the frontier outright: it attains strictly higher power than every baseline on every dataset, with gains over the best AdaPT variant ranging from $+0.030$ (`Hepatitis`) to $+0.346$ (`skin`). While here we treats significance as a property of the location, Supplementary Section P.1 address the case where significance is a property of the distribution that produced the p-value, so the same hypothesis is a true null that happens to sit among signals. This serve as a misspecification analysis of the methods, showing that even for corrupted data the framework present much better results then the standard B-H baseline.

Real-World Datasets. **HIGGS particle physics**: the dataset (Chen et al., 2022) contains 9.8K simulated collision events binned into $N \approx 500$ groups by k-means clustering on kinematic features. For each bin, we compute a one-sided binomial p-value testing whether the rate of events satisfying $m_{wwbb} < 0.8$ exceeds the expected background rate. Ground-truth labels (signal-enriched vs. background-dominated bins) are used only for evaluation. **TCGA differential expression**: we test $N = 3{,}000$ genes between tumor types (BRCA cohort) using DESeq2 p-values from TCGA RNA-seq data. Ground-truth labels are derived from MSigDB Hallmark pathway membership: genes in 22 cancer-relevant pathways are labeled positive; genes in 4 unrelated pathways are labeled negative. FDR and power are evaluated on labeled discoveries. As for implementation details: For HIGGS, we employ a Matérn kernel, encoding nearby bins in kinematic space. For TCGA, we use a diffusion kernel ($\beta = 2$) on the STRING protein-protein interaction network (Szklarczyk et al., 2022) (204,372

edges), widely used in systems biology to capture pathway-level signals (Tripathi et al., 2016); the diffusion parameter and λ_{reg} are automatically tuned via C-V. For AdaPT baselines, covariates are bin centroids in kinematic space for HIGGS, and spectral coordinates of the PPI Laplacian for TCGA (matching the kernel information).

As Table 1 shows, on HIGGS, BH (0.253), SmoothFDR (0.597), all three AdaPT variants (≥ 0.20) and Two-group (0.213) exceed the target on the label-based error rate, while Rule 1 (0.093/0.382) and Rule 2 (0.079/0.343) are the only methods that stay below $q = 0.10$ and recover meaningful power. On TCGA DE, where most baselines stay below the target, the AdaPT variants stall at power ≈ 0.46 and BH/Two-group plateau near 0.65; Rule 1 and Rule 2 reach a 0.15-0.19 power gain over the strongest competitor at comparable FDR.

Table 1: FDR / Power ($\alpha = 0.10$) .

Method	HIGGS	TCGA DE
BH	.253/.564	.065/.643
Two-group	0.213/0.445	0.062/0.651
SmoothFDR	.597/.922	.540/.946
AdaPT-LR	.200/.059	.082/.459
AdaPT-RF	.200/.078	.085/.463
AdaPT-XGB	.223/.426	.075/.461
Rule 1	.093/.382	.061/.801
Rule 2	.079/.343	.063/.838

Inference on Unseen Locations. To validate generalization to untested locations, we performed a holdout evaluation on the semi-synthetic datasets. We selected 70% of locations using the density-aware A-optimal strategy ($\gamma = 0.1$), observed their p-values, and fit the spatial model to predict $\hat{\alpha}(\text{loc})$ at the remaining 30%. Figure 2(c) compares predictions (y-axis) against the true α values computed from the full dataset (x-axis), with 95% posterior CI error bars. The strong diagonal clustering demonstrates accurate spatial interpolation, with 94.8% coverage (close to nominal 95%). These are estimates of the prior at untested locations, used to choose where to test next (Section 6); an FDR-controlled decision at a location still requires observing its p-value.

8 Conclusions and Limitations

The paper introduced a framework for regularized FDR control on arbitrarily structured hypothesis spaces, where structure is encoded through reproducing kernel choice, and instantiated it via two decision rules, *Rule 1* and *Rule 2*. The approach also produce continuous prior maps valid across entire domains with convex optimization guarantees, flexible smoothness regularization and principled hyperparameter selection. Both rules deliver substantial power gains over existing baselines while keeping the error guarantees, which do not require the spatial model to be correct beyond condition (PC) for Rule 1. Moreover, Rule 2 allows for a variant that leverage the framework tractability to allow for a decision rule with significantly more power in the cost of a slack in the fdr guarantee.

Several limitations merit discussion. **First**, our theoretical guarantees apply to the finite set of observed locations; extending decisions to unobserved points relies on the continuous-domain procedure of Section 6 and is subject to the standard caveats of out-of-sample inference. **Second**, every hypothesis contributes only a *single p-value* from its marginal two-group mixture, creating an irreducible information floor that is most pronounced at isolated locations with few neighbors. This limitation applies to any method of this kind; an advantage of our formulation is that the resulting uncertainty can itself be quantified (Section 6). **Third**, the framework relies on kernel operations whose cost becomes a bottleneck for very large hypothesis sets; in Supplementary Section Q we demonstrate that the approach handles up to $500,000$ hypotheses on a single 24 GB GPU. Despite these limitations, we believe the substantial power gains demonstrated in our experiments, combined with the theoretical guarantees and computational efficiency of our framework, make it a valuable addition to the spatial FDR toolkit when structural assumptions are appropriate. Finally, we believe there is great potential in further exploring the relation between RKHS tractability and other FDR control frameworks, particularly covariate-dependent methods, where this relation might lead to deeper uncertainty quantification in FDR correction with covariates, addressing the covariates uncertainty directly.

AI USE STATEMENT

In this work, we used generative AI tools for proving the mathematical claims and to assist in writing the proofs; to implement the proposed method, the baselines, and preprocesses the HIGGS and TCGA data; and to provide feedback on the research methodology and experiments in a pre-submission review of the manuscript. We have used generative AI tools in a a very limited scope to refine the theoretical framework and the mathematical formulation and to propose or refine hypotheses. Additionally, we used generative AI tools to draft parts of the paper, to identify relevant literature, and to edit the paper for readability. We have reviewed all AI-assisted work: the authors checked every proof step by step, tested the AI-written code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. ISSN 1088-6850. doi: 10.1090/s0002-9947-1950-0051437-7. URL http://dx.doi.org/10.1090/S0002-9947-1950-0051437-7.

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.

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, 12 2018. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4), Aug. 2001. ISSN 0090-5364. doi: 10.1214/aos/1013699998. URL http://dx.doi.org/10.1214/aos/1013699998.

G. Blanchard, S. Delattre, and E. Roquain. Testing over a continuum of null hypotheses with false discovery rate control. *Bernoulli*, 20(1), Feb. 2014. ISSN 1350-7265. doi: 10.3150/12-bej488. URL http://dx.doi.org/10.3150/12-BEJ488.

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

T. T. Cai, H. Li, J. Maris, and J. Xie. Optimal false discovery rate control for dependent data. *Statistics and Its Interface*, 4(4):417–430, 2011. ISSN 1938-7997. doi: 10.4310/sii.2011.v4.n4.a1. URL http://dx.doi.org/10.4310/SII.2011.v4.n4.a1.

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.

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.

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.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, Mar. 2007. ISSN 1537-274X. doi: 10.1198/016214506000001437. URL http://dx.doi.org/10.1198/016214506000001437.

R. Heller and S. Rosset. Optimal control of false discovery criteria in the two-group model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(1):133–155, Dec. 2020. ISSN 1467-9868. doi: 10.1111/rssb.12403. URL http://dx.doi.org/10.1111/rssb.12403.

D. Henrion, J. B. Lasserre, and J. Lofberg. Gloptipoly 3: moments, optimization and semidefinite programming, 2007. URL https://arxiv.org/abs/0709.2559.

R. Hettich and K. O. Kortanek. Semi-infinite programming: Theory, methods, and applications. *SIAM Review*, 35(3):380–429, Sept. 1993. ISSN 1095-7200. doi: 10.1137/1035089. URL http://dx.doi.org/10.1137/1035089.

N. Ignatiadis and W. Huber. Covariate powered cross-weighted multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(4):720–751, Aug. 2021. ISSN 1467-9868. doi: 10.1111/rssb.12411. URL http://dx.doi.org/10.1111/rssb.12411.

N. Ignatiadis, B. Klaus, J. B. Zaugg, and W. Huber. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, 13(7):577–580, May 2016. ISSN 1548-7105. doi: 10.1038/nmeth.3885. URL http://dx.doi.org/10.1038/nmeth.3885.

M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences, 2018. URL https://arxiv.org/abs/1807.02582.

R. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. *Proceedings of the Nineteenth International Conference on Machine Learning*, 11, 04 2002.

M. Kozdoba, B. Perets, and S. Mannor. Sobolev space regularised pre density models. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 25494–25533. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/kozdoba24a.html.

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

A. Li and R. F. Barber. Accumulation tests for fdr control in ordered hypothesis testing, 2016a. URL https://arxiv.org/abs/1505.07352.

A. Li and R. F. Barber. Multiple testing with the structure adaptive benjamini-hochberg algorithm, 2016b. URL https://arxiv.org/abs/1606.07926.

M. López and G. Still. Semi-infinite programming. *European Journal of Operational Research*, 180(2):491–518, July 2007. ISSN 0377-2217. doi: 10.1016/j.ejor.2006.08.045. URL http://dx.doi.org/10.1016/j.ejor.2006.08.045.

J. R. Munkres. *Analysis on Manifolds*. CRC Press, Feb. 2018. ISBN 9780429494147. doi: 10.1201/9780429494147. URL http://dx.doi.org/10.1201/9780429494147.

M. Perone Pacifico, C. Genovese, I. Verdinelli, and L. Wasserman. False discovery control for random fields. *Journal of the American Statistical Association*, 99(468):1002–1014, Dec. 2004. ISSN 1537-274X. doi: 10.1198/0162145000001655. URL http://dx.doi.org/10.1198/0162145000001655.

J. O. Royset, E. Polak, and A. Kiureghian. Adaptive approximations and exact penalization for the solution of generalized semi-infinite min-max problems. *SIAM Journal on Optimization*, 14(1):1–34, Jan. 2003. ISSN 1095-7189. doi: 10.1137/s1052623402406777. URL http://dx.doi.org/10.1137/S1052623402406777.

R. Sarkar. *Low Distortion Delaunay Embedding of Trees in Hyperbolic Plane*, page 355–366. Springer Berlin Heidelberg, 2012. ISBN 9783642258787. doi: 10.1007/978-3-642-25878-7_34. URL http://dx.doi.org/10.1007/978-3-642-25878-7_34.

B. Schölkopf, R. Herbrich, and A. J. Smola. *A Generalized Representer Theorem*, page 416–426. Springer Berlin Heidelberg, 2001. ISBN 9783540445814. doi: 10.1007/3-540-44581-1_27. URL http://dx.doi.org/10.1007/3-540-44581-1_27.

A. Schwartzman and X. Lin. The effect of correlation in false discovery rate estimation. *Biometrika*, 98(1):199–214, Feb. 2011. ISSN 1464-3510. doi: 10.1093/biomet/asq075. URL http://dx.doi.org/10.1093/biomet/asq075.

J. G. Scott, R. C. Kelly, M. A. Smith, P. Zhou, and R. E. Kass. False discovery rate regression: an application to neural synchrony detection in primary visual cortex, 2014. URL https://arxiv.org/abs/1307.3495.

A. J. Smola and R. Kondor. *Kernels and Regularization on Graphs*, page 144–158. Springer Berlin Heidelberg, 2003. ISBN 9783540451679. doi: 10.1007/978-3-540-45167-9_12. URL http://dx.doi.org/10.1007/978-3-540-45167-9_12.

G. Still. Generalized semi-infinite programming: numerical aspects. *Optimization*, 49(3):223–242, Jan. 2001. ISSN 1029-4945. doi: 10.1080/02331930108844531. URL http://dx.doi.org/10.1080/02331930108844531.

J. D. Storey. The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6), Dec. 2003. ISSN 0090-5364. doi: 10.1214/aos/1074290335. URL http://dx.doi.org/10.1214/aos/1074290335.

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.

D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. Doncheva, S. Pyysalo, P. Bork, L. Jensen, and C. von Mering. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, Nov. 2022. ISSN 1362-4962. doi: 10.1093/nar/gkac1000. URL http://dx.doi.org/10.1093/nar/gkac1000.

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

W. Tansey, Y. Wang, D. M. Blei, and R. Rabadan. Black box fdr, 2018. URL https://arxiv.org/abs/1806.03143.

S. Tripathi, S. Moutari, M. Dehmer, and F. Emmert-Streib. Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules. *BMC Bioinformatics*, 17(1), Mar. 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-0979-8. URL http://dx.doi.org/10.1186/s12859-016-0979-8.

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.

F. Xia, M. J. Zhang, J. Zou, and D. Tse. Neuralfdr: Learning discovery thresholds from hypothesis features, 2017. URL https://arxiv.org/abs/1711.01312.

M. J. Zhang, F. Xia, and J. Zou. Fast and covariate-adaptive method amplifies detection power in large-scale multiple hypothesis testing. *Nature Communications*, 10(1), July 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11247-0. URL http://dx.doi.org/10.1038/s41467-019-11247-0.

X. Zhang and J. Chen. Covariate adaptive false discovery rate control with applications to omics-wide multiple testing, 2020. URL https://arxiv.org/abs/1909.04811.

A Practical Considerations

Hyperparameter Selection. All hyperparameters (regularization strength λ_{reg}, kernel type, length-scale, smoothness, and the shape parameter a of the alternative density) are selected by held-out likelihood (Supplementary Section L). As Remark E.1 notes, $f(z \mid \text{loc}; \theta)$ is a density for every θ, so held-out log-likelihoods are comparable across hyperparameter values.

Kernel Choice. Kernel choice encodes both the geometry and the smoothness of the function space. For continuous domains we employ Matérn kernels, whose RKHS is norm-equivalent to the Sobolev space of order $\nu + d/2$ (Kanagawa et al., 2018; Kozdoba et al., 2024), so its functions are continuous for every $\nu > 0$ and k times differentiable for integers $k < \nu$; ν sets the smoothness of the prior map, not its validity. For discrete topologies the framework extends seamlessly through graph kernels: diffusion kernels (Kondor and Lafferty, 2002; Smola and Kondor, 2003; Tripathi et al., 2016) replace the L_1 Total Variation penalty of prior methods with an L_2 smoothness penalty $\sum_{(i,j)\in E} W_{ij}(\alpha_i - \alpha_j)^2$ that preserves network modularity (e.g., for gene–gene interactions), and hyperbolic embeddings via Sarkar's theorem (Sarkar, 2012) accommodate hierarchical structures (e.g., gene ontologies) in just two dimensions, where Euclidean kernels would otherwise require dimension exponential in tree depth. Supplementary Section J provides the in-depth treatment for kernels selection.

B Proof of Marginal Density Independence

Proposition (Restated. Marginal Density Independence)**.** *Under the spatially-varying mixture model (Eq. 1), the marginal density of test statistics follows the standard two-group mixture:*

$$f(z) = \bar{\alpha} f_0(z) + (1 - \bar{\alpha}) f_1(z) \tag{16}$$

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

Proof. Consider the observed locations $\{\text{loc}_i\}_{i=1}^N$ as samples drawn from a spatial sampling distribution with density $p(\text{loc})$ over the domain $\mathcal{H}$. The marginal density of a test statistic z, denoted $f(z)$, is obtained by integrating the conditional mixture model (Eq. 1 in main text) over the spatial domain:

$$f(z) = \int_{\mathcal{H}} f(z \mid \text{loc})\, p(\text{loc})\, d\text{loc} \tag{17}$$

$$= \int_{\mathcal{H}} [\alpha(\text{loc}) f_0(z) + (1 - \alpha(\text{loc})) f_1(z)]\, p(\text{loc})\, d\text{loc} \tag{18}$$

Since $f_0(z)$ and $f_1(z)$ do not depend on loc (by our spatially-invariant assumption), we can factor them out of the integral:

$$f(z) = f_0(z) \int_{\mathcal{H}} \alpha(\text{loc})\, p(\text{loc})\, d\text{loc} + f_1(z) \int_{\mathcal{H}} (1 - \alpha(\text{loc}))\, p(\text{loc})\, d\text{loc} \tag{19}$$

Define the global average null probability as:

$$\bar{\alpha} = \mathbb{E}_{\text{loc}}[\alpha(\text{loc})] = \int_{\mathcal{H}} \alpha(\text{loc})\, p(\text{loc})\, d\text{loc} \tag{20}$$

Then the second integral becomes:

$$\int_{\mathcal{H}} (1 - \alpha(\text{loc}))\, p(\text{loc})\, d\text{loc} = \int_{\mathcal{H}} p(\text{loc})\, d\text{loc} - \int_{\mathcal{H}} \alpha(\text{loc})\, p(\text{loc})\, d\text{loc} = 1 - \bar{\alpha} \tag{21}$$

Substituting Equations equation 20 and equation 21 into Equation equation 19:

$$f(z) = \bar{\alpha} f_0(z) + (1 - \bar{\alpha}) f_1(z) \tag{22}$$

This is precisely the standard two-group mixture form used in classical local FDR methods, confirming that the spatial structure affects only the mixing proportion $\bar{\alpha}$, not the functional forms of f_0 and f_1. □

C Proof of FDR Control for Rule 1

Setting. Under the mixture model (Eq. 1) with locations drawn i.i.d. from $p(\mathrm{loc})$ (Proposition 3.1), the triples $(\mathrm{loc}_i, \theta_i, p_i)$ are independent across i, with $P(\theta_i = 0 \mid \mathrm{loc}_i) = \alpha(\mathrm{loc}_i)$ and $p_i \mid \theta_i \sim f_{\theta_i}$. Let $\mathcal{D} = \{(\mathrm{loc}_j, p_j)\}_{j=1}^N$ and $\mathrm{lfdr}_i^* = P(\theta_i = 0 \mid \mathcal{D})$; by independence across i, $\mathrm{lfdr}_i^* = P(\theta_i = 0 \mid \mathrm{loc}_i, p_i)$. Throughout, $\mathrm{lfdr}_{\mathrm{marg}}$ is evaluated at the true $(\bar{\alpha}, f_0, f_1)$.

Theorem C.1 (FDR control for Rule 1)**.** *Let R_1 be the rejection set of Rule 1 (Section 5) under any $\mathcal{D}$-measurable ordering, in particular the ordering by $\mathrm{lfdr}_{\mathrm{spatial}}$ with $\hat{\alpha}$ and its hyperparameters selected on the same data. (a) If (PC) holds, then $\mathrm{FDR}(R_1) \le q$. (b) If $\mathrm{Cov}\big(\mathbf{1}\{i \in R_1\}, \mathrm{lfdr}_i^* - \mathrm{lfdr}_{\mathrm{marg}}(p_i)\big) \le 0$ for every i, then $\mathrm{mFDR}(R_1) = \mathbb{E}|R_1 \cap \mathcal{H}_0| / \mathbb{E}|R_1| \le q$. Neither part requires $\hat{\alpha}$ to be correctly specified, and neither requires sample splitting.*

Proof. Step 1: $\mathrm{lfdr}_{\mathrm{marg}}$ is the posterior given p_i. By Proposition 3.1, $P(\theta_i = 0,\, p_i \in dp) = \bar{\alpha} f_0(p)\, dp$ and p_i has density $\bar{\alpha} f_0 + (1 - \bar{\alpha}) f_1$, so $\mathrm{lfdr}_{\mathrm{marg}}(p_i) = P(\theta_i = 0 \mid p_i)$. Since $\sigma(p_i) \subseteq \sigma(\mathcal{D})$, the tower property gives

$$\mathrm{lfdr}_{\mathrm{marg}}(p_i) = \mathbb{E}\big[\mathrm{lfdr}_i^* \mid p_i\big]. \tag{23}$$

Step 2: conditioning on the data. R_1 is a function of $\mathcal{D}$: the ordering, $\hat{\alpha}$ and its hyperparameters are all computed from it. With $V = |R_1 \cap \mathcal{H}_0|$,

$$\mathbb{E}[V \mid \mathcal{D}] = \textstyle\sum_{i \in R_1} \mathrm{lfdr}_i^*, \tag{24}$$

and no independence between the ordering and the labels is needed. By construction, on every realization,

$$\textstyle\sum_{i \in R_1} \mathrm{lfdr}_{\mathrm{marg}}(p_i) \le q\, |R_1|. \tag{25}$$

Part (a). Since $|R_1|$ is $\mathcal{D}$-measurable, equation 24 gives

$$\mathrm{FDR}(R_1) = \mathbb{E}\Big[\tfrac{\sum_{i \in R_1} \mathrm{lfdr}_{\mathrm{marg}}(p_i)}{|R_1| \vee 1}\Big] + \mathbb{E}\Big[\tfrac{\sum_{i \in R_1} (\mathrm{lfdr}_i^* - \mathrm{lfdr}_{\mathrm{marg}}(p_i))}{|R_1| \vee 1}\Big] \le q,$$

the first term bounded by equation 25 and the second non-positive by (PC).

Part (b). Taking expectations in equation 24, $\mathbb{E}[V] = \mathbb{E}\big[\sum_{i \in R_1} \mathrm{lfdr}_{\mathrm{marg}}(p_i)\big] + \sum_{i=1}^N \mathbb{E}\big[\mathbf{1}\{i \in R_1\}(\mathrm{lfdr}_i^* - \mathrm{lfdr}_{\mathrm{marg}}(p_i))\big]$. By equation 23, $\mathbb{E}[\mathrm{lfdr}_i^* - \mathrm{lfdr}_{\mathrm{marg}}(p_i)] = 0$, so each summand equals the covariance in (b) and is non-positive. With equation 25, $\mathbb{E}[V] \le q\, \mathbb{E}|R_1|$. □

Remark C.2 (When (PC) holds)**.** *Suppose the ordering depends on the data only through $(p_1, \ldots, p_N)$ and independent randomization U, e.g. a uniformly random ordering or the ordering by $\mathrm{lfdr}_{\mathrm{marg}}$. Then R_1 is a function of $(p_1, \ldots, p_N, U)$, and by independence across i, $\mathbb{E}[\mathrm{lfdr}_i^* \mid p_1, \ldots, p_N, U] = \mathbb{E}[\mathrm{lfdr}_i^* \mid p_i] = \mathrm{lfdr}_{\mathrm{marg}}(p_i)$ by equation 23. Conditioning the left-hand side of (PC) on $(p_1, \ldots, p_N, U)$ makes it exactly zero. An ordering that uses the locations moves it away from zero: below zero when it selects hypotheses whose true posterior lies below $\mathrm{lfdr}_{\mathrm{marg}}$, above zero only when it prefers hypotheses more likely null than their p-values alone indicate.*

Remark C.3 (Necessity of (PC))**.** *Let the domain consist of two regions A and B of equal size with $\alpha^* = 1$ on A and $\alpha^* = 1/2$ on B, so that $\bar{\alpha} = 3/4$. Order the hypotheses of A first, by increasing p, and those of B after. For small p, $\mathrm{lfdr}_{\mathrm{marg}}(p) = 3/\{3 + f_1(p)\}$ is small, so the running-average rule rejects the smallest-p hypotheses of A, all of which are null: $\mathrm{FDP} = 1$ on every realization with at least one rejection. The ordering uses locations only, apart from the sort within A, and (PC) fails because $\mathrm{lfdr}_i^* = 1 > \mathrm{lfdr}_{\mathrm{marg}}(p_i)$ on the rejected set. Ordered procedures that count nulls inside the prefix (Barber and Candès, 2015; Li and Barber, 2016a) keep FDR control under any ordering independent of the p-values; the marginal budget does not, which is the price of charging each rejection its marginal posterior instead of counting. An adversary-proof rule with a location-informed ordering is the masked construction of Proposition F.7.*

Remark C.4 (No perturbation slack)**.** $\mathrm{lfdr}_{\mathrm{marg}}(p_i)$ *depends on* p_i *and the global* $(\bar{\alpha}, f_0, f_1)$ *only, not on* $\hat{\alpha}$*. The spatial estimator enters Rule 1 only through the ordering, hence only through (PC). No stability argument is needed, unlike for Rule 2 (Theorem F.5), and the bound holds at every finite* N *with no* $O(\cdot)$ *term.*

Remark C.5 (Source of power: composition, not cardinality)**.** *Let* R_{marg} *be the running-average rule applied along the ordering by* $\mathrm{lfdr}_{\mathrm{marg}}$*, at the same level* q *and with the same* $\mathrm{lfdr}_{\mathrm{marg}}$*. Then* $|R_1| \le |R_{\mathrm{marg}}|$ *on every realization: for every* r*, the* r *smallest values of* $\mathrm{lfdr}_{\mathrm{marg}}$ *have the smallest sum, so if* $\sum_{j\le r} \mathrm{lfdr}_{\mathrm{marg}}(p_{\pi(j)}) \le qr$*, the same holds for the first* r *hypotheses in the marginal ordering, and hence* $r^* \le |R_{\mathrm{marg}}|$*. In particular, Rule 1 makes no rejection whenever the marginal rule makes none. Its gain is therefore in which hypotheses it rejects: along the spatial ordering, a hypothesis with a larger* $\mathrm{lfdr}_{\mathrm{marg}}$ *but a location that lowers its posterior null probability can enter* R_1 *in place of one with a smaller* $\mathrm{lfdr}_{\mathrm{marg}}$ *in a null-rich region. When (PC) holds strictly, part (a) of Theorem C.1 gives* $\mathrm{FDR}(R_1) < q$*, because the marginal accounting overstates the error of an informative prefix.*

D Derivation of Natural Gradient for Point-wise Optimization

D.1 Standard Euclidean Gradient

The objective function (Eq. 5 in main text) decomposes into three terms:

$$\mathcal{L}(\mathbf{c}) = \mathcal{L}_{\text{data}}(\mathbf{c}) + \mathcal{L}_{\text{reg}}(\mathbf{c}) + \mathcal{L}_{\text{bound}}(\mathbf{c}) \tag{26}$$

We compute the gradient of each term separately.

D.1.1 Data Term Gradient

The negative log-likelihood term is:

$$\mathcal{L}_{\text{data}} = -\sum_{i=1}^{N} \log\left[\alpha_i f_0(p_i) + (1-\alpha_i) f_1(p_i)\right] \tag{27}$$

Using the chain rule with $\alpha_i = (K\mathbf{c})_i = \sum_{j=1}^{N} K_{ij} c_j$:

$$\frac{\partial \mathcal{L}_{\text{data}}}{\partial c_j} = -\sum_{i=1}^{N} \frac{1}{\alpha_i f_0(p_i) + (1-\alpha_i) f_1(p_i)} \cdot \frac{\partial}{\partial c_j}\left[\alpha_i f_0(p_i) + (1-\alpha_i) f_1(p_i)\right] \tag{28}$$

$$= -\sum_{i=1}^{N} \frac{f_0(p_i) - f_1(p_i)}{\alpha_i f_0(p_i) + (1-\alpha_i) f_1(p_i)} \cdot \frac{\partial \alpha_i}{\partial c_j} \tag{29}$$

$$= -\sum_{i=1}^{N} \frac{f_0(p_i) - f_1(p_i)}{\alpha_i f_0(p_i) + (1-\alpha_i) f_1(p_i)} \cdot K_{ij} \tag{30}$$

In vector notation, define the residual vector $\mathbf{w} \in \mathbb{R}^N$ with entries:

$$w_i = -\frac{f_0(p_i) - f_1(p_i)}{\alpha_i f_0(p_i) + (1-\alpha_i) f_1(p_i)} \tag{31}$$

Then:

$$\nabla_{\mathbf{c}} \mathcal{L}_{\text{data}} = K\mathbf{w} \tag{32}$$

D.1.2 Regularization Term Gradient (Centered)

The centered RKHS regularization is:

$$\mathcal{L}_{\text{reg}} \;=\; \lambda_{\text{reg}} \|\alpha - \bar{\alpha}\|^2_{\mathcal{H}_K} \;=\; \lambda_{\text{reg}} (\mathbf{c} - \mathbf{c}_{\bar{\alpha}})^T K (\mathbf{c} - \mathbf{c}_{\bar{\alpha}}), \tag{33}$$

where $\mathbf{c}_{\bar{\alpha}} \in \mathbb{R}^N$ is the precomputed anchor coefficient defined by $K\mathbf{c}_{\bar{\alpha}} = \bar{\alpha}\mathbf{1}$ (solved once via conjugate gradient before optimization begins).

Taking the gradient:

$$\nabla_{\mathbf{c}}\mathcal{L}_{\text{reg}} \;=\; 2\lambda_{\text{reg}}K(\mathbf{c} - \mathbf{c}_{\bar{\alpha}}). \tag{34}$$

D.1.3 Boundary Penalty Gradient

The boundary penalty is:

$$\Lambda_{\text{bound}}(\alpha) = \sum_{i=1}^{N} \left[\max(0, \alpha_i - 1)^2 + \max(0, -\alpha_i)^2\right] \tag{35}$$

Define the element-wise gradient:

$$[\nabla_\alpha \Lambda_{\text{bound}}]_i = \frac{\partial}{\partial \alpha_i} \left[\max(0, \alpha_i - 1)^2 + \max(0, -\alpha_i)^2\right] \tag{36}$$

This has the closed form:

$$[\nabla_\alpha \Lambda_{\text{bound}}]_i = \begin{cases} 2(\alpha_i - 1) & \text{if } \alpha_i > 1 \\ 2\alpha_i & \text{if } \alpha_i < 0 \\ 0 & \text{otherwise} \end{cases} \tag{37}$$

Using the chain rule $\frac{\partial \Lambda_{\text{bound}}}{\partial c_j} = \sum_{i=1}^{N} \frac{\partial \Lambda_{\text{bound}}}{\partial \alpha_i} \frac{\partial \alpha_i}{\partial c_j}$:

$$\nabla_{\mathbf{c}}\mathcal{L}_{\text{bound}} = \lambda_{\text{bound}}K\nabla_\alpha \Lambda_{\text{bound}} \tag{38}$$

D.2 Combined Gradient and Factorization

Combining Equations equation 32, equation 34, and equation 38:

$$\begin{aligned} \nabla_{\mathbf{c}}\mathcal{L} &= K\mathbf{w} + 2\lambda_{\text{reg}}K(\mathbf{c} - \mathbf{c}_{\bar{\alpha}}) + \lambda_{\text{bound}}K\nabla_\alpha \Lambda_{\text{bound}} && (39) \\ &= K\left[\mathbf{w} + 2\lambda_{\text{reg}}(\mathbf{c} - \mathbf{c}_{\bar{\alpha}}) + \lambda_{\text{bound}}\nabla_\alpha \Lambda_{\text{bound}}\right] && (40) \end{aligned}$$

This factorization reveals that the Gram matrix K appears as a leading factor, which is the source of the ill-conditioning in standard gradient descent.

D.3 Natural Gradient and Kernel Cancellation

Proof of Lemma 4.1. The natural gradient is defined as:

$$\tilde{\nabla}_\alpha \mathcal{L} = K^{-1}\nabla_{\mathbf{c}}\mathcal{L} \tag{41}$$

Substituting the factored form from Equation equation 40:

$$\begin{aligned} \tilde{\nabla}_\alpha \mathcal{L} &= K^{-1}\left(K\left[\mathbf{w} + 2\lambda_{\text{reg}}(\mathbf{c} - \mathbf{c}_{\bar{\alpha}}) + \lambda_{\text{bound}}\nabla_\alpha \Lambda_{\text{bound}}\right]\right) && (42) \\ &= \left(K^{-1}K\right)\left[\mathbf{w} + 2\lambda_{\text{reg}}(\mathbf{c} - \mathbf{c}_{\bar{\alpha}}) + \lambda_{\text{bound}}\nabla_\alpha \Lambda_{\text{bound}}\right] && (43) \\ &= I\left[\mathbf{w} + 2\lambda_{\text{reg}}(\mathbf{c} - \mathbf{c}_{\bar{\alpha}}) + \lambda_{\text{bound}}\nabla_\alpha \Lambda_{\text{bound}}\right] && (44) \\ &= \mathbf{w} + 2\lambda_{\text{reg}}(\mathbf{c} - \mathbf{c}_{\bar{\alpha}}) + \lambda_{\text{bound}}\nabla_\alpha \Lambda_{\text{bound}} && (45) \end{aligned}$$

where we used $K^{-1}K = I$ (the identity matrix). This completes the proof. □

The complete natural gradient descent update at iteration k is:

Algorithm 1 Natural Gradient Update for Point-wise FDR (centered)

1: **Pre-compute (one time):** $\mathbf{c}_{\bar{\alpha}}$ by solving $K\mathbf{c}_{\bar{\alpha}} = \bar{\alpha}\mathbf{1}$ via CG. ▷ $O(N^2 t_{\text{CG}})$
2: **Input:** Current coefficients $\mathbf{c}^{(k)}$, learning rate η
3: Compute $\boldsymbol{\alpha}^{(k)} = K\mathbf{c}^{(k)}$ ▷ Forward pass: $O(N^2)$
4: Compute residuals: $w_i^{(k)} = -\frac{f_0(p_i) - f_1(p_i)}{\alpha_i^{(k)} f_0(p_i) + (1-\alpha_i^{(k)}) f_1(p_i)}$ ▷ $O(N)$
5: Compute boundary gradients: $[\nabla_\alpha \Lambda_{\text{bound}}]_i$ ▷ $O(N)$
6: Form natural gradient:
7: $\quad \tilde{\nabla}^{(k)} = \mathbf{w}^{(k)} + 2\lambda_{\text{reg}}(\mathbf{c}^{(k)} - \mathbf{c}_{\bar{\alpha}}) + \lambda_{\text{bound}} \nabla_\alpha \Lambda_{\text{bound}}$ ▷ $O(N)$
8: Update: $\mathbf{c}^{(k+1)} = \mathbf{c}^{(k)} - \eta \tilde{\nabla}^{(k)}$ ▷ $O(N)$
9: **Return:** $\mathbf{c}^{(k+1)}$

The total per-iteration complexity is $O(N^2)$, dominated by the kernel matrix-vector product in the forward pass.

E Normalization of the Cross-Validation Criterion

Here we record the normalization property used by the cross-validation criterion of Section A.

Remark E.1 (Normalization)**.** *For every hyperparameter value θ and every location with $\alpha(\text{loc};\theta) \in [0,1]$, $f(z \mid \text{loc};\theta) = \alpha(\text{loc};\theta) f_0(z) + (1 - \alpha(\text{loc};\theta)) f_1(z)$ integrates to one in z, since f_0 and f_1 are densities. Held-out log-likelihoods $\sum_{i \in \text{Test}} \log f(z_i \mid \text{loc}_i;\theta)$ are therefore comparable across θ without normalization corrections.*

Consistency with classical FDR. Because the regularizer is centered at $\bar{\alpha}$, the limit $\lambda_{\text{reg}} \to \infty$ (infinite smoothing) drives $\alpha(\text{loc};\theta) \equiv \bar{\alpha}$. In this limit the model reduces exactly to the classical non-spatial two-group mixture, with the marginal null probability $\bar{\alpha}$ of Supplementary Section H, exactly as in standard local FDR methods (Efron, 2004). (Without centering, the limit would instead drive $\alpha \to 0$, breaking this consistency.)

F FDR Control for Rule 2

Rule 2 compares each hypothesis's score T_i with a mirror score $\tilde{T}_i$, computed from $1 - p_i$ instead of p_i. A null p-value is uniform, so it is equally likely to be p_i or $1 - p_i$: a null is as likely to fall below a cutoff through its own score as through its mirror score. The number of mirror scores below a cutoff therefore estimates the number of false rejections, and the step-up rule equation 11 picks the largest cutoff at which that estimate stays below a fraction q of the rejections. The argument is exact when the fitted surface $\hat{\alpha}$ cannot tell on which side of $1/2$ a null p-value fell. The deployed rule fits $\hat{\alpha}$ on the p-values themselves, so flipping one null p-value, $p_i \mapsto 1 - p_i$, changes the fit slightly and nudges every score. This appendix measures what that costs. After fixing notation and assumptions, four lemmas supply the tools: the bookkeeping of the step-up rule (Lemma F.1), the effect of flipping a null (Lemma F.2), how far a flip moves the fit (Lemma F.3), and which decisions such a move can change (Lemma F.4). The main result, Theorem F.5, bounds the mFDR by q plus the probability that a null lies in a narrow band around the cutoff, with no stability assumption. Corollary F.6 turns this into a rate when the cutoff itself is stable, and Proposition F.7 gives a variant with exact FDR control.

Setting and assumptions

Each hypothesis carries a pair of scores, its own and its mirror's, and the proof tracks the smaller and the larger of the two. It compares the observed data with a copy in which a single null p-value is flipped. Three assumptions are used. (A1) makes the flip harmless in distribution; (A2) ensures that the smaller score of each pair is the one computed from the smaller of p_i and $1 - p_i$; and (A3) collects the regularity conditions used only to bound how far a flip moves the fit.

Setting. Throughout, $f_0 \equiv 1$ (Appendix G), and the score is $g(s,p;\psi) = s/\{s + (1-s)f_1(p;\psi)\}$ for a prior value $s \in [0,1]$, where ψ are the fitted parameters of f_1; $\theta = (\bar{\alpha}, \psi)$ collects the global quantities of the marginal fit. Rule 2 (Eq. equation 11) uses $T_i = g(\hat{\alpha}(\mathrm{loc}_i), p_i; \hat{\psi})$ and $\tilde{T}_i = g(\hat{\alpha}(\mathrm{loc}_i), 1-p_i; \hat{\psi})$, the counts $R(t) = \#\{i : T_i \le t\}$ and $A(t) = \#\{i : \tilde{T}_i \le t\}$, the cutoff $\hat{t} = \max\{t \in \{T_i\} \cup \{\tilde{T}_i\} : (1 + A(t))/(1 \vee R(t)) \le q\}$, with $\hat{t} = -\infty$ if no t qualifies, and the rejection set $\mathcal{R} = \{i : T_i \le \hat{t}\}$. Write $\ell_i = \min(T_i, \tilde{T}_i)$ and $h_i = \max(T_i, \tilde{T}_i)$ for the smaller and larger score of a pair, $B_i = \mathbf{1}\{p_i > 1/2\}$ for the side of $1/2$ on which p_i fell, $\mathcal{H}_0$ for the true nulls, $V = \#\{i \in \mathcal{H}_0 : T_i \le \hat{t}\}$ for the false rejections and $\tilde{V} = \#\{i \in \mathcal{H}_0 : \tilde{T}_i \le \hat{t}\}$ for the nulls counted through their mirror scores. For $i \in \mathcal{H}_0$, $\mathcal{D}^{(i)}$ is the data with p_i replaced by $1 - p_i$, and a superscript (i) marks any quantity recomputed on $\mathcal{D}^{(i)}$, including $\hat{\alpha}$, $\hat{\theta}$ and $\hat{t}$. The regularization λ_{reg} and the kernel are held fixed (Remark F.8).

Assumptions.

- (A1) *Null symmetry.* Given the labels, the p-values are independent, and $p_i \sim U[0,1]$ for $i \in \mathcal{H}_0$.
- (A2) *Side condition.* $f_1(p;\psi) > f_1(1-p;\psi)$ for all $p < 1/2$ and every ψ the fit can return. For the Beta alternative of Appendix G, $f_1(p)/f_1(1-p) = \{p/(1-p)\}^{a-b}$, so (A2) holds whenever $a < b$. As g is decreasing in f_1, (A2) gives $T_i < \tilde{T}_i$ if and only if $p_i < 1/2$, so $B_i = \mathbf{1}\{T_i > \tilde{T}_i\}$.
- (A3) *Regularity* (used only in Lemma F.3). $\kappa^2 = \max_x K(x,x) < \infty$. For every $i \in \mathcal{H}_0$ and every j, all values $\bar{\alpha} + u$ with $\bar{\alpha} \in \{\hat{\bar{\alpha}}, \hat{\bar{\alpha}}^{(i)}\}$ and u between $\hat{f}(\mathrm{loc}_j)$ and $\hat{f}^{(i)}(\mathrm{loc}_j)$ lie in $[\varepsilon, 1-\varepsilon]$. The global estimates are flip-stable, $\|\hat{\theta}^{(i)} - \hat{\theta}\| \le C_s/N$. The score g is L_ψ-Lipschitz in ψ, and $\partial_u g_j$ (defined in the proof of Lemma F.3) is C_θ-Lipschitz in θ, on the fitted range.

TOOLS

The proof rests on one observation. Flipping a null p-value leaves the data just as likely (A1), so on average a null is as often counted through its own score as through its mirror score, and the false rejections match the mirror count, which the step-up rule keeps below q times the rejections. The only error comes from the flip changing everything else: it moves the fit and hence every score, and it can move the cutoff. The four lemmas set this up. Lemma F.1 is the bookkeeping of the step-up rule, Lemma F.2 states the effect of the flip, Lemma F.3 bounds how far the flip moves the fit and the scores, and Lemma F.4 shows that such a move changes a hypothesis's position only if the hypothesis is close to the cutoff.

Lemma F.1 (Ratio bound)**.** *On every realization $A(\hat{t}) \le q|\mathcal{R}|$; hence $\mathbb{E}[\tilde{V}] \le q\,\mathbb{E}|\mathcal{R}|$.*

Proof. If $\hat{t} = -\infty$ then $A(\hat{t}) = |\mathcal{R}| = 0$. Otherwise $1 + A(\hat{t}) \le q\,(1 \vee R(\hat{t}))$; since $q < 1$ this forces $R(\hat{t}) \ge 1$, so $A(\hat{t}) \le q|\mathcal{R}| - 1$. Finally $\tilde{V} \le A(\hat{t})$. □

Lemma F.2 (Flip of a null)**.** *Fix $i \in \mathcal{H}_0$. Under (A1), $\mathcal{D}^{(i)}$ has the same law as $\mathcal{D}$, and $B_i(\mathcal{D}^{(i)}) = 1 - B_i(\mathcal{D})$ almost surely.*

Proof. Given the labels, p_i is independent of the rest of the data and uniform, and $U[0,1]$ is invariant under $p \mapsto 1-p$; averaging over the labels gives the first claim. The map $p_i \mapsto 1 - p_i$ exchanges $\{p_i < 1/2\}$ and $\{p_i > 1/2\}$, and $\{p_i = 1/2\}$ has probability zero. □

Lemma F.3 (Stability of the plug-in surface)**.** *Fix $i \in \mathcal{H}_0$ and assume (A3). Then*

$$\beta_i := \max_j \left|\hat{\alpha}^{(i)}(\mathrm{loc}_j) - \hat{\alpha}(\mathrm{loc}_j)\right| \;\le\; \bar{\beta} := \frac{\kappa^2\,(L_\ell + C_\theta C_s/2)}{\lambda_{\mathrm{reg}}} + \frac{C_s}{N}, \qquad L_\ell = \frac{1}{\varepsilon},$$

and, between $\mathcal{D}$ and $\mathcal{D}^{(i)}$, every T_j, $\tilde{T}_j$, ℓ_j and h_j moves by at most

$$\delta := \frac{\bar{\beta}}{4\varepsilon(1-\varepsilon)} + \frac{L_\psi C_s}{N} = O\Big(\frac{1}{\lambda_{\mathrm{reg}}}\Big) + O\Big(\frac{1}{N}\Big),$$

where for $j = i$ the pair $\{T_i, \tilde{T}_i\}$ is compared as an unordered pair.

Proof. The argument is the uniform-stability argument of (Bousquet and Elisseeff, 2002), adapted to our objective. At the observed locations $\hat{\alpha}(\mathrm{loc}_j) = \hat{\bar{\alpha}} + \hat{f}(\mathrm{loc}_j)$, where $\hat{f} \in \mathcal{H}_K$ minimizes $J(f) = \sum_j g_j(f(\mathrm{loc}_j); \hat{\theta}) + \lambda_{\mathrm{reg}} \|f\|^2_{\mathcal{H}_K}$ with

$$g_j(u;\theta) = -\log\big\{(\bar{\alpha}+u) + (1-\bar{\alpha}-u)\, f_1(p_j;\psi)\big\} + \lambda_{\mathrm{bound}}\big[(\bar{\alpha}+u-1)_+^2 + (-\bar{\alpha}-u)_+^2\big],$$

convex in u; this is objective (5) with $f_0 \equiv 1$ and the centering of Section 4. Let $J^{(i)}$ be the same objective on $\mathcal{D}^{(i)}$ and $\Delta = \hat{f}^{(i)} - \hat{f}$. Both objectives are $2\lambda_{\mathrm{reg}}$-strongly convex, so $J(\hat{f}^{(i)}) - J(\hat{f}) \geq \lambda_{\mathrm{reg}}\|\Delta\|^2$ and $J^{(i)}(\hat{f}) - J^{(i)}(\hat{f}^{(i)}) \geq \lambda_{\mathrm{reg}}\|\Delta\|^2$. Adding the two, the penalty cancels, and with $q_j = g_j(\cdot;\hat{\theta}) - g_j^{(i)}(\cdot;\hat{\theta}^{(i)})$,

$$2\lambda_{\mathrm{reg}}\|\Delta\|^2 \leq \sum_j \big[q_j(\hat{f}^{(i)}(\mathrm{loc}_j)) - q_j(\hat{f}(\mathrm{loc}_j))\big] \leq \kappa\,\|\Delta\| \sum_j \sup|q_j'|,$$

the supremum taken over the segment between $\hat{f}(\mathrm{loc}_j)$ and $\hat{f}^{(i)}(\mathrm{loc}_j)$, using $|\Delta(x)| \leq \kappa\|\Delta\|$. On that segment the boundary term vanishes by (A3), and $|\partial_u g_j| = |1-f_1|/\{s+(1-s)f_1\} \leq \max(1/s, 1/(1-s)) \leq L_\ell$ with $s = \bar{\alpha}+u$. For $j = i$ both p_i and θ change, so $\sup|q_i'| \leq 2L_\ell$; for $j \neq i$ only θ changes, so $\sup|q_j'| \leq C_\theta C_s/N$. Hence $\max_j|\Delta(\mathrm{loc}_j)| \leq \kappa\|\Delta\| \leq \kappa^2(2L_\ell + C_\theta C_s)/(2\lambda_{\mathrm{reg}})$, and $\beta_i \leq |\hat{\bar{\alpha}}^{(i)} - \hat{\bar{\alpha}}| + \max_j|\Delta(\mathrm{loc}_j)| \leq \bar{\beta}$. For the scores, $\partial_s g = f_1/\{s+(1-s)f_1\}^2 \leq 1/\{4s(1-s)\}$, maximizing over $f_1 \geq 0$, and g is L_ψ-Lipschitz in ψ, so for $x \in \{p_j, 1-p_j\}$, $|g(\hat{\alpha}^{(i)}(\mathrm{loc}_j), x; \hat{\psi}^{(i)}) - g(\hat{\alpha}(\mathrm{loc}_j), x; \hat{\psi})| \leq \delta$. For $j \neq i$ this bounds the moves of T_j and $\tilde{T}_j$; for $j = i$ the flip exchanges the two arguments, so the unordered pair moves by the same amount. The minimum ℓ_j and maximum h_j of a pair move by no more than the pair. □

Lemma F.4 (Nudge). *Fix $i \in \mathcal{H}_0$, let $\delta_i = \max_j \max\big(|\ell_j^{(i)} - \ell_j|, |h_j^{(i)} - h_j|\big)$ and $\Delta_i = |\hat{t}^{(i)} - \hat{t}|$, with $\Delta_i = 0$ if both cutoffs are $-\infty$ and $\Delta_i = \infty$ if exactly one is. If $\min\big(|\hat{t}-\ell_j|, |\hat{t}-h_j|\big) > \delta_i + \Delta_i$, then $\mathbf{1}\{\ell_j \leq \hat{t}\} = \mathbf{1}\{\ell_j^{(i)} \leq \hat{t}^{(i)}\}$ and $\mathbf{1}\{h_j \leq \hat{t}\} = \mathbf{1}\{h_j^{(i)} \leq \hat{t}^{(i)}\}$: hypothesis j sits on the same side of the cutoff in both data sets. Under (A3), $\delta_i \leq \delta$ by Lemma F.3.*

Proof. $|(\ell_j^{(i)} - \hat{t}^{(i)}) - (\ell_j - \hat{t})| \leq \delta_i + \Delta_i < |\ell_j - \hat{t}|$, so the two differences have the same sign; likewise for h_j. □

A flip therefore acts on the decisions only through a band around the cutoff, whose width is the nudge δ_i plus the cutoff's own move Δ_i. With count data, attainable scores near the cutoff are separated by a gap, and the band is empty when $\delta_i + \Delta_i$ is below it; with continuous p-values there is no gap, and the band is thin rather than empty when scores do not pile up at the cutoff (Corollary F.6). Both δ_i and Δ_i are computable on the data by refitting with one p-value flipped.

Main result

Putting the tools together: the expected number of false rejections equals the expected mirror count, which Lemma F.1 keeps below q times the expected rejections, plus a correction from nulls whose position relative to the cutoff changes when their own p-value is flipped. By Lemma F.4, only nulls inside the band can contribute to that correction, each by at most one half of its probability of lying there.

Theorem F.5 (mFDR of Rule 2). *Under (A1)–(A2), with $\eta_i = \min\big(|\hat{t} - \ell_i|, |\hat{t} - h_i|\big)$ and $\mathbb{E}|\mathcal{R}| > 0$,*

$$\mathrm{mFDR}(\mathcal{R}) = \frac{\mathbb{E}[V]}{\mathbb{E}|\mathcal{R}|} \;\leq\; q + \frac{1}{2\,\mathbb{E}|\mathcal{R}|} \sum_{i \in \mathcal{H}_0} \Pr\big(\eta_i \leq \delta_i + \Delta_i\big).$$

No stability assumption is used, and $\hat{\alpha}$ need not be correctly specified.

Proof. Step 1: slack identity. Fix $i \in \mathcal{H}_0$. By (A2), $T_i = \ell_i$ when $B_i = 0$ and $T_i = h_i$ when $B_i = 1$, so $\mathbf{1}\{T_i \leq \hat{t}\} - \mathbf{1}\{\tilde{T}_i \leq \hat{t}\} = \mathbf{1}\{\mathcal{A}_i\}(1-2B_i)$ with $\mathcal{A}_i = \{\ell_i \leq \hat{t} < h_i\}$; a null with $h_i \leq \hat{t}$ is counted in both V and $\tilde{V}$ and contributes zero. Summing over the nulls, $V - \tilde{V} = \sum_{i \in \mathcal{H}_0} \mathbf{1}\{\mathcal{A}_i\}(1-2B_i)$.

Step 2: each null's term. Let $s_i = \mathbb{E}[\mathbf{1}\{\mathcal{A}_i\}(1-2B_i)]$ and let $\mathcal{A}_i^{(i)}$ be the same event on $\mathcal{D}^{(i)}$. By Lemma F.2, $s_i = \mathbb{E}[\mathbf{1}\{\mathcal{A}_i^{(i)}\}(1-2B_i(\mathcal{D}^{(i)}))] = -\mathbb{E}[\mathbf{1}\{\mathcal{A}_i^{(i)}\}(1-2B_i)]$. Averaging the two expressions, $s_i = \frac{1}{2}\,\mathbb{E}\big[(\mathbf{1}\{\mathcal{A}_i\} - \mathbf{1}\{\mathcal{A}_i^{(i)}\})(1-2B_i)\big]$, so $|s_i| \le \frac{1}{2}\Pr(\mathcal{A}_i \neq \mathcal{A}_i^{(i)})$. The event $\mathcal{A}_i$ is fixed by the sides of ℓ_i and h_i relative to the cutoff, so by Lemma F.4 it can differ from $\mathcal{A}_i^{(i)}$ only if $\eta_i \le \delta_i + \Delta_i$.

Step 3. By Lemma F.1, $\mathbb{E}[V] = \mathbb{E}[\tilde{V}] + \sum_{i\in\mathcal{H}_0} s_i \le q\,\mathbb{E}|\mathcal{R}| + \frac{1}{2}\sum_{i\in\mathcal{H}_0}\Pr(\eta_i \le \delta_i + \Delta_i)$; divide by $\mathbb{E}|\mathcal{R}|$. □

Theorem F.5 leaves the size of the band as a probability. When a single flip moves the cutoff only a little and the scores do not pile up near it, the band holds a fraction of order δ of the nulls, and the bound becomes a rate.

Corollary F.6 (Rate under threshold stability)**.** *Assume (A1)–(A3) and, in addition: (i) threshold stability, $\Delta_i \le \bar{\Delta}$ for every $i \in \mathcal{H}_0$; (ii) bounded boundary density, $\sum_{i\in\mathcal{H}_0}\Pr(\eta_i \le w) \le \rho\, w\, N$ for $w \le \delta + \bar{\Delta}$; (iii) $\mathbb{E}|\mathcal{R}| \ge r_0 N$. Then*

$$\mathrm{mFDR}(\mathcal{R}) \le q + \frac{\rho}{2r_0}\,(\delta + \bar{\Delta}).$$

If the ratio $(1+A(t))/(1\vee R(t))$ crosses q transversally, so that $\bar{\Delta} = O(\delta) + O(1/N)$, this is $\mathrm{mFDR}(\mathcal{R}) \le q + O(1/\lambda_{\mathrm{reg}}) + O(1/N)$.

Proof. Insert $\delta_i \le \delta$ (Lemma F.3) and (i)–(iii) into Theorem F.5. □

Condition (i) is the one input not derived from the model: $\hat{t}$ is a global function of all scores, and when the ratio stays close to q over a range of t, a single flip can move it across that range. Theorem F.5 does not need it; the corollary uses it only to turn the band into a rate.

Why one flip moves the cutoff by $O(1/N)$. Write $\hat{F}(t) = (1+A(t))/(1\vee R(t))$ for the estimated false-discovery proportion. Flipping one null exchanges its two scores, so $A(t)$ and $R(t)$ each change by at most one for every t, a vertical move of at most $1/R(t)$ in $\hat{F}$; every other score moves by at most δ (Lemma F.3), a horizontal shift of at most δ. If $\hat{F}$ crosses q at $\hat{t}$ with slope $\sigma = \{a(\hat{t}) - q\,r(\hat{t})\}/R(\hat{t})$, where a and r count mirror scores and all scores per unit of t at the cutoff, the crossing moves by at most $\Delta \lesssim \delta + 1/\{a(\hat{t}) - q\,r(\hat{t})\}$. With $N \approx 10^3$, a few hundred nulls and score densities of order one, $a(\hat{t}) - q\,r(\hat{t})$ is a few hundred, so the second term is a few 10^{-3} and shrinks as $1/N$; the pooled scalars $(\hat{\bar{\alpha}}, \hat{\psi})$ move by $O(1/N)$ for the same reason, since a flip changes one summand in an average over N. The argument gives nothing when $a(\hat{t}) \approx q\,r(\hat{t})$, where $\hat{F}$ runs along q instead of crossing it and the cutoff can jump to the next crossing; this is the case that (i) excludes. Both δ_i and Δ_i are computable: refit with p_i flipped and recompute $\hat{t}$.

An exact variant

As explained above- if the fitted surface cannot see on which side of $1/2$ the null p-values fell, flipping a null changes nothing but its own side, and no correction is needed. Fitting on the masked values $\min(p_j, 1-p_j)$ achieves this, at a cost in power, and gives exact FDR control.

Proposition F.7 (Exact FDR for the masked variant)**.** *Let $m_j = \min(p_j, 1-p_j)$. Fit the surface and the global quantities on $\{(\mathrm{loc}_j, m_j)\}_{j=1}^N$ with the folded components $f_0^{\mathrm{m}}(m) = 2$ and $f_1^{\mathrm{m}}(m) = f_1(m) + f_1(1-m)$ on $[0,1/2]$, which keep the mixing weight $\alpha(\mathrm{loc})$, giving $\hat{\alpha}^{\mathrm{m}}$ and $\hat{\psi}^{\mathrm{m}}$; compute $T_j, \tilde{T}_j$ with them. Count each hypothesis on the side of its smaller score, $R^{\mathrm{m}}(t) = \#\{j : T_j \le t,\ T_j < \tilde{T}_j\}$ and $A^{\mathrm{m}}(t) = \#\{j : \tilde{T}_j \le t,\ \tilde{T}_j < T_j\}$, choose $\hat{t}$ as in equation 11 with these counts, and reject $\{j : T_j \le \hat{t},\ T_j < \tilde{T}_j\}$. Under (A1)–(A2), $\mathrm{FDR} \le q$ for every N, with no condition on $\hat{\alpha}^{\mathrm{m}}$.*

Proof. Set $W_j = (1-2B_j)(1-\ell_j)$. By (A2), $R^{\mathrm{m}}(t) = \#\{j : W_j \ge 1-t\}$ and $A^{\mathrm{m}}(t) = \#\{j : W_j \le -(1-t)\}$, so the procedure is the knockoff+ filter of (Barber and Candès, 2015) applied to W

at level q. The magnitudes $|W_j| = 1 - \ell_j$, with $\ell_j = g(\hat{\alpha}^{\mathrm{m}}(\mathrm{loc}_j), m_j; \hat{\psi}^{\mathrm{m}})$ by (A2), are functions of the masked data; by (A1), conditional on the masked data and on the non-null p-values, the null signs $1 - 2B_j$ are i.i.d. fair coins. This is the sign-flip property under which knockoff+ controls FDR at level q (Barber and Candès, 2015). □

The masking is that of AdaPT (Lei and Fithian, 2018), applied once with the spatial model as the learner and without the interactive updates; **?** use the same fit-on-masked-data-then-filter construction for knockoff statistics.

Remarks

Remark F.8 (Hyperparameters)**.** *All results treat λ_{reg} and the kernel as fixed. In our experiments they are selected by cross-validation (Appendix J) and then held fixed; the guarantees are for the selected values.*

Remark F.9 (Counting on both sides)**.** *Rule 2 (Eq. equation 11) counts a hypothesis whose two scores both lie below the cutoff in R and in A. Theorem F.5 and Corollary F.6 cover this, since such a hypothesis contributes zero to the slack identity. Proposition F.7 counts each hypothesis once, on the side of its smaller score, which is what the knockoff+ argument requires; the two counts coincide unless some hypothesis has both scores below the cutoff.*

Remark F.10 (Misspecification)**.** *None of the results requires $\hat{\alpha}$ to be correct: Theorem F.5 and Corollary F.6 depend on the surface only through δ_i and Δ_i, and Proposition F.7 only through masking. Correct specification is what maximizes power (Appendix F).*

G Optimality of the Spatial lfdr Ranking

Theorem G.1 (Optimality of the spatial lfdr ranking)**.** *Assume correct specification: $\hat{\alpha} = \alpha^*$, $(\bar{\alpha}, f_0, f_1)$ are the true marginal quantities, and the triples $(\mathrm{loc}_i, \theta_i, p_i)$ are independent across i (Appendix B). For a level q, let $\Pi_{\mathrm{sp}}(q)$ be the expected number of true discoveries of the sublevel set of $\mathrm{lfdr}_{\mathrm{spatial}}$ whose mFDR equals q, and $\Pi_{\mathrm{p}}(q)$ that of any rule measurable with respect to the p-values $\{p_i\}_{i=1}^N$ alone, in particular any rule ranking by $\mathrm{lfdr}_{\mathrm{marg}}$, with mFDR at most q. Then $\Pi_{\mathrm{sp}}(q) \geq \Pi_{\mathrm{p}}(q)$ for every attainable q, with strict inequality at every q at which, with positive probability, the p-value-only rule rejects a hypothesis while retaining one with a smaller true posterior null probability.*

Proof. Write $\mathcal{D} = \{(\mathrm{loc}_j, p_j)\}_{j=1}^N$, $\theta_i = 1$ for alternatives, and $\mathrm{lfdr}_i^* = P(\theta_i = 0 \mid \mathcal{D}) = P(\theta_i = 0 \mid \mathrm{loc}_i, p_i)$ by independence.

Step 1: the optimal frontier. For a level ℓ let $\Pi^*(\ell) = \sup\{\mathbb{E}[\sum_i \theta_i \delta_i] : \delta\ \mathcal{D}\text{-measurable},\ \mathrm{mFDR}(\delta) \leq \ell\}$. For any $\mathcal{D}$-measurable δ, $\mathbb{E}[\sum_i \theta_i \delta_i] = \mathbb{E}[\sum_i (1 - \mathrm{lfdr}_i^*)\delta_i]$ and $\mathbb{E}[V] = \mathbb{E}[\sum_i \mathrm{lfdr}_i^* \delta_i]$, so $\Pi^*(\ell)$ maximizes the first expression subject to $\mathbb{E}[\sum_i (\mathrm{lfdr}_i^* - \ell)\delta_i] \leq 0$. This is a linear program whose solution rejects a sublevel set $\{i : \mathrm{lfdr}_i^* \leq t\}$ (Sun and Cai, 2007), randomized at the boundary if lfdr^* has atoms; its mFDR is $\rho(t) = \mathbb{E}[\sum_i \mathrm{lfdr}_i^* \mathbf{1}\{\mathrm{lfdr}_i^* \leq t\}]/\mathbb{E}[\sum_i \mathbf{1}\{\mathrm{lfdr}_i^* \leq t\}]$, and as t increases these sets trace Π^*. A level q is attainable if $q = \rho(t)$ for some t.

Step 2: the spatial lfdr attains the frontier. Under correct specification, Bayes' theorem applied to Eq. (1) gives $\mathrm{lfdr}_{\mathrm{spatial}}(p_i, \mathrm{loc}_i) = \alpha^*(\mathrm{loc}_i) f_0(p_i) / \{\alpha^*(\mathrm{loc}_i) f_0(p_i) + (1 - \alpha^*(\mathrm{loc}_i)) f_1(p_i)\} = \mathrm{lfdr}_i^*$. The sublevel sets of $\mathrm{lfdr}_{\mathrm{spatial}}$ are therefore those of Step 1, and $\Pi_{\mathrm{sp}}(q) = \Pi^*(q)$ at every attainable q.

Step 3: p-value-only rules lie on or below the frontier. A rule measurable with respect to the p-values alone is $\mathcal{D}$-measurable with mFDR at most q, so $\Pi_{\mathrm{p}}(q) \leq \Pi^*(q) = \Pi_{\mathrm{sp}}(q)$. The loss has a concrete source: two hypotheses with equal p-values but $\alpha^*(\mathrm{loc}_i) < \alpha^*(\mathrm{loc}_j)$ have $\mathrm{lfdr}_i^* < \mathrm{lfdr}_j^*$, since the lfdr is strictly increasing in its prior argument wherever $f_0, f_1 > 0$, yet a p-value-only rule cannot prefer i over j.

Step 4: strictness by exchange. Suppose that, with positive probability, the p-value-only rule rejects some i while retaining some j with $\mathrm{lfdr}_j^* < \mathrm{lfdr}_i^*$. On this event, choosing the pair by a fixed tie-breaking order so that the new rule is $\mathcal{D}$-measurable, reject j instead of i. The number of rejections

is unchanged and $\mathbb{E}[V]$ decreases by $\mathbb{E}[(\mathrm{lfdr}_i^* - \mathrm{lfdr}_j^*)\mathbf{1}\{\text{event}\}] > 0$, so the modified rule still has mFDR at most q, while its expected number of true discoveries increases by the same amount. Hence $\Pi_{\mathrm{p}}(q) < \Pi^*(q) = \Pi_{\mathrm{sp}}(q)$. □

Remark G.2 (When the gain is strict)**.** *Such inversions require α^* to be non-constant: if $\alpha^* \equiv \bar{\alpha}$, then $\mathrm{lfdr}_i^* = \mathrm{lfdr}_{\mathrm{marg}}(p_i)$ is a function of p_i alone and the marginal ranking has none. Non-constancy is not sufficient: if f_0 and f_1 have disjoint supports, $\mathrm{lfdr}_i^* \in \{0, 1\}$ is already determined by p_i, and again no inversion occurs.*

Remark G.3 (Scope)**.** *The theorem compares rankings at matched mFDR. It does not compare Rule 1 and Rule 2 as deployed: both reject sublevel sets of $\mathrm{lfdr}_{\mathrm{spatial}}$ and differ only in their data-dependent cutoffs.*

H Marginal Density Estimation

The estimation of the marginal component densities, f_0 and f_1, typically follows one of two paradigms, each with distinct limitations. The first approach assumes the inputs are well-calibrated p-values, enforcing a strict Uniform distribution for the null hypothesis ($f_0 \sim U[0, 1]$). While theoretically sound, this direction fails fundamentally when the input data are z-scores or other test statistics, and the calibration itself is rare in practice (Efron, 2004).

The second approach operates on z-values (or generic continuous data) and relies on general mixture models to separate the null and alternative distributions based on their shape. One might assume this direction could handle p-values by simply treating them as bounded data points. However, a nuanced problem arises: when the null hypothesis is truly Uniform, it manifests as a highly "overdispersed" or maximal-entropy background relative to the signal. Standard clustering or separation algorithms, which typically expect compact modes for both classes, fail to identify the signal against this flat, featureless background. This dilemma, where p-value methods cannot handle z-scores, and z-score methods fail on p-values due to null overdispersion, motivated many hybrid strategies, which we did not dive into here.

In this paper, we hold $a < 1$ fixed; it encodes the canonical alternative-density shape (singularity at $p = 0$, monotone decrease on $(0, 1]$). We estimate $\bar{\alpha}$ with Storey's estimator (Storey, 2003) with Storey's Bootstrap selection rule for λ over $\Lambda = \{0.05, 0.10, \ldots, 0.90\}$. That is, for each $\lambda \in \Lambda$ we compute J replicates $\hat{\bar{\alpha}}^{*(j)}(\lambda)$ from $\hat{p}$-values resampled with replacement and estimate

$$\widehat{\mathrm{MSE}}(\lambda) = \frac{1}{J} \sum_{j=1}^{J} \left(\hat{\bar{\alpha}}^{*(j)}(\lambda) - \min_{\lambda' \in \Lambda} \hat{\bar{\alpha}}(\lambda') \right)^2.$$

We then set $\hat{\bar{\alpha}} = \hat{\bar{\alpha}}(\hat{\lambda})$ with $\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \widehat{\mathrm{MSE}}(\lambda)$. The grid minimum serves only as the bootstrap target. Given $\hat{\bar{\alpha}}$, b is estimated by maximum likelihood on the marginal values, which is continuous in b, so a maximizer exists.

I Summary of Guarantees and Assumptions

Table 2 lists, for every result of Appendice 5, what it guarantees and which assumptions it uses. Assumptions are stated in Appendix F; (PC) in Section 5.

J Kernel Selection Principles

Having established both point-wise and entire-domain solution approaches, we now address the question of kernel choice. The reproducing kernel $K(\cdot, \cdot)$ serves a dual role: it defines the smoothness of $\alpha(x)$ through the RKHS norm $\|\alpha\|_{\mathcal{H}_K}^2 = c^T K c$ (or, in centered form, $\|\alpha - \bar{\alpha}\|_{\mathcal{H}_K}^2 = (\mathbf{c} - \mathbf{c}_{\bar{\alpha}})^T K (\mathbf{c} - \mathbf{c}_{\bar{\alpha}})$), and it determines the interpolation behavior between observed locations- essential for predictions at new points in the entire-domain setting. A fundamental requirement for any symmetric function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ to serve as a reproducing kernel is positive definiteness. The natural gradient is defined through the metric K; Proposition 4.1 avoids forming K^{-1}, but the metric must be non-degenerate, so we require the Gram matrix to be strictly positive definite ($K \succ 0$). This

Table 2: What each result guarantees and what it assumes. (A1) conditional independence of the p-values given the labels with $U[0,1]$ nulls; (A2) $f_1(p) > f_1(1-p)$ for $p < 1/2$; (A3) fitted range in $[\varepsilon, 1-\varepsilon]$ and flip-stable $(\bar{\alpha}, \psi)$; (PC) the posterior-comparison condition.

Result	Object	Guarantee	Assumptions used
Prop. 3.1	marginal mixture	$f(z) = \bar{\alpha} f_0 + (1-\bar{\alpha}) f_1$; $\alpha(\cdot)$ and (f_0, f_1) separable	locations i.i.d. from $p(\mathrm{loc})$, or a fixed design with $\bar{\alpha} = N^{-1} \sum_i \alpha(\mathrm{loc}_i)$
Thm. C.1(a)	Rule 1, deployed	$\mathrm{FDR} \le q$ for every N	independence across i; $\mathrm{lfdr}_{\mathrm{marg}}$ at the true $(\bar{\alpha}, f_0, f_1)$; (PC). Not assumed: correctness of $\hat{\alpha}$, sample splitting, any property of the ordering beyond measurability
Thm. C.1(b)	Rule 1	$\mathrm{mFDR} \le q$	as (a), with the covariance form of (PC)
Rem. C.2	marginal ordering	(PC) holds with equality; the rule of Sun and Cai (2007) is the special case	as Thm. C.1
Rem. C.5	Rule 1 vs. marginal rule	$\lvert R_1 \rvert \le \lvert R_{\mathrm{marg}} \rvert$ on every realization	none
Lem. F.1	Rule 2	$\tilde{V} \le A(\hat{t}) \le q \lvert \mathcal{R} \rvert$ on every realization	none
Thm. F.5	Rule 2, deployed	$\mathrm{mFDR} \le q + \frac{1}{2\,\mathbb{E}\lvert\mathcal{R}\rvert} \sum_{i \in \mathcal{H}_0} \Pr(\eta_i \le \delta_i + \Delta_i)$	(A1), (A2). No stability assumption; the band term is not quantified
Cor. F.6	Rule 2, deployed	$\mathrm{mFDR} \le q + \rho(\delta + \bar{\Delta})/(2r_0)$, $\delta = O(1/\lambda_{\mathrm{reg}}) + O(1/N)$	(A1)–(A3); threshold stability $\Delta_i \le \bar{\Delta}$; bounded boundary density ρ; $\mathbb{E}\lvert\mathcal{R}\rvert \ge r_0 N$; λ_{reg} and kernel fixed (Rem. F.8)
Prop. F.7	Rule 2, masked fit	$\mathrm{FDR} \le q$ exactly, for every N	(A1), (A2); $\hat{\alpha}, \hat{\psi}$ fit on $\min(p_i, 1-p_i)$
Thm. G.1	ranking by $\mathrm{lfdr}_{\mathrm{spatial}}$	maximal expected true discoveries at $\mathrm{mFDR} \le q$; strict over p-value-only rules when inversions occur	correct specification (true α^*, f_0, f_1); mFDR criterion; no statement about the cutoffs of Rules 1–2
App. E	CV criterion	$f(z \mid \mathrm{loc}; \theta)$ integrates to one for every θ	$\alpha(\mathrm{loc}; \theta) \in [0,1]$ at the evaluated location

also ensures uniqueness of the solution in coefficient space. Most standard kernels (Gaussian, Matérn) satisfy this property when evaluated on distinct points.

While positive definiteness guarantees a valid RKHS, the *differentiability* of the kernel determines the smoothness properties of functions in that space. This relationship is formalized through Sobolev space theory. A Sobolev space $W^{m,2}(\mathbb{R}^d)$ contains functions whose derivatives up to order m are square-integrable, where m quantifies smoothness. An RKHS can often be identified with a specific Sobolev space, with the kernel's differentiability determining m. A key result is the Sobolev embedding theorem is that if $m > d/2$, then functions in $W^{m,2}(\mathbb{R}^d)$ are guaranteed to be continuous and bounded.

The smoothness requirements depend on the application. If the goal is solely to estimate $\alpha(\mathrm{loc}_i)$ at observed locations, kernel differentiability is not strictly required- the optimization machinery and Representer Theorem only require kernel evaluation at data points. However, when interpolating to new locations in the entire-domain setting, kernel differentiability becomes essential. By choosing a smooth kernel satisfying $m > d/2$, we impose the prior belief that $\alpha(x)$ varies continuously across space, ensuring the function behaves predictably between observed points. For example, the linear kernel $K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$ (with a small ridge added to the Gram matrix, whose rank is at most d) is a valid choice for point-wise estimation, where it effectively uses the correlation between location vectors as a similarity measure. However, for entire-domain generalization, the linear kernel is constrained by its limited smoothness properties. Given our need for a flexible, configurable kernel suitable for both point-wise and entire-domain settings, in our evaluations we employ the **Matérn kernel family** which includes a smoothness parameter $\nu > 0$ that explicitly controls the differentiability. For the Matérn kernel the Sobolev order is $m = \nu + d/2$ (Kanagawa et al., 2018), so the embedding condition $m > d/2$ holds for every $\nu > 0$; larger ν yields k continuous derivatives for every integer $k < \nu$.

J.1 Graph Kernels: Bridging Discrete and Continuous FDR

A significant conceptual advance of our framework is the connection it establishes between smooth FDR methods and graph kernel theory. While we focused on continuous spatial domains $\mathcal{H} \subset \mathbb{R}^d$, many multiple testing problems arise on discrete structures: gene regulatory networks, phylogenetic trees, gene ontology hierarchies, and social networks. Indeed, most current smooth FDR methods operate on such discrete graph representations, requiring topology-specific algorithmic solutions where methods designed for trees cannot be applied to scale-free networks, and techniques for hierarchical structures fail on general graphs. Our RKHS framework provides a unified solution: by choosing an appropriate graph kernel, the same optimization algorithm applies to arbitrary discrete topologies.

For a comprehensive treatment of graph kernels, we refer the reader to the seminal work of Kondor and Lafferty (2002) and its subsequent extensions by Smola and Kondor (2003). Within this framework, the *Graph Laplacian kernel* ($K = L^\dagger$) serves as a canonical example. Its associated RKHS norm, $\|\alpha\|^2_{\mathcal{H}} = \boldsymbol{\alpha}^T L \boldsymbol{\alpha} = \sum_{(i,j)\in E} W_{ij}(\alpha_i - \alpha_j)^2$, explicitly penalizes signal differences across edges using an L_2 metric. This formulation contrasts fundamentally with the Total Variation (TV) penalty ($\sum W_{ij}|\alpha_i - \alpha_j|$) employed by current graph-based FDR methods, whereas TV relies on the L_1 norm to enforce piecewise-constant clustering, the Laplacian kernel promotes smooth variation across the network. To capture longer-range dependencies, this approach naturally extends to *diffusion kernels* ($K = \exp(-\beta L)$) and *random walk kernels*, which incorporate global graph topology and allow distant but well-connected nodes to influence local estimation. A particularly promising direction involves the use of **Hyperbolic Embeddings for Hierarchical Structures**, as many biological and social networks exhibit inherent hierarchical organization : gene ontologies (GO), phylogenetic trees, and protein interaction networks with hub-spoke patterns. For such structures, the seminal result of (Sarkar, 2012) regarding hyperbolic spaces applies. Briefly, Sarkar's Theorem states that any tree with n nodes can be embedded into the 2-dimensional hyperbolic space (Poincaré disk $\mathbb{H}^2$) with arbitrarily low distortion $(1 + \epsilon)$ for any $\epsilon > 0$.This dimension efficiency is crucial for practical application, but more fundamentally, it addresses a geometric incompatibility. Consider gene ontology enrichment testing containing 50,000 terms, but observations at only 500 locations (1% sampling). In Euclidean space, such hierarchical structures face an intrinsic "capacity" problem: trees exhibit exponential volume growth, whereas Euclidean space has only polynomial growth. Consequently, embedding a tree into *any* low-dimensional Euclidean space $\mathbb{R}^d$ must incur significant distortion. Even high-dimensional Euclidean embeddings fail to capture the tree topology faithfully compared to the hyperbolic plane. **In contrast**, by leveraging Sarkar's construction, we can embed these dependencies into just 2 dimensions with near-perfect fidelity, effectively bypassing the limitations of Euclidean kernels for hierarchical data.

K Alternative Approaches for Enforcing $\alpha(x) \in [0, 1]$ Over Continuous Domains

Before developing the convex two-step kernel logistic regression framework, we explored several direct methods for enforcing $\alpha(\text{loc}) \in [0, 1]$ across the entire continuous domain $\mathcal{H}$. This section documents these approaches, their theoretical foundations, and the computational barriers that led us to pursue the convex alternative. We present this analysis to contextualize our methodological choices and guide future research.

K.1 Approach 1: Direct Squashing and Non-Convexity

A common solution for ensuring global constraint satisfaction is to apply a squashing function $\sigma : \mathbb{R} \to [0, 1]$ to an unconstrained function, for example, the logistic function $\sigma(z) = 1/(1 + e^{-z})$. The fundamental issue is that the composition of the logarithm with the mixture under squashing renders the objective non-convex. To see this explicitly, consider the data term for a single observation:

$$\ell(g_i) = -\log\left[\sigma(g_i) f_0(p_i) + (1 - \sigma(g_i)) f_1(p_i)\right] \tag{46}$$

The second derivative with respect to the latent value g_i is:

$$\frac{\partial^2 \ell}{\partial g_i^2} = \frac{\partial}{\partial g_i}\left[\frac{-\sigma'(g_i)(f_0(p_i) - f_1(p_i))}{\sigma(g_i)f_0(p_i) + (1-\sigma(g_i))f_1(p_i)}\right] \tag{47}$$

$$= \frac{-\sigma''(g_i)(f_0 - f_1)(\sigma f_0 + (1-\sigma)f_1) + (\sigma'(g_i))^2(f_0 - f_1)^2}{[\sigma(g_i)f_0(p_i) + (1-\sigma(g_i))f_1(p_i)]^2} \tag{48}$$

For the logistic function, the second derivative $\sigma''(z) = \sigma(z)(1-\sigma(z))(1-2\sigma(z))$ changes sign depending on whether $\sigma(z)$ is above or below 0.5. This means the loss function is neither convex nor concave in g_i, even for a single data point. The global objective, being a sum over N such terms plus a convex regularizer, inherits this non-convexity.

The non-convex landscape leads to multiple stationary points, many of which are local minima. Different initializations of $\mathbf{c}$ can converge to qualitatively different solutions with vastly different objective values. Moreover, the objective surface often exhibits regions where the gradient is near-zero but the Hessian has both positive and negative eigenvalues, causing optimization algorithms to stall. In our experiments, we found that the direct squashing approach failed to converge within reasonable iteration budgets, oscillating between different regions of parameter space. Even when optimization converged, the resulting $\alpha(\text{loc})$ functions often exhibited pathological behavior such as rapid oscillations between extreme values (near 0 and 1) in regions with sparse data.

K.2 Approach 2: The Semi-Infinite Programming (SIP) Formulation

Enforcing $\alpha(x) \in [0,1]$ for all $x \in \mathcal{H}$ is a **semi-infinite programming (SIP)** problem (Hettich and Kortanek, 1993), an optimization with finite decision variables but infinitely many constraints:

$$\begin{aligned} \min_{\mathbf{c}\in\mathbb{R}^N} \quad & -\sum_{i=1}^{N} \log\left[\alpha(x_i)f_0(p_i) + (1-\alpha(x_i))f_1(p_i)\right] + \lambda\|\alpha - \bar{\alpha}\|^2_{\mathcal{H}_K} \\ \text{s.t.} \quad & 0 \le \alpha(x) \le 1, \quad \forall x \in \mathcal{H} \end{aligned} \tag{49}$$

where $\alpha(x) = \sum_{j=1}^{N} c_j K(x, x_j)$.

Classical SIP methods include: (1) discretization with adaptive refinement, (2) exchange methods that iteratively add violated constraints, (3) reduction to finite equivalent constraints via problem structure, and (4) barrier/penalty methods with sampling-based approximation. Here, each method corresponds to a classical SIP approach: local reduction uses KKT finite reduction (Still, 2001), polynomial SDP applies moment relaxations (Henrion et al., 2007), and the barrier method employs interior point penalties (Royset et al., 2003). However, SIP theory assumes convex objectives and our non-convex mixture likelihood eliminates convergence guarantees, and constraint violation patterns depend on kernel choice in complex ways.

Method 1: Local Reduction via Critical Points. Extreme constraint violations can only occur at critical points in the interior or at domain boundaries. For upper boundary violations ($\alpha(x) > 1$), critical points satisfy:

$$\alpha(x) = \sum_{j=1}^{N} c_j K(x, x_j) = 1 \tag{50}$$

$$\nabla_x \alpha(x) = \sum_{j=1}^{N} c_j \nabla_x K(x, x_j) = 0 \tag{51}$$

$$\nabla^2_{xx} \alpha(x) \preceq 0 \tag{52}$$

For Gaussian RBF kernels, this yields a system of $d+1$ nonlinear equations in d unknowns, solvable via Newton-Raphson when well-conditioned. By Bézout's theorem, a system of d polynomial equations of degree $\mathcal{O}(d)$ admits $\mathcal{O}(2^d d^d)$ real solutions. Cost per critical point: $\mathcal{O}(d^3)$ per Newton iteration gives a total cost of:

$$\text{Cost}_{\text{local}} = \mathcal{O}(2^d d^{d+3} N) \tag{53}$$

So, while under compactness of $\mathcal{H}$ and non-degeneracy of K, all constraint-violating critical points are identified in finite time, the solution is intractable for large d due to exponential scaling and ill-conditioning. Moreover, Jacobian conditioning deteriorates as $\kappa(J) \sim \mathcal{O}(N^{d/(d+2)})$, causing numerical instability. In practice, we tried few common solvers, all weren't able to solve the set of equations fro $d \geq 10$.

Method 2: Polynomial SDP Relaxation. Approximate the non-polynomial mixture likelihood via Taylor expansion, then apply semidefinite programming relaxations for global certificates. Expand $\log(h_i(\alpha))$ around $\alpha = 0.5$ with $\beta_i = \alpha(x_i) - 0.5$ and mixing ratio $r_i = (f_0(p_i) - f_1(p_i))/m_i$:

$$\log(h_i) \approx \log(m_i) + \beta_i r_i - \frac{(\beta_i r_i)^2}{2} + \frac{(\beta_i r_i)^3}{3} - \frac{(\beta_i r_i)^4}{4} \tag{54}$$

Substituting $\beta_i = \sum_j c_j K(x_i, x_j) - 0.5$ yields a degree-4 polynomial in $\mathbf{c}$. The Lasserre SDP hierarchy constructs moment matrices $M_k(y)$ of dimension $\binom{N+k}{k}$ that provide increasingly tight convex relaxations.

For bounded densities with $\max_i |f_0(p_i)|, |f_1(p_i)| \leq B$, Taylor error is $\mathcal{O}(|\beta_i|^5)$. As relaxation order $k \to \infty$, SDP converges to the global polynomial optimum. Nevertheless, the computational memory requirements for moment matrix:

$$\text{Memory} = 8 \cdot \frac{\binom{N+k}{k}(\binom{N+k}{k} + 1)}{2} \text{ bytes} \tag{55}$$

Concrete limits:

$$\begin{aligned} N = 50, k = 2: &\quad 7.03 \text{ MB} \\ N = 100, k = 2: &\quad 106 \text{ MB} \\ N = 50, k = 4: &\quad 400 \text{ GB} \end{aligned}$$

Making this solution non-practical.

K.3 Approach 3: Barrier Method with Tail-Aware Sampling

Finally, here we describe the barrier method which we compare against in the evaluations section. Transform the semi-infinite constraint set into a penalized objective using logarithmic barrier functions, then approximate the resulting domain integral via importance sampling that concentrates samples in regions most likely to violate constraints. The classical barrier method for constrained optimization replaces hard constraints with smooth penalty terms that approach infinity at the boundary. For the semi-infinite programming problem equation 49 we reformulate as:

$$\mathcal{L}_\nu(\mathbf{c}) = -\sum_{i=1}^{N} \log(h_i(\alpha(x_i))) + \lambda \|\alpha - \bar{\alpha}\|^2_{\mathcal{H}_K} - \frac{1}{\nu} \int_{\mathcal{H}} [\log(\alpha(x)) + \log(1 - \alpha(x))]\, dx \tag{56}$$

where $\nu > 0$ is the barrier parameter controlling penalty strength and the integral is over the Lebesgue measure on $\mathcal{H} \subseteq \mathbb{R}^d$. The barrier terms $-\log(\alpha(x))$ and $-\log(1 - \alpha(x))$ create increasingly steep penalties as $\alpha(x)$ approaches 0 or 1, respectively. As $\nu \to \infty$, these penalties force $\alpha(x)$ to remain strictly within $(0, 1)$ throughout the domain. The method proceeds by solving a sequence of problems with increasing ν:

$$\nu_0 = 1, \quad \nu_{k+1} = \beta \nu_k, \quad \beta \in [1.1, 1.5] \tag{57}$$

Starting from small ν_0 (weak constraints) allows easier optimization, while gradually increasing ν tightens constraints. Each iteration warm-starts from the previous solution. The barrier objective contains the integral:

$$I(\mathbf{c}) = \int_{\mathcal{H}} [\log(\alpha(x)) + \log(1 - \alpha(x))]\, dx \tag{58}$$

This integral has **no closed form for general kernels and domains**, with direct numerical quadrature (e.g., Gaussian quadrature grids) becomes intractable in large d.

K.3.1 Monte Carlo Approximation via Importance Sampling.

We approximate the integral using Monte Carlo integration with M samples $\{z_m\}_{m=1}^M$ drawn from a proposal distribution $q(x)$:

$$I(\mathbf{c}) = \int_{\mathcal{H}} \frac{f(x)}{q(x)} q(x)\, dx \approx \frac{1}{M} \sum_{m=1}^{M} \frac{f(z_m)}{q(z_m)} \tag{59}$$

where $f(x) = \log(\alpha(x)) + \log(1 - \alpha(x))$ and $q(x)$ is the sampling density. The key question is: *what distribution q minimizes variance and ensures constraint violation detection?* For that, we construct a kernel density estimate from the N observed data locations:

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} K_{\text{KDE}}(x, x_i) \tag{60}$$

where $K_{\text{KDE}}(x, y) = \frac{1}{h^d} K_0\left(\frac{x-y}{h}\right)$ is a probability kernel (typically Gaussian) with bandwidth parameter $h > 0$. The bandwidth h critically determines sampling quality. We employ two standard methods: *1. Scott's Rule*:

$$h_{\text{Scott}} = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} N^{-\frac{1}{d+4}} \hat{\sigma} \tag{61}$$

where $\hat{\sigma}$ is the empirical standard deviation of the data locations (computed per-dimension and averaged). *2. Cross-Validation* (optimal, expensive):

$$h_{\text{CV}} = \arg\min_h \frac{1}{N} \sum_{i=1}^{N} \left(\hat{p}_{-i}(x_i; h) - \delta(x_i)\right)^2 \tag{62}$$

where $\hat{p}_{-i}(x; h)$ is the leave-one-out KDE excluding point x_i, and δ is the Dirac delta. In practice, this minimizes integrated squared error via grid search over candidate h values.

The Critical Observation is that naive sampling from $\hat{p}(x)$ concentrates samples where data is dense. However, constraint violations $\alpha(x) \notin [0, 1]$ may occur most violently in *tail regions* far from training data, where RKHS extrapolation becomes unreliable. Sampling from $\hat{p}(x)$ thus *misses exactly the regions we need to monitor*. To address this, we design a hybrid distribution that balances three competing objectives:

$$p_{\text{hybrid}}(x) = \rho_1 \hat{p}(x) + \rho_2 q_{\text{tail}}(x) + \rho_3 u(x) \tag{63}$$

where: **Data-Dense Zone** ($\rho_1 = 0.3$) are sampled from KDE $\hat{p}(x)$. These regions contribute most to the data-fit term in $\mathcal{L}_\nu$. **Tail Zone** ($\rho_2 = 0.5$) are samples from inverse density:

$$q_{\text{tail}}(x) = \frac{w(x)}{\int_{\mathcal{H}} w(y) dy}, \quad w(x) = \frac{1}{\hat{p}(x) + \epsilon} \tag{64}$$

where $\epsilon > 0$ (typically $\epsilon = 10^{-6}$) prevents division by zero. These are points where violent violations occur, for which we use rejection sampling with acceptance probability $\propto (\hat{p}(x) + \epsilon)^{-1}$. **Uniform Zone** ($\rho_3 = 0.2$) are sampled uniformly:

$$u(x) = \frac{1}{\text{Vol}(\mathcal{H})} \tag{65}$$

Finally, the choice $(\rho_1, \rho_2, \rho_3) = (0.3, 0.5, 0.2)$ comes from hyperparameter optimization.

K.3.2 Practical Limitations

Despite strong theoretical foundations and favorable asymptotic complexity, the barrier method proved impractical in our experiments:

1. **Sampling Sensitivity:** The mixing weights (ρ_1, ρ_2, ρ_3) require problem-specific tuning. Too much tail allocation ($\rho_2 > 0.6$) introduces excessive gradient noise; too little ($\rho_2 < 0.3$) misses violations.

2. **Bandwidth Selection:** KDE bandwidth h critically affects tail sampling quality. Scott's rule often oversmooths for $d \geq 5$; cross-validation is expensive.
3. **Barrier Schedule:** The growth rate β and starting value ν_0 require careful tuning. Aggressive schedules ($\beta > 1.5$) cause convergence failure; conservative schedules ($\beta < 1.1$) waste computation.
4. **Gradient Variance:** Even with $M = 10^4$ samples, Monte Carlo variance in the barrier gradient necessitates small learning rates ($\eta \sim 10^{-3}$), requiring hundreds of iterations per barrier level.
5. **Non-Convexity Persists:** The method still optimizes the non-convex mixture likelihood, inheriting all local minima issues. Multiple random initializations are required, multiplying computational cost.
6. **Numerical Instability:** Near constraint boundaries ($\alpha \approx 0$ or $\alpha \approx 1$), the terms $1/\alpha$ and $1/(1-\alpha)$ become numerically unstable, requiring careful regularization ($\alpha \leftarrow \text{clip}(\alpha, 10^{-8}, 1 - 10^{-8})$).

L Practical Cross-Validation Procedure

In our experiments (Section 7 in main text), we employ the following procedure:

Algorithm 2 Hyperparameter Selection via Cross-Validation

Require: Data $\{(p_i, \text{loc}_i)\}_{i=1}^N$, hyperparameter grid Θ, number of folds K
Ensure: Optimal hyperparameters θ^*

```
Partition data into K folds: D = D_1 ∪ ··· ∪ D_K
for each θ ∈ Θ do
    Initialize cumulative log-likelihood: L_CV(θ) = 0
    for fold k = 1, ..., K do
        Train on D_train^(k) = D \ D_k
        Obtain coefficients c^(k)(θ) by solving Eq. 5
        Compute test log-likelihood:
            L_k(θ) = Σ_{i∈D_k} log f(z_i | loc_i; θ, c^(k))
        Update: L_CV(θ) ← L_CV(θ) + L_k(θ)
    end for
end for
return θ* = arg max_{θ∈Θ} L_CV(θ)
```

M Posterior Variance, Fisher Information, and Prior Variance

Symbol convention. To avoid collision with the mixture density $h_i = \alpha_i f_0(p_i) + (1-\alpha_i) f_1(p_i)$ of §D, we denote the per-observation Fisher information by $\mathcal{I}_i$ throughout this section.

M.1 Fisher Information of the Mixture

Proposition M.1 (Fisher Information)**.** *For the mixture $f(p \mid \alpha) = \alpha f_0(p) + (1-\alpha) f_1(p)$, the observed Fisher information is:*

$$\mathcal{I}(\alpha; p) \;=\; \left(\frac{\partial \log f(p \mid \alpha)}{\partial \alpha}\right)^2 \;=\; \frac{[f_0(p) - f_1(p)]^2}{[\alpha f_0(p) + (1-\alpha) f_1(p)]^2}. \tag{66}$$

At observation i, we write $\mathcal{I}_i := \mathcal{I}(\alpha_i; p_i)$.

Proof. Direct computation:

$$\frac{\partial \log f}{\partial \alpha} \;=\; \frac{f_0(p) - f_1(p)}{\alpha f_0(p) + (1-\alpha) f_1(p)} \quad\Longrightarrow\quad \mathcal{I}(\alpha; p) \;=\; \left(\frac{f_0(p) - f_1(p)}{\alpha f_0(p) + (1-\alpha) f_1(p)}\right)^2. \tag{67}$$

□

Remark M.2 (Fisher Information Behavior)**.** *The Fisher information $\mathcal{I}_i$ exhibits characteristic behavior as a function of α:*

- *When $\alpha \approx 0.5$ and $f_0(p) \approx f_1(p)$: $\mathcal{I}_i \approx 0$ (low information, high statistical uncertainty).*
- *When $\alpha \approx 0$ or $\alpha \approx 1$: One distribution dominates, yielding larger $\mathcal{I}_i$ (more information, lower statistical uncertainty).*

This creates the "Fisher floor" phenomenon: even with perfect spatial coverage, locations with $\alpha \approx 0.5$ have inherently high posterior uncertainty due to the mixture model ambiguity.

M.2 Hessian of the Centered Objective

The Hessian of $\mathcal{L}(\mathbf{c})$ at $\hat{\mathbf{c}}$ is the second derivative of the three terms.

Data term. From $\mathcal{L}_{\text{data}} = -\sum_i \log h_i$ with $h_i = \alpha_i f_0(p_i) + (1-\alpha_i) f_1(p_i)$ and $\alpha_i = (K\mathbf{c})_i$, the second derivative gives

$$\nabla^2 \mathcal{L}_{\text{data}} \;=\; K^T \text{diag}(\mathcal{I}_i)\, K, \tag{68}$$

where $\mathcal{I}_i$ is the per-observation Fisher information of Proposition M.1.

Centered regularization term. For $\mathcal{L}_{\text{reg}} = \lambda_{\text{reg}} (\mathbf{c} - \mathbf{c}_{\bar{\alpha}})^T K (\mathbf{c} - \mathbf{c}_{\bar{\alpha}})$,

$$\nabla^2 \mathcal{L}_{\text{reg}} \;=\; 2\lambda_{\text{reg}} K. \tag{69}$$

(Centering is a linear shift in $\mathbf{c}$ and does not affect curvature.)

Boundary term. With $\mathbf{b}_i = 2 \cdot I(\alpha_i \notin [0,1])$,

$$\nabla^2 \mathcal{L}_{\text{bound}} \;=\; \lambda_{\text{bound}}\, K^T \text{diag}(\mathbf{b})\, K. \tag{70}$$

Total Hessian.

$$H \;=\; K^T \text{diag}(\mathcal{I}_i)\, K \;+\; 2\lambda_{\text{reg}}\, K \;+\; \lambda_{\text{bound}}\, K^T \text{diag}(\mathbf{b})\, K. \tag{71}$$

M.3 Posterior Variance Theorem

Theorem M.3 (Posterior Predictive Variance)**.** *Under the Laplace approximation $\mathbf{c} \mid \textit{data} \sim \mathcal{N}(\hat{\mathbf{c}}, H^{-1})$ with H given by equation 71, the posterior variance of $\alpha(loc)$ is:*

$$\text{Var}[\alpha(\textit{loc}) \mid \textit{data}] = \mathbf{k}(\textit{loc})^T H^{-1} \mathbf{k}(\textit{loc}), \tag{72}$$

where $\mathbf{k}(\textit{loc}) = [K(\textit{loc}, \textit{loc}_{s_1}), \ldots, K(\textit{loc}, \textit{loc}_{s_{N_0}})]^T$.

Proof. Since $\alpha(\text{loc}) = \mathbf{k}(\text{loc})^T \mathbf{c}$ is linear in $\mathbf{c} \sim \mathcal{N}(\hat{\mathbf{c}}, H^{-1})$:

$$\text{Var}[\alpha(\text{loc})] = \mathbf{k}(\text{loc})^T \text{Var}[\mathbf{c}] \mathbf{k}(\text{loc}) = \mathbf{k}(\text{loc})^T H^{-1} \mathbf{k}(\text{loc}) \tag{73}$$

by standard properties of linear transformations of Gaussian random vectors. □

M.4 Properties of the Prior Prediction Variance

Proposition M.4 (Properties of Prior Prediction Variance)**.** *For a strictly positive definite kernel K and regularization parameter $\lambda > 0$, the prior prediction variance:*

$$\sigma^2_{\textit{prior}}(\textit{loc} \mid S) = K(\textit{loc}, \textit{loc}) - \mathbf{k}(\textit{loc})^T (K_S + \lambda I)^{-1} \mathbf{k}(\textit{loc}) \tag{74}$$

satisfies:

(i) ***Boundedness****: $0 \le \sigma^2_{\textit{prior}}(\textit{loc} \mid S) \le K(\textit{loc}, \textit{loc})$ for all $\textit{loc} \in \mathcal{H}$.*

(ii) ***Monotonicity****: If $S_1 \subseteq S_2$, then $\sigma^2_{\textit{prior}}(\textit{loc} \mid S_2) \le \sigma^2_{\textit{prior}}(\textit{loc} \mid S_1)$ for all $\textit{loc} \in \mathcal{H}$.*

(iii) ***Distance decay****: For stationary kernels* $K(loc, loc') = k(\|loc - loc'\|)$ *with* $\lim_{r\to\infty} k(r) = 0$*:*

$$\lim_{\min_{s\in S} \|loc - loc_s\| \to \infty} \sigma^2_{prior}(loc \mid S) = K(loc, loc). \tag{75}$$

Proof. **(i) Boundedness.** The lower bound follows from non-negativity of variance. For the upper bound, since $K_S + \lambda I \succ 0$, its inverse is positive definite. Therefore:

$$\mathbf{k}^T (K_S + \lambda I)^{-1} \mathbf{k} \geq 0 \implies \sigma^2_{\text{prior}} = K(\text{loc}, \text{loc}) - \mathbf{k}^T (K_S + \lambda I)^{-1} \mathbf{k} \leq K(\text{loc}, \text{loc}). \tag{76}$$

(ii) Monotonicity. Let $S_1 \subseteq S_2$ with $|S_1| = n$, $|S_2| = n + m$. Partition:

$$K_{S_2} = \begin{bmatrix} K_{S_1} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}, \quad \mathbf{k}_{S_2} = \begin{bmatrix} \mathbf{k}_{S_1} \\ \mathbf{k}_{\text{new}} \end{bmatrix}. \tag{77}$$

By the Schur complement formula:

$$(K_{S_2} + \lambda I)^{-1} = \begin{bmatrix} M & N \\ N^T & P \end{bmatrix} \tag{78}$$

where $M = (K_{S_1} + \lambda I)^{-1} + (K_{S_1} + \lambda I)^{-1} K_{12} P K_{21} (K_{S_1} + \lambda I)^{-1}$ and $P = [(K_{22} + \lambda I) - K_{21}(K_{S_1} + \lambda I)^{-1} K_{12}]^{-1} \succ 0$.

Computing the quadratic form,

$$\mathbf{k}_{S_2}^T (K_{S_2} + \lambda I)^{-1} \mathbf{k}_{S_2} = \mathbf{k}_{S_1}^T (K_{S_1} + \lambda I)^{-1} \mathbf{k}_{S_1} \; + \; v^T P \, v, \tag{79}$$

where $v = \mathbf{k}_{\text{new}} - K_{21}(K_{S_1} + \lambda I)^{-1} \mathbf{k}_{S_1}$. Since $P \succ 0$, $v^T P v \geq 0$, so

$$\mathbf{k}_{S_2}^T (K_{S_2} + \lambda I)^{-1} \mathbf{k}_{S_2} \;\geq\; \mathbf{k}_{S_1}^T (K_{S_1} + \lambda I)^{-1} \mathbf{k}_{S_1}, \tag{80}$$

which by definition of σ^2_{prior} yields $\sigma^2_{\text{prior}}(\text{loc} \mid S_2) \leq \sigma^2_{\text{prior}}(\text{loc} \mid S_1)$.

(iii) Distance decay. For stationary kernels $K(\text{loc}, \text{loc}') = k(\|\text{loc} - \text{loc}'\|)$ with $k(r) \to 0$, all entries of $\mathbf{k}(\text{loc})$ tend to 0 as $\min_{s\in S} \|\text{loc} - \text{loc}_s\| \to \infty$. Since $K_S + \lambda I \succeq \lambda I$, the operator norm $\|(K_S + \lambda I)^{-1}\|_{\text{op}} \leq 1/\lambda$, so

$$\left|\mathbf{k}^T (K_S + \lambda I)^{-1} \mathbf{k}\right| \;\leq\; \|(K_S + \lambda I)^{-1}\|_{\text{op}} \cdot \|\mathbf{k}\|^2 \;\leq\; \frac{\|\mathbf{k}\|^2}{\lambda} \;\to\; 0, \tag{81}$$

hence $\sigma^2_{\text{prior}}(\text{loc} \mid S) \to K(\text{loc}, \text{loc})$. □

N Density-Aware A-Optimal Design

This section provides the algorithm referenced in Section 6 for selecting N_0 measurement locations that minimize average prediction variance across the spatial domain.

Algorithm 3 Density-Aware Greedy A-Optimal Location Selection

Require: Locations $\{\text{loc}_i\}_{i=1}^{M}$, kernel K, budget N_0, regularization λ, density parameter γ
Ensure: Selected subset S

Compute kernel matrix: $K_{\text{full}} \in \mathbb{R}^{M \times M}$
Compute local density: $\rho_i = \sum_{j \neq i} K(\text{loc}_i, \text{loc}_j)$ for all i
Compute weights: $w_i = \rho_i^{\gamma} / \sum_j \rho_j^{\gamma}$ ▷ Normalize
Initialize: $S \leftarrow \emptyset$, $U \leftarrow \{1, \ldots, M\}$
for $k = 1$ to N_0 **do**
    best_score $\leftarrow +\infty$, best_idx $\leftarrow$ null
    for each $j \in U$ **do**
        $S_{\text{temp}} \leftarrow S \cup \{j\}$
        $K_S \leftarrow K_{\text{full}}[S_{\text{temp}}, S_{\text{temp}}]$
        $L \leftarrow \text{cholesky}(K_S + \lambda I)$
        weighted_var $\leftarrow 0$
        for each $i \in U \setminus \{j\}$ **do**
            $\mathbf{k}_i \leftarrow K_{\text{full}}[i, S_{\text{temp}}]$
            $\mathbf{x}_i \leftarrow L^{-T}(L^{-1}\mathbf{k}_i)$ ▷ Solve using Cholesky
            $\sigma_i^2 \leftarrow K_{\text{full}}[i, i] - \mathbf{k}_i^T \mathbf{x}_i$
            weighted_var $\leftarrow$ weighted_var $+ w_i \cdot \sigma_i^2$
        end for
        if weighted_var $<$ best_score **then**
            best_score $\leftarrow$ weighted_var, best_idx $\leftarrow j$
        end if
    end for
    $S \leftarrow S \cup \{\text{best_idx}\}$, $U \leftarrow U \setminus \{\text{best_idx}\}$
end for
return S

Computational Complexity. The algorithm performs N_0 iterations. In iteration k, evaluating $(M-k)$ candidates each requiring Cholesky factorization $O(k^3)$ and $(M-k)$ variance computations $O(k^2)$ gives $O(M^2k^2)$ per iteration. Total complexity: $O(M^2N_0^3)$.

Density Parameter Selection. The parameter γ controls the cluster-isolation tradeoff:

- $\gamma > 0$: Prioritizes dense regions (population centers). Useful when inference targets are concentrated in clusters.
- $\gamma = 0$: Uniform weighting (balanced coverage). Recommended default for general applications.
- $\gamma < 0$: Prioritizes isolated points (rare events). Useful for outlier detection or when signals are expected in sparse regions.

For most applications, $\gamma \in [-0.5, 0.5]$ provides reasonable balance. Cross-validation over γ is possible but computationally expensive.

O WEIGHTED-BH PROCEDURES WITH SIDE INFORMATION

Three procedures use per-hypothesis side information inside the Benjamini–Hochberg framework rather than inside a two-group model. We describe each in plain terms and state where it differs from our framework.

IHW. IHW (Ignatiadis et al., 2016), with finite-sample theory in Ignatiadis and Huber (2021), is Benjamini–Hochberg with data-driven weights. Each hypothesis carries one scalar covariate that is independent of its p-value under the null. The covariate is cut into a small number of bins, every bin receives a weight, and the weights are chosen to maximize the number of rejections of BH run on the weighted p-values, with the weights constrained to average one. To stop the weights from overfitting, the weights applied to one fold of hypotheses are learned from the other folds; this cross-weighting is

what gives finite-sample FDR control under independence or positive dependence. IHW estimates neither a null proportion nor an alternative density; its gain comes from moving the BH budget toward informative bins. IHW needs a scalar covariate and a binning, so graph or tree structure can enter only through an embedding. It is a p-value weighting rule; we estimate the null prior as a function on the domain and rank by the posterior.

SABHA. SABHA (Li and Barber, 2016b) is Benjamini–Hochberg with weights that estimate the local null proportion under a structural constraint. Each p-value is censored at a level τ (0.5 in their experiments): only whether p_i exceeds τ is used. The censored frequencies are projected onto a constraint set that encodes the structure (ordered, block, or low total variation are the worked cases) and floored at a small constant, giving an estimate $\hat{q}_i$ of the null proportion at each hypothesis. BH is then run on $\hat{q}_i\, p_i$. Because the weights use only the censored part of each p-value, FDR is controlled in finite samples at the target level times a factor slightly above one; the excess is governed by the Rademacher complexity of the constraint set. SABHA's estimate of the null proportion uses one bit per p-value (above or below τ) and a user-chosen constraint set; ours uses the full mixture likelihood and an RKHS norm that can encode any kernel, including graphs and trees. SABHA weights p-values; we rank by the posterior. SABHA's finite-sample guarantee carries an explicit inflation factor; Rule 1 holds at level q (under PC).

LAWS. LAWS (Cai et al., 2021) is designed for spatial domains. It estimates the local sparsity level, the probability that the hypothesis at a location is non-null, by kernel smoothing over neighbouring locations of a screening indicator (whether a p-value exceeds a threshold τ), a localized version of Storey's estimator; no null or alternative density is estimated. Each p-value is divided by the local odds of being non-null, and the weighted p-values are thresholded by a step-up rule whose estimated false-discovery proportion sums the local sparsity levels times the threshold. FDR control is asymptotic (ours finite), only under weak dependence, and consistency of the sparsity estimate. STRAW (Wang et al., 2023) generalizes the weight to a power of the odds and estimates the sparsity by smoothing one minus the local FDR. LAWS's estimator is a bandwidth-based smoother of a screening indicator on a metric space; ours is a penalized maximum-likelihood fit in an RKHS with any kernel. LAWS is a weighted-p-value rule with asymptotic control; we rank by the spatial posterior and budget by the marginal one (Rule 1) or by mirror counting (Rule 2), with non-asymptotic statements.

P Additional Experiments

P.1 Sensitivity of the decision rules to label corruption

This experiment isolates the decision rules from the estimation of the prior. We simulate the semi-synthetic protocol of Section 7 with $N = 1{,}000$ hypotheses per dataset, a signal cluster and a null cluster holding 25% of the hypotheses each and a structureless null background holding the remainder, nulls drawn from $\mathcal{U}[0,1]$ and alternatives from $\mathrm{Beta}(0.3, 3)$. A fraction C of the labels inside each cluster is flipped before the p-values are drawn, so that the signal cluster contains a proportion C of nulls and the null cluster a proportion C of alternatives, while the overall null proportion stays at $\bar{\alpha} \approx 0.75$. False discovery proportion and power are computed against the labels that generated the p-values, at $q = 0.10$, on ten independent datasets for each $C \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Figure 3 joins the three points of each dataset by a line.

BH sits at $\bar{\alpha}\, q$ for every C, as its theorem prescribes. Rule 1 rejects about as many hypotheses as BH at every C (Remark C.5 bounds its cardinality by the marginal rule) but with a far smaller false discovery proportion, from 0.00 at $C = 0$ to 0.04 at $C = 0.5$, since its ordering places the signal cluster first. Rule 2's power decreases as the clusters dilute, from 1.00 at $C = 0$ to 0.92, 0.84, 0.74, 0.70 and 0.64 at $C = 0.1$ to 0.5, and remains well above BH's (0.36–0.39) even at $C = 0.5$, where the two clusters are identical in composition and the only remaining structure is that the background is null. Its false discovery proportion is at or slightly above q for $C \geq 0.2$ (means 0.098, 0.087, 0.110, 0.097), which is the marginal-FDR-with-slack regime of Theorem F.5 rather than the exact control of the masked variant (Proposition F.7).

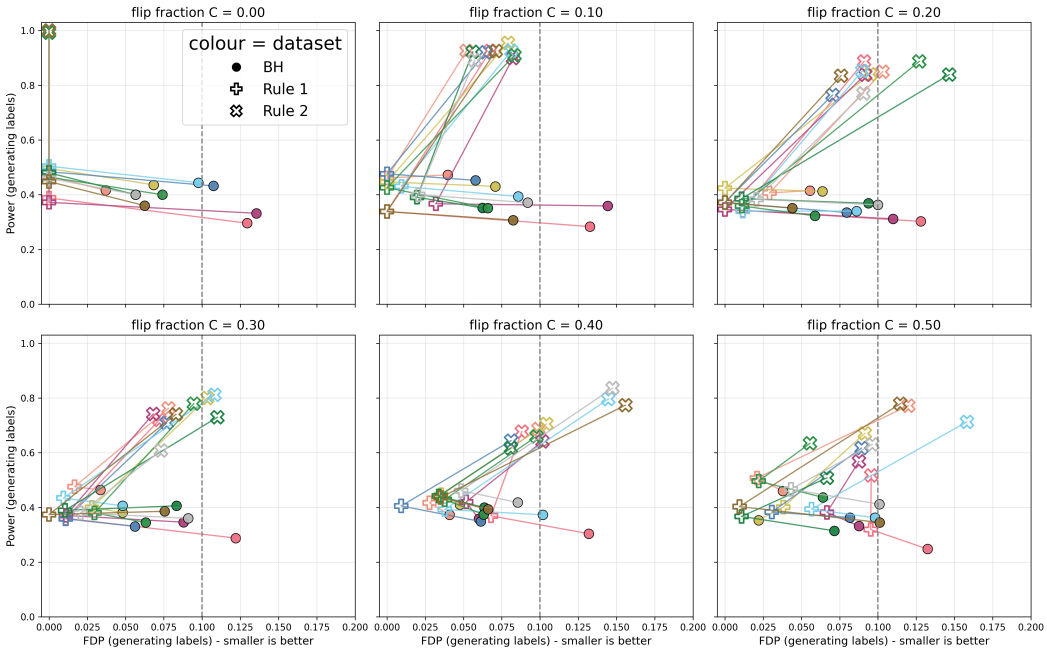

Figure 3: False discovery proportion versus power of BH, Rule 1 and Rule 2 under increasing label corruption C inside the clusters. Colour identifies the dataset; each line joins the three rules on one dataset. Rule 1 keeps its false discovery proportion far below q at every C; Rule 2 loses power as the clusters dilute but stays above BH even at $C = 0.5$.

Q Scalability Benchmarks

Table 3: GPU benchmarks on a single RTX 3090 (24 GB), **100 optimization iterations.**

N	Method	Optimization (100 iter)	GPU Memory
1,000	Dense f32	0.04 s	29 MB
5,000	Dense f32	0.02 s	513 MB
10,000	Dense f32	0.05 s	2.0 GB
20,000	Dense f32	0.20 s	8.0 GB
30,000	Dense f32	0.42 s	18.0 GB
40,000	Dense f16	0.42 s	16.0 GB
50,000	Dense f16	0.65 s	15.0 GB
75,000	Tiled	33 s	5.4 GB
100,000	Tiled	59 s	5.4 GB
200,000	Tiled	3 m 58 s	5.4 GB
500,000	Tiled	24 m 47 s	5.5 GB

Dense float32 (full kernel in memory) handles $N \leq 30$K with the complete pipeline, optimization, Hessian, and Cholesky, in under 4 seconds. **Dense float16** halves memory, extending the dense method to $N = 50$K with optimization still under 1 second. **Tiled matrix-vector products** compute kernel tiles on-the-fly without storing the full Gram matrix, keeping GPU memory flat at ~ 5.4 GB regardless of N, reaching $N = 500$K with **zero approximation error**. The tradeoff is purely computational: tiled is $\sim 50\times$ slower than dense at the same N, but any problem size fits on any GPU.